ESTÁNDARES
para Pruebas Educativas y Psicológicas

American Educational Research Association
American Psychological Association
National Council on Measurement in Education

ESTÁNDARES

para Pruebas Educativas y Psicológicas

American Educational Research Association
American Psychological Association
National Council on Measurement in Education

Este volumen es la traducción en Español de *Standards for Educational and Psychological Testing*, edición de 2014. Esta traducción se debe citar como sigue: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas* (M. Lieve, Trans.). Washington, DC: American Educational Research Association. (Original work published 2014)

Publicado por la
American Educational Research Association
1430 K St., NW, Suite 1200
Washington, DC 20005, EE. UU.

Impreso en los Estados Unidos de América

Preparado por el
Comité Conjunto sobre los *Estándares para Pruebas Educativas y Psicológicas* de la American Educational Research Association, la American Psychological Association y el National Council on Measurement in Education

ISBN 978-0-935302-74-5

Library of Congress Control Number: 2018937716

TABLA DE CONTENIDO

PARTE III
APLICACIONES DE LAS PRUEBAS

PREFACIO

La presente edición de los *Estándares para Pruebas Educativas y Psicológicas* está patrocinada por la American Educational Research Association (AERA; Asociación Estadounidense de Investigación Educativa), la American Psychological Association (APA; Asociación Estadounidense de Psicología) y el National Council on Measurement in Education (NCME; Consejo Nacional de Medición en Educación). Documentos anteriores de las organizaciones patrocinadoras también sirvieron de guía para el desarrollo y uso de pruebas. El primero fue las *Recomendaciones Técnicas para las Pruebas Psicológicas y las Técnicas de Diagnóstico*, elaborado por un comité de la APA y publicadas por la APA en 1954. El segundo fue las *Recomendaciones Técnicas para Pruebas de Rendimiento*, elaborado por un comité que representaba a la AERA y al National Council on Measurement Used in Education (NCMUE; Consejo Nacional sobre Medición Utilizada en Educación) y publicadas por la National Education Association (Asociación Nacional de Educación) en 1955.

El tercero, que reemplazó a los dos anteriores, fue elaborado por un comité conjunto que representaba a la AERA, la APA y el NCME y fue publicado por la APA en 1966. Se trató de la primera edición de los *Estándares para Pruebas Educativas y Psicológicas*, también conocidos como los *Estándares*. Las tres ediciones posteriores de los *Estándares* fueron elaboradas por comités conjuntos que representaban a la AERA, la APA y el NCME, publicadas en 1974, 1985 y 1999.

El actual Comité Directivo de los *Estándares* fue formado por la AERA, la APA y el NCME, las tres organizaciones patrocinadoras, en 2005, integrado por un representante de cada organización. Las responsabilidades del comité incluyeron determinar si era necesaria una revisión de los *Estándares* de 1999 y luego crear el cargo, presupuesto y cronograma de trabajo para un comité conjunto; designar copresidentes y miembros del comité conjunto; supervisar los asuntos financieros y un fondo de desarrollo; y realizar otras tareas relacionadas con la revisión y la publicación de los *Estándares*.

Comité Directivo de los *Estándares*

Wayne J. Camara (Presidente), designado por la APA
David Frisbie (2008—presente), designado por el NCME
Suzanne Lane, designada por la AERA
Barbara S. Plake (2005—2007), designada por el NCME

La presente edición de los *Estándares* fue desarrollada por el Comité Conjunto sobre los *Estándares para Pruebas Educativas y Psicológicas*, designado por el Comité de Directivo de los Estándares en 2008. Los miembros del Comité Conjunto son miembros de al menos una de las tres organizaciones patrocinadoras, AERA, APA y NCME. El Comité Conjunto tuvo a su cargo la revisión de los *Estándares* y la elaboración del documento final para su publicación. Su primera reunión tuvo lugar en enero de 2009.

Comité Conjunto sobre los Estándares para Pruebas Educativas y Psicológicas

Barbara S. Plake (Copresidente)
Lauress L. Wise (Copresidente)
Linda L. Cook
Fritz Drasgow
Brian T. Gong
Laura S. Hamilton
Jo-Ida Hansen
Joan L. Herman
Michael T. Kane
Michael J. Kolen
Antonio E. Puente
Paul R. Sackett
Nancy T. Tippins
Walter D. Way
Frank C. Worrell

Cada organización patrocinadora designó uno o dos intermediarios, algunos de los cuales eran miembros del Comité Conjunto, para actuar como canales de comunicación entre las

organizaciones patrocinadoras y el comité durante el proceso de revisión.

Intermediarios para el Comité Conjunto

AERA: Joan L. Herman
APA: Michael J. Kolen y Frank C. Worrell
NCME: Steve Ferrara

Marianne Ernesto (APA) se desempeñó como directora del proyecto para el Comité Conjunto, y Dianne L. Schneider (APA) actuó como la coordinadora del proyecto. Gerald Sroufe (AERA) proporcionó asistencia administrativa para el Comité Directivo. El asesor legal de la APA se ocupó de la revisión legal externa de los Estándares. Daniel R. Eignor y James C. Impara revisaron los *Estándares* en cuanto a precisión técnica y coherencia entre los capítulos.

En 2008, cada una de las tres organizaciones patrocinadoras emitió una convocatoria para presentar comentarios sobre los *Estándares* de 1999. En función de una revisión de los comentarios recibidos, el Comité Directivo identificó cuatro áreas principales de contenido en las que debía concentrarse la revisión: avances tecnológicos en materia de pruebas, aumento del uso de pruebas para rendición de cuentas y establecimiento de políticas educativas, acceso para todas las poblaciones de individuos examinados, y cuestiones asociadas con pruebas en el centro de trabajo. Además, el comité prestó especial atención a asegurar una voz común y al uso coherente de lenguaje técnico entre los capítulos.

En enero de 2011, se puso a disposición una versión preliminar de los *Estándares* revisados para revisión y comentarios del público. Las organizaciones que presentaron comentarios sobre la versión preliminar y/o comentarios en respuesta a la convocatoria de 2008 se indican a continuación. Muchas personas de cada organización aportaron comentarios, al igual que muchos miembros particulares de AERA, APA y NCME. El Comité Conjunto consideró cada comentario en su revisión de los *Estándares*. Estas revisiones razonadas de diversos puntos de observación profesionales ayudaron al Comité Conjunto en la elaboración de las revisiones finales de la presente edición de los *Estándares*.

Los comentarios provinieron de las siguientes organizaciones:

Organizaciones patrocinadoras

American Educational Research Association
American Psychological Association
National Council on Measurement in Education

Asociaciones profesionales

American Academy of Clinical Neuropsychology (Academia Estadounidense de Neuropsicología Clínica)
American Board of Internal Medicine (Consejo Estadounidense de Medicina Interna)
American Counseling Association (Asociación Estadounidense de Asesoramiento)
American Institute of CPAs (Instituto Estadounidense de Contadores Públicos Certificados), Equipo de Exámenes
Consejo para el Avance de la Psicología en el Interés Público de la APA
Consejo de Asuntos Educativos de la APA
Consejo de Asuntos Profesionales de la APA
Consejo de Asuntos Científicos de la APA
Consejo de Políticas y Planificación de la APA
Comité sobre Edad Avanzada de la APA
Comité sobre Niños, Jóvenes y Familias de la APA
Comité de Asuntos de las Minorías Étnicas de la APA
Comité de Relaciones Internacionales en Psicología de la APA
Comité de Asuntos Legales de la APA
Comité sobre Pruebas y Evaluación Psicológicas de la APA
Comité sobre Estado Socioeconómico de la APA
Sociedad para la Psicología de la Mujer de la APA (División 35)
División de Evaluación, Medición y Estadística de la APA (División 5)
División de Psicología Escolar de la APA (División 16)
Comité de Ética de la APA
Sociedad para la Psicología Industrial y Organizativa de la APA (División 14)
Sociedad de Psicología Clínica de Niños y Adolescentes de la APA (División 53)
Sociedad de Psicología de Asesoramiento de la APA (División 17) Asian American Psychological Association (Asociación Asiáticoestadounidense de Psicología) Association of Test Publishers (Asociación de Editores de Pruebas)

District of Columbia Psychological Association (Asociación Psicológica del Distrito de Columbia)
Massachusetts Neuropsychological Society (Sociedad Neuropsicológica de Massachusetts)
Massachusetts Psychological Association (Asociación Psicológica de Massachusetts)
National Academy of Neuropsychology (Academia Nacional de Neuropsicología)
National Association of School Psychologists (Asociación Nacional de Psicólogos Escolares)
National Board of Medical Examiners (Consejo Nacional de Examinadores Médicos)
National Council of Teachers of Mathematics (Consejo Nacional de Profesores de Matemáticas)
Junta Directiva del NCME
Comité sobre Cuestiones de Diversidad y Pruebas del NCME
Comité sobre Uso de Estándares y Pruebas del NCME

Compañías que realizan pruebas

ACT
Alpine Testing Solutions
The College Board
Educational Testing Service
Harcourt Assessment, Inc.
Hogan Assessment Systems
Pearson
Prometric
Vangent Human Capital Management
Wonderlic, Inc.

Instituciones académicas y de investigación

Centro para la Evaluación Educativa, Universidad de Massachusetts
Centro para la Equidad y la Excelencia en Educación de la Universidad George Washington
Human Resources Research Organization (HumRRO; Organización de Investigación en Recursos Humanos) Centro Nacional de Resultados Educativos,
Universidad de Minnesota

Organizaciones de acreditación

American Registry of Radiologic Technologists (Registro Estadounidense de Tecnólogos en Radiología) National Board for Certified Counselors (Consejo Nacional de Asesores Certificados) National Board of Medical Examiners (Consejo Nacional de Examinadores Médicos)

Otras instituciones

Departamento de Educación de California Consejo Asesor de Igualdad en el Empleo
Fair Access Coalition on Testing (Coalición de Acceso Justo sobre Pruebas)
Instituto de Evaluación e Ingeniería de Avanzada, México
Autoridad de Calificaciones y Planes de Estudio, Departamento de Educación del Reino Unido
Performance Testing Council (Consejo de Pruebas de Desempeño)

Cuando el Comité Conjunto completó su revisión final de los *Estándares*, presentó la revisión a las tres organizaciones patrocinadores para su aprobación y aval. Cada organización tuvo su propio órgano rector y mecanismo de aprobación, así como una declaración sobre el significado de su aprobación:

> **AERA**: La aprobación de los *Estándares* por parte de la AERA significa que el Consejo adopta el documento como política de la AERA.
>
> **APA**: La aprobación de los *Estándares* por parte de la APA significa que el Consejo de Representantes adopta el documento como política de la APA.
>
> **NCME**: Los *Estándares para Pruebas Educativas y Psicológicas* han sido avalados por el NCME, y este aval conlleva un imperativo ético para todos los miembros del NCME de adherir a estos estándares en la práctica de la medición.

Si bien los *Estándares* son prescriptivos, no contienen mecanismos de aplicación. Los Estándares se formularon con la intención de ser coherentes con otros estándares, pautas y códigos de conducta publicados por las tres organizaciones patrocinadoras.

Comité Conjunto sobre los *Estándares para Pruebas Educativas y Psicológicas*

INTRODUCCIÓN

La evaluación y las pruebas educativas y psicológicas se encuentran entre los aportes más importantes que las ciencias cognitivas y del comportamiento han hecho a nuestra sociedad, al brindar fuentes fundamentales y significativas de información sobre individuos y grupos. No todas las pruebas están bien desarrolladas, ni todas las prácticas de desarrollo de pruebas son sensatas o beneficiosas, pero existe amplia evidencia que documenta la utilidad de las pruebas bien construidas y bien interpretadas. Las pruebas bien construidas que son válidas para sus fines previstos presentan el potencial de brindar beneficios sustanciales para los examinandos y los usuarios de las pruebas. Su uso adecuado puede dar lugar a mejores decisiones sobre individuos y programas que las que se generarían sin su uso y también pueden proporcionar un camino hacia un acceso más amplio y equitativo a la educación y el empleo. El uso inadecuado de las pruebas, por otra parte, puede dar lugar a un daño considerable para los examinandos y otras partes afectadas por las decisiones basadas en las pruebas. La intención de los *Estándares para Pruebas Educativas y Psicológicas* es promover prácticas sólidas de desarrollo de pruebas y brindar una base para evaluar la calidad de esas prácticas. Los *Estándares* están dirigidos a profesionales que especifican, desarrollan o seleccionan pruebas y para quienes interpretan los resultados de las pruebas o evalúan su calidad técnica.

La finalidad de los *Estándares*

La finalidad de los *Estándares* es proporcionar criterios para el desarrollo y la evaluación de pruebas y prácticas de desarrollo de pruebas y brindar pautas para evaluar la validez de las interpretaciones de los puntajes de las pruebas para los usos previstos de las pruebas. Si bien esas evaluaciones deberían depender ampliamente del juicio profesional, los *Estándares* brindan un marco de referencia para garantizar que se aborden cuestiones relevantes. Todos los desarrolladores, patrocinadores, editores y usuarios profesionales de pruebas deben hacer esfuerzos razonables para cumplir y seguir los *Estándares* y deben alentar a los demás a hacerlo. Todos los estándares aplicables deben ser cumplidos por todas las pruebas y en todos los usos de las pruebas a menos que exista un motivo profesional sólido que demuestre por qué un estándar no es relevante o técnicamente viable en un caso en particular.

Los *Estándares* no intentan proporcionar respuestas psicométricas a preguntas de política pública respecto del uso de pruebas. En general, los *Estándares* proponen que, dentro de límites viables, se ponga a disposición información técnica de modo que los involucrados en las decisiones sobre políticas puedan estar plenamente informados.

Descargo de responsabilidad legal

Los *Estándares* no constituyen una declaración de requisitos legales, y el cumplimiento con los *Estándares* no sustituye el asesoramiento legal. Numerosas leyes, regulaciones, normas y decisiones judiciales federales, estatales y locales se relacionan con algunos aspectos del uso, la producción, el mantenimiento y el desarrollo de pruebas y resultados de pruebas e imponen estándares que pueden ser diferentes para los diferentes tipos de pruebas. La revisión de estas cuestiones legales excede el alcance de los *Estándares*, cuyo propósito distintivo es establecer los criterios para prácticas sólidas de desarrollo de pruebas desde la perspectiva de profesionales de las ciencias cognitiva y del comportamiento. En los casos en que al parecer uno o más estándares abordan una cuestión respecto de la cual los requisitos legales establecidos pueden ser especialmente relevantes, el estándar, comentario o material introductorio puede tomar nota de ese hecho. La falta de referencia específica a requisitos legales, no obstante, no implica la ausencia de un requisito legal relevante.

Al aplicar estándares a nivel internacional, las diferencias legales pueden dar lugar a cuestiones adicionales o requerir un tratamiento diferente de las cuestiones.

En algunas áreas, como la recopilación, análisis y uso de datos y resultados de pruebas para diferentes subgrupos, la ley puede tanto requerir que los participantes en el proceso de prueba hagan determinadas acciones como prohibir que esos participantes hagan otras acciones. Asimismo, debido a que la ciencia de las pruebas es una disciplina en evolución, es posible que las revisiones recientes de los *Estándares* no se reflejen en autoridades legales existentes, incluidas decisiones judiciales y pautas de organismos. En todas las situaciones, los participantes en el proceso de prueba deberían obtener el consejo de un asesor respecto de los requisitos legales aplicables.

Además, si bien las organizaciones patrocinadoras no pueden hacer cumplir los *Estándare*s, las autoridades de regulación y los tribunales los han reconocido en reiteradas oportunidades como el establecimiento de estándares profesionales generalmente aceptados que siguen los desarrolladores y usuarios de pruebas y otros procedimientos de selección. El cumplimiento o incumplimiento de los *Estándares* puede utilizarse como evidencia relevante de responsabilidad legal en procedimientos judiciales y regulatorios. Los *Estándares*, por lo tanto, merecen la consideración atenta de todos los participantes en el proceso de prueba.

Ninguna parte de los *Estándares* tiene por objeto constituir asesoramiento legal. Además, los editores niegan toda responsabilidad generada por la participación en el proceso de prueba.

Pruebas y usos de las pruebas a los que se aplican estos Estándares

Una prueba es un dispositivo o procedimiento en el cual se obtiene y posteriormente se evalúa y califica una muestra del comportamiento de un individuo examinado en un dominio especificado, a través de un proceso estandarizado. Si bien el término *prueba* en ocasiones se reserva a instrumentos en los que las respuestas se evalúan según su corrección o calidad, y los términos *escala* e *inventario* se utilizan para medidas de actitudes, interés y disposiciones, los *Estándares* utilizan el único término prueba para referirse a todos esos dispositivos evaluativos.

En ocasiones se hace una distinción entre pruebas y evaluaciones. *Evaluación* es un término más amplio que prueba; comúnmente se refiere a un proceso que integra la información de la prueba con información de otras fuentes (p. ej., información de otras pruebas, inventarios y entrevistas; o de los antecedentes sociales, educativos, laborales, de salud o psicológicos de la persona). La aplicabilidad de los *Estándares* a un dispositivo o método de evaluación se determina por el contenido y no se altera por el término aplicado a este (p. ej., prueba, evaluación, escala, inventario). Los *Estándares* no deben utilizarse como una lista de comprobación, como se destaca en la sección "Precauciones que deben considerarse al utilizar los *Estándares*" al final de este capítulo.

Las pruebas difieren en una serie de dimensiones: el modo en que se presentan los materiales de la prueba (p. ej., papel y lápiz, administración oral o por computadora); el grado con el que se estandarizan los materiales de estímulo; el tipo de formato de respuesta (selección de una respuesta de un conjunto de alternativas, en oposición a la producción de una respuesta en forma libre); y el grado con el que se diseñan los materiales de la prueba para reflejar o simular un contexto en particular. En todos los casos, no obstante, las pruebas estandarizan el proceso mediante el cual se evalúan y califican las respuestas de los examinandos a los materiales de la prueba. Como se observó en versiones anteriores de los *Estándares*, se necesitan los mismos tipos generales de información para juzgar la solidez de los resultados obtenidos del uso de todas las variedades de pruebas.

La demarcación precisa entre dispositivos de medición utilizados en los campos de las pruebas educativas y psicológicas que se encuadran y no se encuadran dentro del alcance de los Estándares es difícil de identificar. Si bien los *Estándares* se aplican de manera más directa a medidas estandarizadas generalmente reconocidas como "pruebas", como medidas de habilidad, aptitud, rendimiento, actitudes, intereses, personalidad,

funcionamiento cognitivo y salud mental, los Estándares también pueden aplicarse con utilidad en diversos grados a una amplia variedad de técnicas de evaluación menos formales. La aplicación rigurosa de los *Estándares* a evaluaciones de empleo no estandarizadas (como algunas entrevistas de trabajo) o a la amplia variedad de muestras de comportamiento no estructurado utilizadas en algunas formas de evaluación clínica y psicológica escolar (p. ej., una entrevista de admisión) o a pruebas hechas por instructores que se utilizan para evaluar el desempeño estudiantil en educación y capacitación, por lo general no es posible. Resulta útil distinguir entre dispositivos que reivindican los conceptos y técnicas del campo de las pruebas educativas y psicológicas y los dispositivos que representan ayudas no estandarizadas o menos estandarizadas a las decisiones evaluativas diarias. Si bien los principios y conceptos subyacentes a los *Estándares* pueden aplicarse con éxito a las decisiones diarias —como cuando un empresario entrevista a un solicitante de empleo, un gerente evalúa el desempeño de subordinados, un profesor desarrolla una evaluación en el aula para monitorear el progreso de los estudiantes hacia una meta educativa, o un entrenador evalúa a un futuro deportista—, sería excesivo esperar que quienes toman esas decisiones sigan los estándares del campo de las pruebas educativas y psicológicas. Por el contrario, un sistema de entrevistas estructurado desarrollado por un psicólogo y acompañado por afirmaciones de que se ha determinado que el sistema es predictivo del desempeño laboral en diversos otros contextos se encuadra dentro del alcance de los *Estándares*. Adherir a los *Estándares* se vuelve más crítico a medida que aumentan los riesgos para el examinando y la necesidad de proteger al público.

Participantes en el proceso de prueba

La evaluación y las pruebas educativas y psicológicas involucran y afectan significativamente a individuos, instituciones y a la sociedad en su conjunto. Los individuos afectados incluyen estudiantes, padres, familias, profesores, administradores educativos, solicitantes de puestos de trabajo, empleados, clientes, pacientes, supervisores, ejecutivos y evaluadores, entre otros. Las instituciones afectadas incluyen escuelas, universidades, empresas, la industria, clínicas psicológicas y organismos gubernamentales. Los individuos y las instituciones se benefician cuando las pruebas los ayudan a alcanzar sus metas. La sociedad, a su vez, se beneficia cuando las pruebas contribuyen al logro de metas individuales e institucionales.

Hay muchos participantes en el proceso de prueba, que incluyen, entre otros, los siguientes: (a) los que preparan y desarrollan la prueba; (b) los que publican y comercializan la prueba; (c) los que administran y califican la prueba; (d) los que interpretan los resultados de la prueba para los clientes; (e) los que utilizan los resultados de la prueba para algún fin de toma de decisiones (incluidos los responsables de formular políticas y quienes utilizan datos para informar la política social); (f) los que se someten a la prueba por elección, instrucción o necesidad; (g) los que patrocinan las pruebas, como juntas que representan a instituciones u organismos gubernamentales que tienen contrato con una desarrollador de pruebas para un instrumento o servicio específico; y (h) los que seleccionan o revisan las pruebas, evaluando sus méritos comparativos o la aptitud para los usos propuestos. En general, quienes participan en el proceso de prueba deben tener conocimiento adecuado de las pruebas y evaluaciones para permitirles tomar buenas decisiones sobre qué pruebas usar y cómo interpretar los resultados de las pruebas.

Los intereses de las diversas partes involucradas en el proceso de prueba pueden ser congruentes o no. Por ejemplo, cuando se toma una prueba para fines de asesoramiento o para una colocación laboral, los intereses del individuo y de la institución suelen coincidir. Por el contrario, cuando una prueba se utiliza para hacer una selección de entre muchos individuos para un puesto altamente competitivo o para ingresar en un programa educativo o de capacitación, es posible que las preferencias de un solicitante no coincidan con las de un empleador o responsable de admisiones. De manera similar, cuando las pruebas son ordenadas por un tribunal, los intereses

del examinando pueden ser diferentes de los de la parte que solicita la orden judicial.

Los individuos o instituciones pueden cumplir varios roles en el proceso de prueba. Por ejemplo, en clínicas el examinando suele ser el beneficiario previsto de los resultados de la prueba. En algunas situaciones, el administrador de la prueba es un representante del desarrollador de la prueba, y en ocasiones el administrador de la prueba es también el usuario de la prueba. Cuando una organización prepara sus propias pruebas de empleo, es tanto el desarrollador como el usuario. A veces, una prueba es desarrollada por un autor de la prueba, pero es luego publicada, comercializada y distribuida por un editor independiente, aunque el editor puede desempeñar un rol activo en el proceso de desarrollo de la prueba. Los roles, a su vez, también pueden subdividirse. Por ejemplo, tanto una organización como un evaluador profesional pueden desempeñar un rol en la provisión de un centro de evaluaciones. Dada esta mezcla de roles, suele ser difícil asignar la responsabilidad precisa de abordar diversos estándares a participantes específicos en el proceso de prueba. Los usos de pruebas y prácticas de desarrollo de pruebas se mejoran en la medida en que las personas involucradas tienen niveles adecuados de conocimientos en evaluación.

Las pruebas son diseñadas, desarrolladas y utilizadas de diversas maneras. En algunos casos, son desarrolladas y "publicadas" para usarse fuera de la organización que las produce. En otros casos, al igual que las evaluaciones educativas, son diseñadas por el organismo educativo estatal y desarrolladas por contratistas para uso exclusivo y a menudo por única vez del estado y en realidad no se "publican". A lo largo de los *Estándares*, utilizamos el término general *desarrollador de la prueba*, en lugar del término más específico *editor de la prueba*, para hacer referencia a las personas involucradas en el diseño y desarrollo de pruebas en toda la gama competa de escenarios de desarrollo de pruebas.

Los *Estándares* parten de la premisa de que las pruebas y evaluaciones efectivas requieren que todos los profesionales del proceso de desarrollo de la prueba tengan el conocimiento, las habilidades y las capacidades necesarias para cumplir sus roles, así como un conocimiento de factores personales y contextuales que pueden influir en el proceso de desarrollo de la prueba. Por ejemplo, los desarrolladores de pruebas y los que seleccionan pruebas e interpretan los resultados de las pruebas necesitan un conocimiento adecuado de los principios psicométricos como validez y confiabilidad. También deben obtener cualquier credencial de experiencia supervisada y de ejercicio obligatoria por ley que corresponda, que se requiera para cumplir de manera competente con todos los aspectos del proceso de desarrollo de la prueba en el que participen. Todos los profesionales en el proceso de desarrollo de la prueba deben seguir las pautas éticas de su profesión.

Alcance de la revisión

Este volumen funciona como una revisión de los *Estándares para Pruebas Educativas y Psicológicas* de 1999. El proceso de revisión comenzó con la designación de un Comité Directivo compuesto por representantes de las tres organizaciones patrocinadoras responsables de supervisar la dirección general de la iniciativa: la American Educational Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME). Para brindar orientación para la revisión, el Comité Directivo solicitó y resumió comentarios sobre los *Estándares* de 1999 de miembros de las organizaciones patrocinadores y en 2009 convocó al Comité Conjunto para la Revisión de los *Estándares* de 1999 para que efectuara la revisión propiamente dicha. El Comité Conjunto también estuvo compuesto por miembros de las tres organizaciones patrocinadoras y el Comité Directivo le encargó abordar cinco áreas principales: considerar las cuestiones de rendición de cuentas para uso de pruebas en política educativa; ampliar el concepto de accesibilidad de las pruebas para todos los individuos examinados; representar de manera más completa el rol de las pruebas en el centro de trabajo; ampliar el rol de la tecnología en el desarrollo de pruebas; y disponer una mejor estructura organizativa para comunicar los estándares.

Para responder a esta tarea, se tomaron varias medidas:

- Los capítulos "Pruebas y evaluación educativas" y "Pruebas en la evaluación de programas y política pública", en la versión de 1999, se reescribieron para atender a las cuestiones asociadas con los usos de pruebas para fines de rendición de cuentas en materia educativa.
- Se escribió un nuevo capítulo, "Imparcialidad en las pruebas" para hacer hincapié en la accesibilidad y la imparcialidad como cuestiones fundamentales en las pruebas. A lo largo de todos los capítulos de los *Estándares* se hilvanan cuestiones específicas sobre imparcialidad.
- El capítulo "Pruebas relacionadas con empleo y acreditación" (ahora, "Pruebas y acreditación en el centro de trabajo") se reorganizó para identificar de manera más clara cuándo un estándar es relevante para el empleo y/o acreditación.
- En todo el volumen se consideró el impacto de la tecnología. Uno de los principales problemas que se identificaron en relación con la tecnología fue la tensión entre el uso de algoritmos patentados y la necesidad de que los usuarios de pruebas pueden evaluar aplicaciones complejas en áreas como calificación automatizada de ensayos, administración y calificación de tipos de ítems innovadores y pruebas basadas en computadora. Estos problemas se consideran en el capítulo "Diseño y desarrollo de pruebas".
- Se contrató a un editor de contenidos para que ayudara con la precisión y claridad técnicas de cada capítulo y con la coherencia de lenguaje entre los capítulos. Como se observa a continuación, los capítulos de la Parte I ("Fundamentos") y de la Parte II ("Operaciones") ahora tienen un "estándar global" y temas en los que se organizan los estándares individuales. Además, se actualizó el glosario de los *Estándares para Pruebas Educativas y Psicológicas* de 1999. Como se observó anteriormente, un cambio importante en la organización de este volumen tiene que ver con la conceptualización de la imparcialidad. La edición de 1999 tenía una parte dedicada a este tema, con capítulos separados titulados "Imparcialidad en las pruebas y uso de pruebas", "Pruebas a personas de características lingüísticas diversas" y "Pruebas a personas con discapacidades". En la presente edición, los temas abordados en esos capítulos se combinan en un único capítulo integral, y el capítulo se encuentra en la Parte I. Este cambio se hizo para destacar que la imparcialidad exige que todos los examinandos sean tratados con imparcialidad. La imparcialidad y la accesibilidad, la oportunidad no obstruida para que todos los individuos examinados demuestren su situación en el o los constructos que se miden, son relevantes para hacer interpretaciones válidas de los puntajes para todos los individuos y subgrupos en la población prevista de examinandos. Debido a que las cuestiones relacionadas con la imparcialidad en las pruebas no se restringen a individuos con características lingüísticas diversas o con discapacidades, el capítulo se amplió para recoger experiencias de pruebas adecuadas para todos los individuos. Si bien los ejemplos del capítulo suelen referirse a individuos con características lingüísticas y culturales diversas y a individuos con discapacidades, también incluyen ejemplos relevantes al género y a adultos mayores, personas de diversos orígenes étnicos y raciales, y niños pequeños, para ilustrar los posibles obstáculos a una evaluación imparcial y equitativa para todos los individuos examinados.

Organización del volumen

La Parte I de los *Estándares*, "Fundamentos", contiene estándares de validez (cap. 1); confiabilidad/precisión y errores de medición (cap. 2); e imparcialidad en las pruebas (cap. 3). La Parte II, "Operaciones", aborda el diseño y desarrollo de pruebas (cap. 4); puntajes, escalas, normas, vinculación de puntajes y puntajes de corte (cap. 5); administración de pruebas, calificación, presentación

de reportes e interpretación (cap. 6); documentación de apoyo para las pruebas (cap. 7); los derechos y responsabilidades de los examinandos (cap. 8); y los derechos y responsabilidades de los usuarios de las pruebas (cap. 9). La Parte III, "Aplicaciones de las pruebas" trata aplicaciones específicas en pruebas y evaluación psicológicas (cap. 10); pruebas y acreditación en el centro de trabajo (cap. 11); pruebas y evaluación educativas (cap. 12); y usos de pruebas para evaluación de programas, estudios de políticas y rendición de cuentas (cap. 13). Asimismo, se incluye un glosario, que ofrece definiciones de términos según se utilizan específicamente en este volumen.

Cada capítulo comienza con un texto introductorio que brinda los antecedentes para los estándares que siguen. Si bien en ocasiones el texto introductorio es prescriptivo, no debe interpretarse como la imposición de estándares adicionales.

Categorías de estándares

El texto de cada estándar y cualquier comentario que lo acompañe incluyen las condiciones en las que un estándar es relevante. Dependiendo del contexto y la finalidad del desarrollo o uso de la prueba, algunos estándares serán más destacados que otros. Además, algunos estándares tienen un alcance amplio, al establecer cuestiones o requisitos relevantes para casi todas las pruebas o contextos de pruebas, y otros estándares tienen un alcance más acotado. Sin embargo, todos los estándares son importantes en los contextos a los que se aplican. Cualquier clasificación que parezca elevar la importancia general de algunos estándares por sobre otros podría invitar a desatender determinados estándares que deben abordarse en situaciones particulares. En lugar de diferenciar los estándares utilizando rótulos de prioridad, como "primario", "secundario" o "condicional" (como se utilizaron en los *Estándares* de 1985), esta edición destaca que a menos que un estándar se considere claramente irrelevante, inapropiado o técnicamente inviable para un uso en particular, todos los estándares deben cumplirse, lo que hace que todos sean esencialmente "primarios" para ese contexto.

A menos que se especifique lo contrario en un estándar o comentario, y con las advertencias que se describen a continuación, los estándares deben cumplirse antes del uso operativo de la prueba. Cada estándar debe considerarse atentamente para determinar su aplicabilidad al contexto de prueba en consideración. En un caso determinado, es posible que haya un motivo profesional sólido por el que sea inadecuado adherir al estándar. También es posible que haya ocasiones en las que la viabilidad técnica influya a que un estándar pueda cumplirse o no antes del uso operativo de la prueba. Por ejemplo, algunos estándares pueden requerir análisis de datos que no están disponibles en el momento del uso operativo inicial de la prueba. En algunos casos, es posible que los análisis cuantitativos tradicionales no sean viables debido a tamaños de muestra pequeños. Sin embargo, puede haber otras metodologías que podrían utilizarse para reunir información para respaldar el estándar, como metodologías para muestras pequeñas, estudios cualitativos, grupos focales e, incluso, análisis lógico. En esos casos, los desarrolladores y usuarios de la prueba deben hacer un esfuerzo de buena fe para proporcionar los tipos de datos requeridos en el estándar para respaldar las interpretaciones válidas de los resultados de la prueba para sus fines previstos. Si los desarrolladores, usuarios y, cuando corresponda, patrocinadores de la prueba han considerado que un estándar es inaplicable o técnicamente inviable, deben poder explicar, si se les solicita, el fundamento de su decisión. Sin embargo, no existe expectativa de que la documentación de todas esas decisiones esté habitualmente disponible.

Presentación de estándares individuales

Los estándares individuales se presentan después de un texto introductorio que presenta algunos conceptos claves para interpretar y aplicar los estándares. En muchos casos, los estándares propiamente dichos están acompañados de uno o más comentarios. Estos comentarios tienen por objeto ampliar, aclarar o brindar ejemplos para contribuir a la interpretación del significado de

los estándares. Los estándares a menudo le indican a un desarrollador o usuario que implemente determinadas acciones. Según el tipo de prueba, en ocasiones no está claro en el enunciado de un estándar a quién está dirigido el estándar. Por ejemplo, el Estándar 1.2 en el capítulo "Validez" indica:

> Se debe presentar una razón fundamental para cada interpretación prevista de los puntajes de la prueba para un uso determinado, junto con un resumen de la evidencia y la teoría que inciden en la interpretación prevista.

La parte responsable de implementar este estándar es la parte o persona que está articulando la interpretación recomendada de los puntajes de la prueba. Esta puede ser un usuario de la prueba, un desarrollador de la prueba o alguien que esté planeando usar los puntajes de la prueba para un fin en particular, como tomar decisiones de clasificación u otorgamiento de licencias. A menudo no es posible especificar en el enunciado de un estándar quién es responsable de dichas acciones; se tiene la intención de que la parte o persona que realiza la acción especificada en el estándar sea la parte responsable de adherir al estándar.

Algunos de los estándares individuales y el texto introductorio se refieren a grupos y subgrupos. El término *grupo* por lo general se utiliza para identificar a la población completa de individuos examinados, referida como el *grupo previsto de individuos examinados*, el *grupo previsto de examinandos*, la *población prevista de individuos examinados*, o la *población*. Un *subgrupo* incluye miembros de un grupo más amplio que son identificables de alguna manera que sea relevante para el estándar que se aplica. Cuando los datos o los análisis se indican para varios subgrupos, por lo general se los denomina *subgrupos dentro del grupo previsto de individuos examinados*, *grupos de la población prevista de individuos examinados*, o *subgrupos relevantes*.

Al aplicar los *Estándares*, es importante tener presente que los subgrupos referentes previstos para los estándares individuales son específicos del contexto. Por ejemplo, los subgrupos étnicos referentes que se considerarán durante la fase de diseño de una prueba dependerían de la composición étnica esperada del grupo de prueba previsto. Además, muchos más subgrupos podrían ser relevantes para un estándar relacionado con el diseño de preguntas imparciales de la prueba que para un estándar que se relacione con adaptaciones del formato de una prueba. Los usuarios de los *Estándares* deberán ejercer su juicio profesional al decidir qué subgrupos en particular son relevantes para la aplicación de un estándar específico.

Al decidir qué subgrupos son relevantes para un estándar en particular, pueden considerarse, entre otros, los siguientes factores: evidencia creíble que sugiera que un grupo puede enfrentar obstáculos particulares irrelevantes del constructo para evaluar el desempeño, leyes o regulaciones que designan a un grupo como relevante para interpretaciones de puntajes, y grandes cantidades de individuos en el grupo dentro de la población general. Dependiendo del contexto, los subgrupos relevantes podrían incluir, por ejemplo, hombres y mujeres, individuos de diferente nivel socioeconómico, individuos diferentes en cuanto a raza y/u origen étnico, individuos con diferentes orientaciones sexuales, individuos con características lingüísticas y culturales diversas (en particular cuando las pruebas se realizan a nivel internacional), individuos con discapacidades, niños pequeños o adultos mayores.

Se brindan numerosos ejemplos en los *Estándares* para aclarar puntos o proporcionar ilustraciones de cómo aplicar un estándar en particular. Muchos de los ejemplos se extraen de investigaciones con estudiantes con discapacidades o personas de grupos de lenguaje o culturales diversos; una cantidad menor, de investigaciones con grupos identificables, como niños pequeños o adultos. También se realizó un esfuerzo mayor para proporcionar ejemplos de contextos educativos, psicológicos e industriales.

Los estándares en cada capítulo de las Partes I y II ("Fundamentos" y "Operaciones") son introducidos por un estándar global, diseñado para transmitir la intención central del capítulo. Estos estándares globales están siempre numerados

con .0 tras el número de capítulo. Por ejemplo, el estándar global en el capítulo 1 está numerado 1.0. Los estándares globales resumen los principios rectores aplicables a todas las pruebas y usos de pruebas. Además, los temas y estándares en cada capítulo están ordenados para guardar coherencia con la secuencia del material en el texto introductorio del capítulo. Debido a que algunos usuarios de los *Estándares* pueden consultar solo los capítulos directamente relevantes para una aplicación determinada, ciertos estándares se repiten en diferentes capítulos, en especial en la Parte III, "Aplicaciones de las pruebas". Cuando ocurre esa repetición, la esencia del estándar es la misma. Solo se cambia la redacción, el área de aplicación o el nivel de elaboración en el comentario.

Precauciones que deben considerarse al utilizar los *Estándares*

Además del descargo de responsabilidad legal establecido anteriormente, varias precauciones son importantes si se quieren evitar malas interpretaciones, aplicaciones incorrectas o usos indebidos de los *Estándares*:

- Evaluar la aceptabilidad de una prueba o aplicación de una prueba no depende de la satisfacción literal de cada estándar en este documento, y la aceptabilidad de una prueba o aplicación de una prueba no puede determinarse utilizando una lista de comprobación. Circunstancias específicas afectan la importancia de los estándares individuales, y los estándares individuales no deben considerarse en forma aislada. Por lo tanto, evaluar la aceptabilidad depende de lo siguiente: (a) el juicio profesional que se basa en un conocimiento de la ciencia del comportamiento, psicometría, y los estándares relevantes en el campo profesional al que se aplica la prueba; (b) el grado con el que el desarrollador y el usuario de la prueba hayan satisfecho la intención del estándar; (c) los dispositivos de medición alternativos que estén inmediatamente disponibles; (d) evidencia de investigaciones y empírica respecto de la viabilidad de cumplir el estándar; y (e) leyes y regulaciones aplicables.
- Cuando las pruebas están sobre el tapete en procedimientos judiciales y otras situaciones que requieren el dictamen de peritos, es importante que el juicio profesional se base en el corpus aceptado de conocimientos al determinar la relevancia de estándares particulares en una situación dada. La intención de los *Estándares* es ofrecer orientación para dichos juicios.
- Las afirmaciones de los desarrolladores de pruebas o usuarios de pruebas respecto de que una prueba, manual o procedimiento satisface o sigue los estándares en este volumen deben hacerse con cuidado. Es apropiado que los desarrolladores o usuarios indiquen que se hicieron esfuerzos por adherir a los *Estándares*, y que proporcionen documentos que describan y respalden esos esfuerzos. No deben hacerse afirmaciones generales sin evidencia que las sustente.
- Los estándares se relacionan con un campo de rápida evolución. En consecuencia, existe la necesidad continua de monitorear cambios en el campo y revisar este documento a medida que se desarrollan conocimientos. El uso de versiones anteriores de los *Estándares* puede constituir un perjuicio para los usuarios de pruebas y los examinandos.
- No es la intención de los *Estándares* requerir el uso de métodos técnicos específicos. Por ejemplo, en los casos en que se mencionen requisitos de presentación de reportes estadísticos específicos, siempre debe entenderse la frase "o un equivalente generalmente aceptado".

PARTE I

I Fundamentos

1. VALIDEZ

ANTECEDENTES

La validez se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para usos propuestos de las pruebas. La validez es, por lo tanto, la consideración más fundamental al desarrollar y evaluar pruebas. El proceso de validación involucra acumular evidencia pertinente para proporcionar una base científica sólida para las interpretaciones de puntajes propuestas. Lo que se evalúa son las interpretaciones de los puntajes de la prueba para los usos propuestos, no la prueba propiamente dicha. Cuando los puntajes de la prueba se interpretan en más de una manera (p. ej., tanto para describir el nivel actual del atributo que se mide del examinando como para hacer una predicción sobre un futuro resultado), cada interpretación prevista debe validarse. Los enunciados sobre la validez deben referirse a interpretaciones particulares para usos especificados. Es incorrecto usar la frase no calificada "la validez de la prueba".

La evidencia de la validez de una interpretación dada de puntajes de la prueba para un uso especificado es una condición necesaria para el uso justificado de la prueba. Cuando existe evidencia suficiente de validez, la decisión en cuanto a administrar efectivamente o no una prueba en particular suele tener en cuenta otras consideraciones. Estas incluyen consideraciones sobre costo-beneficio, enmarcadas en subdisciplinas diferentes como análisis de utilidad o como consideración de consecuencias negativas del uso de la prueba, y una ponderación de cualquier consecuencia negativa frente a las consecuencias positivas del uso de la prueba.

La validación lógicamente comienza con un enunciado explícito de la interpretación propuesta de los puntajes de la prueba, junto con una razón fundamental para la relevancia de la interpretación para el uso propuesto. La interpretación propuesta incluye especificar el constructo que la prueba intenta medir. El término *constructo* se utiliza en los *Estándares* para referirse al concepto o característica para cuya medición se diseña una prueba. Casi nunca, o nunca, existe un solo significado posible que puede atribuirse al puntaje de una prueba o a un patrón de respuestas de la prueba. Por lo tanto, siempre corresponde a los desarrolladores y usuarios de la prueba especificar la interpretación del constructo que se hará en función del puntaje o patrón de respuestas.

Entre los ejemplos de constructos que se utilizan actualmente en evaluación se incluyen rendimiento matemático, capacidad cognitiva general, actitudes de identidad racial, depresión y autoestima. Para apoyar el desarrollo de la prueba, la interpretación del constructo propuesta se elabora describiendo su alcance y extensión y delineando los aspectos del constructo que se representarán. La descripción detallada proporciona un marco conceptual para la prueba, delineando el conocimiento, habilidades, capacidades, rasgos, intereses, procesos, competencias o características a evaluar. Idealmente, el marco indica cómo el constructo según lo representado debe distinguirse de otros constructos y cómo debe relacionarse con otras variables.

El marco conceptual se forma en parte por las maneras en que se utilizarán los puntajes de la prueba. Por ejemplo, una prueba de rendimiento matemático podría usarse para colocar a un estudiante en un programa de instrucción adecuado, para respaldar un diploma de escuela secundaria o para informar una decisión sobre admisiones universitarias. Cada uno de estos usos implica una interpretación un tanto diferente de los puntajes de la prueba de rendimiento matemático: que un estudiante se beneficiará con una intervención de instrucción en particular, que un estudiante ha dominado un plan de estudios especificado, o que es probable que un estudiante tenga éxito con el trabajo de nivel universitario. De manera similar, una prueba de responsabilidad podría

utilizarse para asesoramiento psicológico, para informar una decisión sobre empleo, o para el fin científico básico de elaborar el constructo de responsabilidad. Cada uno de estos usos potenciales da forma al marco especificado y la interpretación propuesta de los puntajes de la prueba y también puede tener implicaciones para el desarrollo y la evaluación de la prueba. La validación puede verse como un proceso de construir y evaluar argumentos a favor y en contra de la interpretación prevista de los puntajes de la prueba y su relevancia para el uso propuesto. El marco conceptual señala las clases de evidencia que podrían reunirse para evaluar la interpretación propuesta teniendo en cuenta los fines de la prueba. A medida que la validación continúa y surge nueva evidencia respecto de las interpretaciones que pueden y no pueden extraerse de los puntajes de la prueba, es posible que se necesiten revisiones en la prueba, en el marco conceptual que la forma e, incluso, en el constructo subyacente de la prueba.

La amplia variedad de pruebas y circunstancias hace que sea normal que algunos tipos de evidencia sean especialmente críticos en un caso determinado, mientras que otros tipos serán menos útiles. Las decisiones sobre qué tipos de evidencia son importantes para el argumento de validación en cada caso pueden aclararse desarrollando un conjunto de proposiciones o afirmaciones que respalden la interpretación propuesta para el fin particular de la prueba. Por ejemplo, cuando se utiliza una prueba de rendimiento matemático para evaluar la preparación para un curso avanzado, la evidencia para las siguientes proposiciones podría ser relevante: (a) que determinadas habilidades son prerrequisito para el curso avanzado; (b) que el dominio de contenido de la prueba guarda coherencia con estas habilidades de prerrequisito; (c) que los puntajes de la prueba pueden generalizarse entre conjuntos de ítems relevantes; (d) que los puntajes de la prueba no están indebidamente influenciados por variables auxiliares, como la capacidad de escritura; (e) que el éxito en el curso avanzado puede evaluarse válidamente, y (f) que los examinandos con puntajes altos en la prueba serán más exitosos en el curso avanzado que los examinandos con puntajes bajos en la prueba. Ejemplos de proposiciones en otros contextos de pruebas podrían incluir, por ejemplo, la proposición de que los examinandos con puntajes altos de ansiedad general experimentan ansiedad significativa en una serie de contextos, la proposición de que el puntaje de un niño en una escala de inteligencia se relaciona fuertemente con el desempeño académico del niño, o la proposición de que un cierto patrón de puntajes en una batería neuropsicológica indica afectación que es característica de lesión cerebral. El proceso de validación evoluciona a medida que se articulan estas proposiciones y se reúne evidencia para evaluar su solidez.

La identificación de las proposiciones implícitas por una interpretación propuesta de la prueba puede facilitarse considerando hipótesis rivales que pueden desafiar la interpretación propuesta. También es útil considerar las perspectivas de diferentes partes interesadas, la experiencia existente con pruebas y contextos similares, y las consecuencias previstas del uso propuesto de la prueba. El hallazgo de consecuencias imprevistas del uso de la prueba también puede dar lugar a una consideración de hipótesis rivales. A menudo pueden considerarse hipótesis rivales plausibles considerando si una prueba mide menos o más que su constructo propuesto. Se hace referencia a dichas consideraciones como *infrarrepresentación de constructo* (o *deficiencia de constructo*) y *varianza irrelevante de constructo* (o *contaminación de constructo*), respectivamente.

La infrarrepresentación de constructo se refiere al grado en el cual una prueba no logra capturar aspectos importantes del constructo. Implica un significado acotado de los puntajes de la prueba porque la prueba no muestrea adecuadamente algunos tipos de contenido, no involucra algunos procesos psicológicos o no obtiene algunas maneras de responder que abarca el constructo previsto. Pensemos, por ejemplo, en una prueba que tiene por objeto ser una medida completa de ansiedad. Una prueba en particular podría infrarrepresentar el constructo previsto porque mide solo las reacciones psicológicas y no los componentes emocionales, cognitivos o situacionales. En otro ejemplo, una prueba de comprensión de

lectura que tiene por objeto medir la capacidad de los niños para leer e interpretar historias con comprensión podría no contener una variedad suficiente de pasajes de lectura o podría ignorar un tipo común de material de lectura.

La irrelevancia de constructo se refiere al grado en el cual los puntajes de la prueba se ven afectados por procesos que son externos al fin previsto de la prueba. Lo puntajes de la prueba pueden estar sistemáticamente influenciados en alguna medida por procesos que no son parte del constructo. En caso de una prueba de comprensión de lectura, estos podrían incluir material muy por encima o por debajo del nivel que se pretende evaluar, una reacción emocional al contenido de la prueba, familiaridad con el tema de los pasajes de lectura de la prueba, o la habilidad de escritura necesaria para elaborar una respuesta. Dependiendo de la definición detallada del constructo, el conocimiento de vocabulario o la velocidad de lectura también podrían ser componentes irrelevantes. En una prueba diseñada para medir la ansiedad, un sesgo de respuesta para reportar un nivel menor de la propia ansiedad podría considerarse una fuente de varianza irrelevante de constructo. En el caso de una prueba matemática, esto podría incluir una sobredependencia de las habilidades de comprensión de lectura que podría faltarles a los estudiantes de lengua inglesa. En una prueba diseñada para medir el conocimiento en ciencias, la internalización de los examinandos de estereotipos de género sobre las mujeres en las ciencias podría ser una fuente de varianza irrelevante del constructo.

Casi todas las pruebas dejan afuera elementos que algunos potenciales usuarios consideran que deberían medirse e incluyen algunos elementos que algunos potenciales usuarios consideran inapropiados. La validación involucra la atención minuciosa a posibles distorsiones en el significado que surgen de la representación inadecuada del constructo y también a aspectos de la medición, como el formato de la prueba, las condiciones de administración o el nivel de lenguaje, que pueden limitar o calificar significativamente la interpretación de los puntajes de la prueba para diversos grupos de examinandos. Es decir, el proceso de validación puede conducir a revisiones en la prueba, en el marco conceptual de la prueba, o en ambos. Las interpretaciones extraídas de la prueba revisada deberían volver a validarse.

Cuando se han identificado proposiciones que respaldarían la interpretación propuesta de los puntajes de la prueba, se puede continuar con la validación obteniendo evidencia empírica, estudiando bibliografía relevante y/o realizando análisis lógicos para evaluar cada una de las proposiciones. La evidencia empírica puede incluir tanto evidencia local, producida dentro de contextos donde se utilizará la prueba, como evidencia de aplicaciones de prueba similares en otros contextos. El uso de evidencia existente de pruebas y contextos similares puede mejorar la calidad del argumento de validez, en especial cuando los datos para la prueba y el contexto en cuestión son limitados.

Debido a que una interpretación para un uso dado suele depender de más de una proposición, la evidencia sólida en respaldo de una parte de la interpretación de ninguna manera reduce la necesidad de evidencia que respalde otras partes de la interpretación. Por ejemplo, cuando una prueba de empleo se considera para selección, una fuerte relación predictor-criterio en un contexto de empleo habitualmente no es suficiente para justificar el uso de la prueba. También se debería considerar lo apropiada y significativa que sea la medida del criterio, lo apropiados que sean los materiales y procedimientos de la prueba para la toda la variedad de solicitantes y la coherencia del respaldo para la interpretación propuesta entre los grupos. El juicio profesional orienta las decisiones respecto de formas específicas de evidencia que pueden respaldar mejor la interpretación prevista para el uso especificado. Como en todas las tareas científicas, la calidad de la evidencia es primordial. Algunas evidencias sólidas respecto de una proposición en particular son mejores que numerosas evidencias de calidad cuestionable. La determinación de que la interpretación de una prueba dada para un fin específico se justifica se basa en el juicio profesional de que la preponderancia de la evidencia disponible respalda esa interpretación. La calidad y cantidad de evidencia suficiente para alcanzar este juicio puede diferir para los usos

de la prueba según los riesgos involucrados en la prueba. Es posible que una interpretación dada no se justifique ya sea como resultado de evidencia insuficiente que la respalde o como resultado de evidencia creíble en contra de esta.

La validación es responsabilidad conjunta del desarrollador de la prueba y del usuario de la prueba. El desarrollador de la prueba es responsable de suministrar evidencia relevante y una razón fundamental que respalde cualquier interpretación de puntajes de la prueba para usos especificados previstos por el desarrollador. El usuario de la prueba es en última instancia responsable de evaluar la evidencia en el contexto en particular en el que se usará la prueba. Cuando el usuario de una prueba propone una interpretación o uso de puntajes de la prueba que difiere de los respaldados por el desarrollador de la prueba, la responsabilidad de brindar evidencias de validez que respalden esa interpretación para el uso especificado es del usuario. Debe observarse que pueden hacerse aportes importantes a la evidencia de validación a medida que otros investigadores reporten conclusiones de investigaciones que se relacionen con el significado de los puntajes en la prueba.

Fuentes de evidencia de validación

Las siguientes secciones describen diversas fuentes de evidencia que podrían utilizarse en la evaluación de la validez de una interpretación propuesta de puntajes de la prueba para un uso en particular. Estas fuentes de evidencia pueden iluminar diferentes aspectos de la validez, pero no representan tipos distintos de validez. La validez es un concepto unitario. Es el grado en que toda la evidencia acumulada respalda la interpretación prevista de los puntajes de una prueba para el uso propuesto. Al igual que los *Estándares* de 1999, esta edición hace referencia a los tipos de evidencia de validación, más que a tipos distintos de validez. Para destacar esta distinción, el tratamiento a continuación no sigue la nomenclatura histórica (es decir, el uso de los términos *validez de contenido* o *validez predictiva*).

Como se destaca en el análisis de la sección anterior, no se requiere cada tipo de evidencia presentado a continuación en todos los contextos. En lugar de ello, se necesita el respaldo de cada proposición subyacente a una interpretación de la prueba propuesta para un uso especificado. Una proposición de que una prueba es predictiva de un criterio dado puede respaldarse sin evidencia de que la prueba toma muestrea un domino de contenido en particular. Por el contrario, una proposición de que una prueba cubre una muestra representativa de un plan de estudios en particular puede ser respaldada sin evidencia de que la prueba predice un criterio dado. Sin embargo, un conjunto más complejo de proposiciones, p. ej., que una prueba abarque un dominio especificado y por lo tanto sea predictiva de un criterio que refleja un dominio relacionado, requerirá evidencia que respalde ambas partes de este conjunto de proposiciones. También se espera que los desarrolladores de la prueba demuestren que los puntajes no están indebidamente influenciados por varianza irrelevante de constructo (véase el cap. 3 para un tratamiento detallado de cuestiones relacionadas con varianza irrelevante de constructo). En general, el respaldo adecuado de las interpretaciones propuestas para usos específicos requerirá múltiples fuentes de evidencia.

La postura desarrollada anteriormente también subraya el hecho de que, si una prueba dada se interpreta de distintas maneras para distintos usos, también es probable que difieran las proposiciones que sustentan estas interpretaciones para diferentes usos. Se necesita el respaldo de las proposiciones que sustentan cada interpretación para un uso específico. La evidencia que respalda la interpretación de puntajes en una prueba de rendimiento matemático para colocar estudiantes en cursos subsiguientes (es decir, evidencia de que la interpretación de la prueba es válida para su fin previsto) no permite inferir validez para otros fines (p. ej., promoción o evaluación del profesor).

Evidencia basada en el contenido de la prueba

Se puede obtener evidencia de validación importante de un análisis de la relación entre el contenido de una prueba y el constructo que se intenta medir. El contenido de la prueba hace referencia a los temas, la redacción y el formato

de los ítems, tareas o preguntas de una prueba. La administración y el puntaje también pueden ser relevantes para la evidencia basada en el contenido. Los desarrolladores de la prueba suelen trabajar a partir de una especificación del dominio de contenido. La especificación del contenido describe cuidadosamente el contenido en detalle, a menudo con una clasificación de áreas de contenido y tipos de ítems. La evidencia basada en el contenido de la prueba puede incluir análisis lógicos o empíricos de la adecuación con la que el contenido de la prueba representa el dominio de contenido y de la relevancia del dominio de contenido para la interpretación propuesta de los puntajes de la prueba. La evidencia basada en el contenido también puede provenir de juicios expertos de la relación entre partes de la prueba y el constructo. Por ejemplo, en el desarrollo de una prueba para el otorgamiento de una licencia, pueden especificarse los principales aspectos que son relevantes para la finalidad para la cual se regula la ocupación, y se puede pedir a expertos en esa ocupación que asignen ítems de prueba a las categorías definidas por esas facetas. Estos u otros expertos pueden luego juzgar la representatividad del conjunto de ítems elegido.

Algunas pruebas se basan en observaciones sistemáticas del comportamiento. Por ejemplo, una lista de las tareas que constituyen un dominio de un puesto de trabajo puede desarrollarse a partir de observaciones del comportamiento en un puesto, junto con juicios de expertos en el tema. Los juicios expertos pueden utilizarse para evaluar la importancia relativa, criticidad y/o frecuencia de las diversas tareas. Una prueba de muestra de trabajo puede entonces construirse a partir de un muestreo aleatorio o estratificado de tareas calificadas altamente en estas características. La prueba luego puede administrarse en condiciones estandarizadas en un contexto fuera del trabajo.

Lo apropiado de un dominio de contenido dado se relaciona con las inferencias específicas que se harán de los puntajes de la prueba. Por consiguiente, al considerar una prueba disponible para un fin distinto del fin para el que se desarrolló en primer término, es especialmente importante evaluar lo adecuado del dominio de contenido original para el nuevo fin propuesto. Por ejemplo, una prueba dada para fines de investigación para comparar el rendimiento estudiantil en diferentes estados en un dominio dado puede correctamente también cubrir material que reciba atención escasa o nula en el plan de estudios. Los responsables de formular políticas pueden entonces evaluar el rendimiento estudiantil con respecto tanto al contenido ignorado como al contenido abordado. Por otra parte, cuando se evalúa el dominio estudiantil de un plan de estudios dictado a los fines de informar decisiones sobre estudiantes individuales, como promoción o graduación, el marco que elabora un dominio de contenido es adecuadamente limitado a lo que los estudiantes han tenido la oportunidad de aprender del plan de estudios según fuera dictado.

La evidencia sobre el contenido puede usarse, en parte, para abordar preguntas sobre diferencias en el significado o la interpretación de los puntajes de la prueba entre subgrupos relevantes de examinandos. Resulta de especial interés la medida en que la infrarrepresentación de constructo o la irrelevancia de constructo pueden dar una ventaja injusta o desventaja a uno o más subgrupos de examinandos. Por ejemplo, en una prueba de empleo, el uso de vocabulario más complejo que el necesario para el puesto de trabajo puede ser una fuente de varianza irrelevante de constructo para los estudiantes de lengua inglesa u otros. La revisión atenta del constructo y del dominio de contenido de la prueba por parte de un panel diverso de expertos puede señalar posibles fuentes de dificultad (o facilidad) irrelevante que requieren mayor investigación.

La evidencia de validación orientada al contenido se encuentra en el centro del proceso en el ámbito educativo conocido como *alineación*, que involucra evaluar la correspondencia entre estándares de aprendizaje para estudiantes y el contenido de la prueba. Las cuestiones de muestreo de contenido en el proceso de alineación incluyen evaluar si el contenido de la prueba muestrea adecuadamente el domino propuesto en los estándares del plan de estudios, si las demandas cognitivas de los ítems de la prueba se corresponden con el nivel reflejado en los estándares de aprendizaje de

los estudiantes (p. ej., estándares de contenido) y si la prueba evita la inclusión de características irrelevantes para el estándar que es el objetivo previsto de cada ítem de la prueba.

Evidencia basada en los procesos de respuesta

Algunas interpretaciones de constructos involucran suposiciones más o menos explícitas sobre los procesos cognitivos empleados por los examinandos. Análisis teóricos y empíricos de los procesos de respuesta de los examinandos pueden proporcionar evidencia respecto de la adecuación entre el constructo y la naturaleza detallada del desempeño o respuesta efectivamente empleada por los examinandos. Por ejemplo, si una prueba tiene por objeto evaluar el razonamiento matemático, es importante determinar si los examinandos están, en realidad, razonando sobre el material dado en lugar de seguir un algoritmo estándar aplicable solo a los ítems específicos en la prueba.

La evidencia basada en los procesos de respuesta por lo general proviene de análisis de respuestas individuales. Preguntar a los examinandos de diversos grupos que componen la población examinada prevista sobre sus estrategias de desempeño o repuestas a ítems en particular puede arrojar evidencia que enriquezca la definición de un constructo. Mantener registros que monitoreen el desarrollo de una respuesta a una tarea de escritura, mediante borradores escritos sucesivos o revisiones monitoreadas electrónicamente, por ejemplo, también proporciona evidencia del proceso. La documentación de otros aspectos del desempeño, como los movimientos de los ojos o rapidez al responder, también puede ser relevante para algunos constructos. Las inferencias sobre procesos involucrados en el desempeño también pueden desarrollarse analizando la relación entre pates de la prueba y entre la prueba y otras variables. Grandes diferencias individuales pueden ser reveladoras y pueden llevar a la reconsideración de ciertos formatos de prueba.

La evidencia de los procesos de respuesta puede contribuir a responder preguntas sobre diferencias en el significado o interpretación de puntajes de pruebas entre subgrupos relevantes de examinandos. Los estudios de proceso en los que participan examinandos de diferentes subgrupos pueden ayudar a determinar en qué medida las capacidades irrelevantes o auxiliares al constructo pueden influir de manera diferencial en el desempeño de los examinandos en la prueba.

Los estudios de procesos de respuesta no se limitan al examinando. Las evaluaciones suelen depender de observadores o jueces para que registren y/o evalúen los desempeños o productos de los examinandos. En esos casos, la evidencia de validación relevante incluye la medida en que los procesos de observadores o jueces son coherentes con la interpretación prevista de puntajes. Por ejemplo, si se espera que los jueces apliquen criterios particulares al calificar los desempeños de los examinandos, es importante determinar si están, de hecho, aplicando los criterios apropiados y no siendo influenciados por factores que son irrelevantes para la interpretación prevista (p. ej., la calidad de la caligrafía es irrelevante para juzgar el contenido de un ensayo escrito). Por lo tanto, la validación puede incluir estudios empíricos de cómo los observadores o jueces registran y evalúan datos junto con análisis de lo adecuado que son estos procesos para la interpretación prevista o la definición del constructo.

Si bien la evidencia sobre los procesos de respuesta puede ser central en contextos en los que las afirmaciones explícitas sobre procesos de respuesta son hechas por desarrolladores de la prueba o en los que las inferencias sobre respuestas son hechas por usuarios de la prueba, hay muchos otros casos en los que las afirmaciones sobre los procesos de respuesta no son parte del argumento de validez. En algunos casos, múltiples procesos de respuesta están disponibles para resolver los problemas de interés, y el constructo de interés solo tiene que ver con que el problema se resuelva de manera correcta. Para dar un ejemplo simple, puede haber múltiples caminos posibles para obtener la solución correcta a un problema matemático.

Evidencia basada en la estructura interna

Los análisis de la estructura interna de una prueba pueden indicar el grado en que las relaciones entre ítems de la prueba y componentes de la prueba

se ajustan al constructo sobre el que se basan las interpretaciones propuestas de puntajes de la prueba. El marco conceptual para una prueba puede implicar una sola dimensión de comportamiento, o puede plantear varios componentes; se espera que cada uno de ellos sea homogéneo, pero también son distintos unos de otros. Por ejemplo, una medida de malestar en una encuesta de salud podría evaluar tanto la salud física como emocional. La medida en que las interrelaciones entre ítems confirman las presunciones del marco sería relevante para la validez.

Los tipos específicos de análisis y su interpretación dependen de cómo se utilizará la prueba. Por ejemplo, si una aplicación en particular planteó una serie de componentes de la prueba cada vez más difíciles, se proporcionaría evidencia empírica de la medida en que los patrones de respuesta cumplieron con esta expectativa. Una teoría que planteara la unidimensionalidad requeriría evidencia de homogeneidad de ítems. En este caso, la cantidad de ítems y las interrelaciones entre ítems forman la base para una estimación de confiabilidad del puntaje, pero un índice de este tipo sería inadecuado para pruebas con una estructura interna más compleja.

Algunos estudios de la estructura interna de las pruebas se han diseñado para mostrar si ítems en particular pueden funcionar de manera diferente para subgrupos de examinados identificables (p. ej., subgrupos raciales/étnicos o de género). Se produce un *funcionamiento diferencial de los ítems* cuando diferentes grupos de examinandos con capacidad general similar, o nivel similar en un criterio adecuado, tienen, en promedio, respuestas sistemáticamente diferentes a un ítem en particular. Esta cuestión se analiza en el capítulo 3. Sin embargo, el funcionamiento diferencial de los ítems no siempre es una falla o debilidad. Subconjuntos de ítems que tienen una característica específica en común (p. ej., contenido específico, representación de tarea) pueden funcionar de manera diferente para diferentes grupos de examinandos con puntajes similares. Esto indica una clase de multidimensionalidad que puede esperarse o puede ajustarse al marco de la prueba.

Evidencia basada en relaciones con otras variables

En muchos casos, la interpretación prevista para un uso dado implica que el constructo debería relacionarse con algunas otras variables y, como resultado, análisis de la relación de los puntajes de la prueba con variables externas a la prueba proporcionan otra fuente importante de evidencia de validación. Las variables externas pueden incluir medidas de algunos criterios que se espera que la prueba prediga, así como relaciones con otras pruebas propuestas para medir los mismos constructos, y pruebas que miden constructos relacionados o diferentes. Las medidas distintas de los puntajes de la prueba, tal como criterios de desempeño, suelen utilizarse en contextos laborales. Las variables categóricas, incluidas variables de membrecía de grupos, se vuelven relevantes cuando la teoría que sustenta un uso propuesto de la prueba sugiere que las diferencias del grupo deberían estar presentes o ausentes si una interpretación propuesta de los puntajes de la prueba debe sustentarse. La evidencia basada en las relaciones con otras variables proporciona evidencia sobre el grado en que estas relaciones son coherentes con el constructo que sustenta las interpretaciones propuestas de los puntajes de la prueba.

Evidencia convergente y discriminante. Las relaciones entre los puntajes de la prueba y otras medidas que tienen por objeto evaluar los mismos constructos o similares proporcionan evidencia convergente, mientras que las relaciones entre los puntajes de la prueba y medidas supuestamente de constructos diferentes proporcionan evidencia discriminante. Por ejemplo, dentro de algunos marcos teóricos, podría esperarse que los puntajes en una prueba de selección múltiple de comprensión de lectura se relacionen estrechamente (evidencia convergente) con otras medidas de comprensión de lectura basadas en otros métodos, como las respuestas a ensayos. Al contrario, podría esperarse que los puntajes de la prueba se relacionen menos estrechamente (evidencia discriminante) con medidas de otras habilidades, como el razonamiento lógico. Las relaciones entre

diferentes métodos de medición del constructo pueden ser especialmente útiles para refinar y elaborar el significado y la interpretación del puntaje.

La evidencia de relaciones con otras variables puede involucrar evidencia experimental como correlacional. Podrían diseñarse estudios, por ejemplo, para investigar si los puntajes en una medida de ansiedad mejoran como resultado de algún tratamiento psicológico o si los puntajes en una prueba de rendimiento académico diferencian entre grupos con instrucción y sin instrucción. Si los aumentos del desempeño debido a orientación a corto plazo se ven como una amenaza para la validez, sería útil investigar si los grupos con orientación y sin orientación tienen desempeños diferentes.

Relaciones prueba-criterio. La evidencia de la relación de puntajes de la prueba con un criterio relevante puede expresarse de distintas maneras, pero la pregunta fundamental siempre es ¿con qué exactitud los puntajes de la prueba predicen el desempeño del criterio? El grado de exactitud y el rango de puntajes dentro del que se necesita exactitud dependen del fin para el que se utilice la prueba.

La variable del criterio es una medida de algún atributo o resultado que es operativamente distinto de la prueba. Por lo tanto, la prueba no es una medida de un criterio, sino una medida planteada como un potencial predictor de ese criterio de interés. Si una prueba predice un criterio dado en un contexto dado, es una hipótesis comprobable. Los criterios que son de interés son determinados por los usuarios de la prueba, por ejemplo, administradores en un sistema escolar o gerentes de una empresa. La elección del criterio y los procedimientos de medición utilizados para obtener puntajes de criterios son de primordial importancia. La credibilidad del estudio prueba-criterio depende de la relevancia, confiabilidad y validez de la interpretación basada en la medida del criterio para una aplicación de prueba dada.

Históricamente, se han distinguido dos diseños, a menudo llamados predictivo y concurrente, para evaluar las relaciones prueba-criterio. Un estudio predictivo indica la fortaleza de la relación entre los puntajes de la prueba y los puntajes de criterios que se obtienen en un momento posterior. Un estudio concurrente obtiene puntajes de la prueba e información del criterio aproximadamente al mismo tiempo. Cuando efectivamente se contempla la predicción, como en la admisión académica o los contextos laborales, o en la planificación de programas de rehabilitación, los estudios predictivos pueden conservar las diferencias temporales y otras características de la situación práctica. La evidencia concurrente, que evita cambios temporales, es particularmente útil para pruebas de psicodiagnóstico o en la investigación de medidas alternas de algún constructo especificado para el que ya existe un procedimiento de medición aceptado. La elección de una estrategia de investigación predictiva o concurrente en un dominio dado es también provechosamente informada por evidencia de investigaciones previas respecto de la medida en que los estudios predictivos y concurrentes en ese dominio arrojan los mismos o diferentes resultados.

Los puntajes de la prueba a veces se usan para asignar a individuos a diferentes tratamientos de una manera que sea ventajosa para la institución y/o para los individuos. Entre los ejemplos se incluirían asignar a individuos a diferentes puestos en una organización, o determinar si colocar a un estudiante dado en una clase de apoyo o una clase regular. En ese contexto, se necesita evidencia para juzgar la pertinencia de utilizar una prueba cuando se clasifica o asigna a una persona a un puesto en vez de otro o a un tratamiento en vez de otro. El respaldo de la validez del procedimiento de clasificación se proporciona mostrando que la prueba es útil para determinar qué personas probablemente se beneficien de manera diferente con un tratamiento u otro. Es posible que las pruebas sean sumamente predictivas del desempeño para diferentes programas educativos o puestos sin proporcionar la información necesaria para hacer un juicio comparativo de la eficacia de las asignaciones o tratamientos. En general, las normas de decisión para la selección o asignación también están influenciadas por la cantidad de personas que se aceptarán o las cantidades que pueden admitirse en categorías de asignación alternativas (véase el cap. 11).

También se usa la evidencia sobre relaciones con otras variables para investigar preguntas de predicción diferencial entre subgrupos. Por ejemplo, una conclusión de que la relación de los puntajes de la prueba con una variable de criterio relevante difiere entre subgrupo y otro puede implicar que el significado de los puntajes no es el mismo para miembros de los diferentes grupos, tal vez debido a infrarrepresentación de constructo o fuentes de varianza irrelevante de constructo. Sin embargo, la diferencia también puede implicar que el criterio tiene diferente significado para diferentes grupos. Las diferencias en las relaciones prueba-criterio también pueden surgir de un error de medición, en especial cuando las medias de los grupos difieren, de modo que dichas diferencias no necesariamente indican diferencias en el significado de los puntajes. Véase el análisis de imparcialidad en el capítulo 3 para una consideración más amplia de posibles cursos de acción cuando los puntajes tienen diferentes significados para diferentes grupos.

Generalización de validez. Una cuestión importante en los contextos educativos y laborales es el grado en que la evidencia de validación basada en relaciones prueba-criterio puede generalizarse a una nueva situación sin estudios adicionales de validez en esa nueva situación. Cuando una prueba se usa para predecir los mismos criterios o criterios similares (p. ej., desempeño de un determinado puesto) en momentos diferentes o en lugares diferentes, suele determinarse que las correlaciones prueba-criterio observadas varían sustancialmente. En el pasado, se ha considerado que esto implica que siempre se requieren estudios de validación locales. Más recientemente, se han desarrollado varios enfoques sobre la generalización de evidencia de otros contextos, siendo el metaanálisis el más utilizado en la bibliografía publicada. En particular, los metaanálisis han demostrado que, en algunos dominios, mucha de esta variabilidad puede deberse a artefactos estadísticos como fluctuaciones en el muestreo y variaciones entre estudios de validación en los rangos de los puntajes de las pruebas y en la confiabilidad de las medidas de los criterios. Cuando se tienen en cuenta estas y otras influencias, es posible que se determine que la variabilidad restante en los coeficientes de validez es relativamente pequeña. Por lo tanto, es posible que sean útiles los resúmenes estadísticos de estudios de validación anteriores en la estimación de las relaciones prueba-criterio en una nueva situación. Esta práctica se denomina estudio de generalización de validez.

En algunas circunstancias, existe un fundamento sólido para utilizar la generalización de validez. Este sería el caso cuando la base de datos metaanalítica es amplia, cuando los datos metaanalíticos representan adecuadamente el tipo de situación a la que se desea generalizar y cuando la corrección para artefactos estadísticos produce un patrón claro y coherente de evidencia de validación. En esas circunstancias, el valor informativo de un estudio de validez local puede ser relativamente limitado, si no efectivamente confuso, en especial si el tamaño de su muestra es pequeño. En otras circunstancias, el salto inferencial requerido para la generalización sería mucho más grande. La base de datos metaanalítica puede ser pequeña, las conclusiones pueden ser menos coherentes o la nueva situación puede involucrar características marcadamente diferentes de las representadas en la base de datos metaanalítica. En esas circunstancias, la evidencia de validación específica de la situación será relativamente más informativa. Si bien la investigación sobre la generalización de validez muestra que los resultados de un solo estudio de validación local pueden ser bastante imprecisos, hay situaciones en las que un solo estudio, realizado cuidadosamente, con un tamaño de muestra adecuado, proporciona suficiente evidencia para respaldar o rechazar el uso de la prueba en una nueva situación. Esto destaca la importancia de examinar atentamente el valor informativo comparativo de los estudios acotados frente a los metaanalíticos.

Cuando se llevan a cabo estudios de la generabilidad de evidencia de validación, los estudios anteriores que se incluyen pueden variar de acuerdo con varios aspectos situacionales. Algunas de las principales facetas son (a) diferencias en la manera en que se mide el constructo predictor, (b) el tipo

de puesto de trabajo o plan de estudio involucrado, (c) el tipo de medida de criterio utilizado, (d) el tipo de examinandos, y (e) el período en el que se realizó el estudio. En cualquier estudio de generalización de validez, cualquier cantidad de estas facetas podría variar, y un objetivo principal del estudio es determinar empíricamente la medida en que la variación en estas facetas afecta las correlaciones prueba-criterio obtenidas.

La medida en que la evidencia de validación predictiva o concurrente puede generalizarse a nuevas situaciones es en gran medida una función de investigación acumulada. Si bien la evidencia de generalización a menudo puede ayudar a sustentar una afirmación de validez en una nueva situación, el alcance de datos disponibles limita el grado en que puede sustentarse la afirmación.

La discusión anterior se concentra en el uso de bases de datos acumulativas para estimar relaciones predictor-criterio. Las técnicas metaanalíticas también pueden usarse para resumir otras formas de datos relevantes a otras inferencias que se pueden querer extraer de los puntajes de la prueba en una aplicación en particular, como los efectos de la orientación y los efectos de determinadas alteraciones en las condiciones de la prueba para examinandos con discapacidades especificadas. Reunir evidencia sobre en qué medida las conclusiones de validez pueden generalizarse entre grupos de examinandos es una parte importante del proceso de validación. Cuando la evidencia sugiere que pueden hacerse inferencias a partir de puntajes de la prueba para algunos subgrupos, pero no para otros, intentar opciones como las analizadas en el capítulo 3 puede reducir el riesgo de uso parcial de la prueba.

Evidencia de validación y consecuencias de las pruebas

Algunas consecuencias del uso de pruebas surgen directamente de la interpretación de los puntajes de la prueba para usos previstos por el desarrollador de la prueba. El proceso de validación implica reunir evidencia para evaluar la solidez de estas interpretaciones propuestas para sus usos previstos.

Otras consecuencias también pueden ser parte de una afirmación que se extiende más allá de la interpretación o el uso de puntajes previsto por el desarrollador de la prueba. Por ejemplo, una prueba de rendimiento estudiantil podría proporcionar datos para un sistema cuyo objeto sea identificar y mejorar las escuelas con bajo rendimiento. La afirmación de que los resultados de las pruebas, utilizados de esta manera, darán por resultado una mejora en el aprendizaje estudiantil puede depender de proposiciones sobre el sistema o la intervención propiamente dicha, más allá de las proposiciones basadas en el significado de la prueba misma. Las consecuencias pueden señalar la necesidad de evidencia sobre componentes del sistema que irán más allá de la interpretación de los puntajes de la prueba como una medida válida del rendimiento estudiantil.

Aun así, otras consecuencias son imprevistas, y a menudo negativas. Por ejemplo, las pruebas educativas a nivel estatal o de distrito escolar sobre asignaturas seleccionadas pueden llevar a los profesores a concentrarse en esas asignaturas a expensas de otras. Para citar otro ejemplo, una prueba desarrollada para medir el conocimiento necesario para un determinado puesto de trabajo puede dar lugar a tasas de aprobación más bajas para un grupo que para otro. Las consecuencias imprevistas merecen un examen detenido. Si bien no todas las consecuencias pueden preverse, en algunos casos los factores como experiencias previas en otros contextos ofrecen una base para prever y abordar de manera proactiva las consecuencias imprevistas. Véase el capítulo 12 para consultar ejemplos adicionales de contextos educativos. En algunos casos, las acciones para abordar una consecuencia dan lugar a otras consecuencias. Un ejemplo involucra la noción de "oportunidades perdidas", como en el caso de pasar a calificación por computadora de los ensayos de estudiantes para aumentar la coherencia en las calificaciones, con lo cual se renuncia a los beneficios educativos de abordar el mismo problema capacitando a los profesores para calificar de manera más coherente.

Estos tipos de consideración de consecuencias de las pruebas se analizan más adelante.

Interpretación y usos de puntajes de la prueba previstos por los desarrolladores de la prueba. Las pruebas por lo general se administran con la

expectativa de que se concentrará algún beneficio a partir de la interpretación y el uso de los puntajes previstos por los desarrolladores de la prueba. Algunos de los muchos beneficios posibles que podrían citarse son la selección de terapias eficaces, asignación de trabajadores en puestos adecuados, prevenir que individuos no calificados ingresen en una profesión, o mejora de las prácticas de instrucción en el aula. Una finalidad fundamental de la validación es indicar si es probable que estos beneficios específicos se concreten. Por lo tanto, en el caso de una prueba utilizada en decisiones sobre colocación, la validación sería informada por evidencia de que colocaciones alternativas, de hecho, son beneficiosas de manera diferencial para las personas y la institución. En el caso de pruebas de empleo, si el editor de una prueba asevera que el uso de la prueba dará por resultado una reducción de los costos de capacitación de empleados, mejora de la eficiencia de la fuerza de trabajo o algún otro beneficio, entonces la validación sería informada por evidencia que sustente esa proposición.

Es importante destacar que la validez de las interpretaciones de los puntajes de las pruebas depende no solo de los usos de los puntajes de las pruebas sino específicamente de las afirmaciones que sustentan la teoría de acción para estos usos. Por ejemplo, consideremos un distrito escolar que quiere determinar la preparación de los niños para el jardín de infancia, y entonces administra una batería de pruebas y descarta a los estudiantes con puntajes bajos. Si los puntajes más altos, efectivamente, predicen un desempeño más alto en tareas clave del jardín de infancia, la afirmación de que el uso de los puntajes de la prueba para seleccionar resultados en desempeño más alto en estas tareas clave está respaldada y la interpretación de los puntajes de las pruebas como un predictor de preparación para el jardín de infancia sería válido. Sin embargo, si se hiciera la afirmación de que el uso de los puntajes de las pruebas para la selección daría por resultado el mayor beneficio para los estudiantes, la interpretación de los puntajes de las pruebas como indicadores de preparación para el jardín de infancia no podría ser válida porque los estudiantes con puntajes bajos podrían efectivamente beneficiarse más con el acceso al jardín de infancia. En este caso, se necesita evidencia diferente para respaldar diferentes afirmaciones que podrían hacerse sobre el mismo uso de la prueba de selección (por ejemplo, evidencia de que los estudiantes por debajo de un determinado puntaje de corte se beneficiarían más con otra asignación que con la asignación al jardín de infancia). El desarrollador de la prueba es responsable de la validación de la interpretación de que los puntajes de la prueba determinan las habilidades de preparación indicadas. El distrito escolar es responsable de la validación de la interpretación adecuada de los puntajes de la prueba de preparación y de la evaluación de la política de usar la prueba de preparación para las decisiones de colocación/admisión.

Afirmaciones hechas sobre el uso de la prueba que no se basan directamente en interpretaciones de los puntajes de la prueba. A veces se hacen afirmaciones sobre los beneficios de las pruebas que van más allá de las interpretaciones directas o usos de los puntajes de la prueba propiamente dichos que son especificados por los desarrolladores de la prueba. Las pruebas educativas, por ejemplo, pueden defenderse con el fundamento de que su uso mejorará la motivación de los estudiantes para aprender o fomentará cambios en las prácticas de instrucción en el aula al responsabilizar a los educadores de resultados de aprendizaje valorados. Cuando esas afirmaciones son centrales para la razón fundamental adelantada para las pruebas, el examen directo de las consecuencias de la prueba necesariamente cobra aún más importancia. Quienes hacen esas afirmaciones son responsables de la evaluación de las afirmaciones. En algunos casos, esa información puede obtenerse de datos existentes reunidos para fines distintos de la validación de la prueba; en otros casos se necesitará nueva información para abordar el impacto del programa de pruebas.

Consecuencias que son imprevistas. La interpretación de los puntajes de la prueba para un uso dado puede dar por resultado consecuencias imprevistas. Una distinción clave es entre

consecuencias que surgen de una fuente de error en la interpretación prevista de los puntajes de la prueba para un uso dado y las consecuencias que no resultan de un error en la interpretación de los puntajes de la prueba. A continuación, se dan ejemplos de cada una.

Como se analiza con cierta extensión en el capítulo 3, un dominio en el que a veces se observan consecuencias negativas imprevistas del uso de las pruebas involucra diferencias de puntajes de la prueba para grupos definidos en términos de raza/origen étnico, género, edad y otras características. En esos casos, no obstante, es importante distinguir entre evidencia que es directamente relevante para la validez y evidencia que puede informar decisiones sobre política social, pero queda fuera del terreno de la validez. Por ejemplo, se han planteado inquietudes sobre el efecto de las diferencias de grupos en los puntajes de las pruebas en la selección y promoción laborales, la colocación de niños en clases de educación especial y el acotamiento del plan de estudios de la escuela para excluir objetivos de aprendizaje que no se evalúan. Si bien la información sobre las consecuencias de las pruebas puede influir en las decisiones sobre el uso de la prueba, esas consecuencias, de por sí, no le restan valor a la validez de las interpretaciones previstas de los puntajes de la prueba. En cambio, los juicios de validez o falta de esta a la luz de las consecuencias de las pruebas dependen de una investigación más minuciosa de las fuentes de esas consecuencias.

Por ejemplo, una conclusión de diferentes tasas de contratación para miembros de diferentes grupos como una consecuencia de utilizar una prueba de empleo. Si la diferencia se debe exclusivamente a una distribución desigual de las habilidades que la prueba pretende medir, y si esas habilidades son, de hecho, factores de contribución importantes para el desempeño laboral, entonces encontrar diferencias entre los grupos de por sí no implica ninguna falta de validez para la interpretación prevista. Sin embargo, si la prueba midiera diferencias de habilidades no relacionadas con el desempeño laboral (p. ej., una prueba de lectura sofisticada para un puesto de trabajo que requería solo alfabetización funcional mínima), o si las diferencias se debieran a la sensibilidad de la prueba ante alguna característica del examinando que no tenía por objeto ser parte del constructo de la prueba, entonces la interpretación prevista de los puntajes de la prueba como predictores del desempeño laboral en una manera comparable para todos los grupos de solicitantes se consideraría inválida, incluso si los puntajes de la prueba se correlacionaran positivamente con alguna medida de desempeño laboral. Si una prueba cubre la mayoría del dominio de contenido relevante, pero omite algunas áreas, la cobertura de contenido podría considerarse inadecuada para algunos fines. Sin embargo, si se determina que excluir algunos componentes que podrían evaluarse de inmediato tiene un impacto notable en las tasas de selección para grupos de interés (p. ej., se determina que las diferencias entre subgrupos son menores en componentes excluidos que en componentes incluidos), la interpretación prevista de los puntajes de la prueba como predictores del desempeño laboral en una manera comparable para todos los grupos de solicitantes se consideraría inválida. Por lo tanto, la evidencia sobre consecuencias es relevante para la validez cuando puede trazarse hacia una fuente de invalidez como la infrarrepresentación de constructo o componentes irrelevantes de constructo. La evidencia sobre consecuencias que no puede trazarse así no es relevante para la validez de las interpretaciones previstas de los puntajes de la prueba.

En otro ejemplo, consideremos el caso en el que la investigación respalda el uso por parte de un empleador de una prueba en particular en el dominio de la personalidad (es decir, la prueba demuestra que es predictiva de un aspecto del posterior desempeño laboral), pero se determina que algunos solicitantes se forman una opinión negativa de la organización debido a la percepción de que la prueba invade la privacidad personal. Por lo tanto, hay una consecuencia negativa imprevista del uso de la prueba, pero que no se debe a un defecto en la interpretación prevista de los puntajes de la prueba como predictor del desempeño posterior. Ante esta situación, algunos empleadores pueden concluir que esta consecuencia negativa es un motivo para discontinuar el uso de

la prueba; otros pueden concluir que los beneficios obtenidos al seleccionar a solicitantes superan esta consecuencia negativa. Como muestra este ejemplo, una consideración de consecuencias puede influir en una decisión sobre el uso de la prueba, aunque la consecuencia sea independiente de la validez de la interpretación prevista de los puntajes de la prueba. El ejemplo también muestra que diferentes responsables de tomar decisiones pueden hacer diferentes juicios de valor sobre el impacto de las consecuencias en el uso de la prueba.

El hecho de que la evidencia de validación respalde la interpretación prevista de los puntajes de la prueba para usar en la selección de solicitantes no significa que entonces se requiera el uso de la prueba: Cuestiones distintas de la validez, incluyendo restricciones legales, pueden tener un papel importante y, en algunos casos, determinante en las decisiones sobre el uso de la prueba. Las restricciones legales también pueden limitar la discreción de un empleador para descartar puntajes de la prueba que ya se han administrado, cuando esa decisión se basa en diferencias en los puntajes para subgrupos de diferentes razas, orígenes étnicos o géneros.

Téngase en cuenta que las consecuencias imprevistas también pueden ser positivas. Si se invierte el ejemplo anterior de examinandos que se forman una impresión negativa de una organización sobre la base del uso de una prueba en particular, una prueba diferente puede ser percibida favorablemente por los solicitantes, lo que lleva a una impresión positiva de la organización. Un uso determinado de una prueba puede dar por resultado múltiples consecuencias, algunas positivas y algunas negativas.

En resumen, las decisiones sobre el uso de la prueba son adecuadamente informadas por la evidencia de validación sobre las interpretaciones previstas de los puntajes de la prueba para un uso dado, por evidencia que evalúa afirmaciones adicionales sobre consecuencias del uso de la prueba que no surgen directamente de interpretaciones de los puntajes de la prueba y por juicios de valor sobre consecuencias positivas y negativas imprevistas del uso de la prueba.

Integración de la evidencia de validación

Un argumento de validez sólido integra diversos aspectos de la evidencia en una explicación coherente del grado en que la evidencia existente y la teoría respaldan la interpretación prevista de los puntajes de la prueba para usos específicos. Abarca evidencia reunida a partir de nuevos estudios y evidencia disponible de investigación anterior. El argumento de validez puede indicar la necesidad de refinar la definición del constructo, puede sugerir revisiones en la prueba u otros aspectos del proceso de desarrollo de la prueba, y puede indicar áreas que necesitan mayor investigación.

Comúnmente se observa que el proceso de validación nunca termina, dado que siempre hay información adicional que puede reunirse para comprender más cabalmente una prueba y las inferencias que pueden extraerse de esta. En este sentido, una inferencia de validez es similar a cualquier inferencia científica. Sin embargo, la interpretación de una prueba para un uso dado se basa en evidencia para un conjunto de proposiciones que conforman el argumento de validez, y en algún momento la evidencia de validación permite un juicio breve de la interpretación prevista que está bien respaldada y puede defenderse. En algún momento, el esfuerzo de proporcionar suficiente evidencia de validación para respaldar una interpretación de una prueba dada para un uso específico termina (al menos provisionalmente, a la espera de que surja un fundamento sólido para cuestionar ese juicio). Los requisitos legales pueden exigir que el estudio de validación se actualice a la luz de factores como cambios en la población de la prueba o métodos de prueba alternativos recientemente desarrollados.

La cantidad y el carácter de la evidencia requerida para respaldar un juicio provisional de validez suele variar entre áreas y también dentro un área a medida que avanza la investigación sobre un tema. Por ejemplo, los estándares predominantes de evidencia pueden variar con los riesgos involucrados en el uso o interpretación de los puntajes de la prueba. Los riesgos más elevados pueden conllevar estándares más elevados

de evidencia. Para dar otro ejemplo, en áreas en que la recopilación de datos tiene un costo más alto, podría ser necesario basar las interpretaciones en menor cantidad de datos que en áreas en que la recopilación de datos tiene un costo menor.

En última instancia, la validez de una interpretación prevista de los puntajes de la prueba se basa en toda la evidencia disponible relevante para la calidad técnica de un sistema de prueba. Diferentes componentes de la evidencia de validación se describen en capítulos siguientes de los *Estándares*, e incluyen evidencia de la construcción cuidadosa de la prueba; confiabilidad adecuada de los puntajes; administración y calificación adecuadas de la prueba; precisión en el establecimiento de escala de puntajes, equiparación, y fijación de estándares; y atención cuidadosa a la imparcialidad para todos los examinandos, según corresponda a la interpretación de la prueba en cuestión.

ESTÁNDARES DE VALIDEZ

Los estándares en este capítulo comienzan con un estándar global (numerado 1.0), que se ha diseñado para transmitir la intención central o enfoque principal del capítulo. El estándar global también puede verse como el principio rector del capítulo, y es aplicable a todas las pruebas y usuarios de pruebas. Todos los estándares posteriores se han separado en tres unidades temáticas denominadas de la siguiente manera:

1. Establecimiento de usos e interpretaciones previstos
2. Cuestiones respecto de las muestras y contextos utilizados en la validación
3. Formas específicas de evidencia de validación

Estándar 1.0

Debe establecerse la articulación clara de cada interpretación prevista de los puntajes de la prueba para un uso especificado, y debe proporcionarse evidencia de validación apropiada que respalde cada interpretación prevista.

Unidad 1. Establecimiento de usos e interpretaciones previstos

Estándar 1.1

El desarrollador de la prueba debe establecer claramente cómo se tiene previsto que se interpreten y en consecuencia se utilicen los puntajes de la prueba. Las poblaciones para las que está prevista la prueba deben definirse claramente, y el constructo o los constructos que la prueba tiene por objeto evaluar deben describirse claramente.

Comentario: Los enunciados sobre validez deben referirse a interpretaciones particulares y usos consecuentes. Es incorrecto usar la frase no calificada "la validez de la prueba". Ninguna prueba permite interpretaciones que sean válidas para todos los fines o en todas las situaciones. Cada interpretación recomendada para un uso dado requiere validación. El desarrollador de la prueba debe especificar en lenguaje claro la población para la que está prevista la prueba, el constructo que tiene previsto medir, los contextos en los que se emplearán los puntajes de la prueba y los procesos mediante los que la prueba se administrará y calificará.

Estándar 1.2

Se debe presentar una razón fundamental para cada interpretación prevista de los puntajes de la prueba para un uso determinado, junto con un resumen de la evidencia y la teoría que inciden en la interpretación prevista.

Comentario: La razón fundamental debe indicar qué proposiciones son necesarias para investigar la interpretación prevista. El resumen debe combinar análisis lógico con evidencia empírica para respaldar la razón fundamental de la prueba. La evidencia puede proceder de estudios realizados a nivel local, en el contexto en el que se usará la prueba; de estudios previos específicos; o de síntesis estadísticas completas de estudios disponibles que reúnan claramente los criterios de calidad del estudio especificado. Ningún tipo de evidencia es intrínsecamente preferible a otros, sino que la calidad y relevancia de la evidencia para la interpretación prevista de los puntajes de la prueba para un uso dado determinan el valor de una clase de evidencia en particular. Una presentación de evidencia empírica en cualquier momento debe dar la debida importancia a todas las conclusiones relevantes en la bibliografía científica, incluidas las que no son coherentes con la interpretación o uso previstos. Los desarrolladores de la prueba tienen la responsabilidad de respaldar sus propias recomendaciones, pero los usuarios de la prueba tienen la responsabilidad máxima de evaluar la calidad de la evidencia de validación proporcionada y su relevancia para la situación local.

Estándar 1.3

Si la validez para alguna interpretación común o probable para un uso dado no se ha evaluado, o

si dicha interpretación no es coherente con la evidencia disponible, ese hecho debe aclararse y se debe advertir enfáticamente a los posibles usuarios sobre hacer interpretaciones sin fundamento.

Comentario: Si la experiencia pasada sugiere que es probable que una prueba se use de manera inadecuada para determinadas clases de decisiones o determinadas clases de examinandos, se deben hacer advertencias específicas contra dichos usos. Se requiere juicio profesional para evaluar la medida en que la evidencia de validación existente respalda un uso determinado de la prueba.

Estándar 1.4

Si el puntaje de una prueba se interpreta para un uso determinado de una manera que no ha sido validada, corresponde al usuario justificar la nueva interpretación para ese uso, proporcionando una razón fundamental y reuniendo nueva evidencia, si fuera necesario.

Comentario: Se requiere juicio profesional para evaluar la medida en que la evidencia de validación existente se aplica en la nueva situación y al nuevo grupo de examinandos y para determinar qué nueva evidencia puede ser necesaria. La cantidad y las clases de nueva evidencia requeridas pueden estar influenciadas por experiencia con usos o interpretaciones de pruebas anteriores similares o por la cantidad, calidad y relevancia de datos existentes.

Una prueba que ha sido alterada o administrada de maneras que cambian el constructo subyacente a la prueba para uso con subgrupos de la población requiere evidencia de la validez de la interpretación hecha sobre la base de la prueba modificada (véase el cap. 3). Por ejemplo, si una prueba se adapta para usarse con individuos con una discapacidad en particular de una manera que cambia el constructo subyacente, la prueba modificada debe tener su propia evidencia de validación para la interpretación prevista.

Estándar 1.5

Cuando se indica claramente o se deja implícito que una interpretación recomendada de los puntajes de la prueba para un determinado uso dará un resultado específico, se debe presentar el fundamento para prever ese resultado, junto con la evidencia relevante.

Comentario: Si se asevera, por ejemplo, que interpretar y usar puntajes en una prueba dada para la selección de empleados dará por resultado la reducción de errores de los empleados o de costos de capacitación, debe proporcionarse evidencia que respalde esa aseveración. Una afirmación dada puede ser respaldada por un argumento lógico o teórico, así como también por datos empíricos. Debe darse la debida importancia a las conclusiones en la bibliografía científica que pueden no ser coherentes con la expectativa indicada.

Estándar 1.6

Cuando el uso de una prueba se recomienda aduciéndose que la prueba o el programa de pruebas propiamente dicho dará por resultado algún beneficio indirecto, además de la utilidad de la información de la interpretación de los puntajes de la prueba propiamente dichos, quien hace la recomendación debe explicitar la razón fundamental para prever el beneficio indirecto. Deben proporcionarse los argumentos lógicos o teóricos y la evidencia empírica para el beneficio indirecto. Debe darse la debida importancia a cualquier conclusión contradictoria en la bibliografía científica, incluyendo conclusiones que sugieran resultados indirectos importantes que no sean los pronosticados.

Comentario: Por ejemplo, se han defendido determinados programas de pruebas educativas aduciéndose que tendrían una influencia conveniente en las prácticas de instrucción en el aula o que aclararían la comprensión de los estudiantes de la clase o nivel de rendimiento que se espera que alcancen. En la medida en que dichas afirmaciones entren en la justificación para un programa de pruebas, se vuelven parte del argumento para el uso de la prueba. Se debe examinar la evidencia para dichas afirmaciones —junto con evidencia sobre la validez de la interpretación prevista de los puntajes de la prueba sobre las consecuencias

negativas imprevistas del uso de la prueba— al tomar una decisión general sobre el uso de la prueba. Debe darse la debida importancia a la evidencia contra dichas predicciones, por ejemplo, evidencia de que en algunas condiciones las pruebas educativas pueden tener un efecto negativo en la instrucción en el aula.

Estándar 1.7

Si se afirma que el desempeño en una prueba, o una decisión tomada a partir de este, se ve esencialmente afectado por la práctica y la orientación, entonces se debe documentar la propensión del desempeño en la prueba a cambiar con estas formas de instrucción.

Comentario: Los materiales para ayudar en la interpretación de los puntajes deben resumir evidencia que indique el grado en que puede esperarse la mejora con la práctica u orientación. Además, los materiales escritos para los examinandos deben proporcionar orientación práctica sobre el valor de las actividades de preparación de la prueba, incluida la orientación.

Unidad 2. Cuestiones respecto de las muestras y contextos utilizados en la validación

Estándar 1.8

La composición de cualquier muestra de examinandos de la cual se obtiene evidencia de validación debe describirse con tanto detalle como sea práctico y aceptable, incluidas características sociodemográficas y de desarrollo relevantes.

Comentario: Las conclusiones estadísticas pueden estar influenciadas por factores que afectan la muestra en la que se basan los resultados. Cuando la muestra tiene por objeto representar una población, esa población debe describirse, y debe prestarse atención a cualquier factor sistemático que pueda limitar la representatividad de la muestra. Los factores que podrían esperarse razonablemente que afecten los resultados incluyen autoselección, atrición, capacidad lingüística, condición de discapacidad, y criterios de exclusión, entre otros. Si los participantes en un estudio de validez son pacientes, por ejemplo, los diagnósticos de los pacientes son importantes, así como otras características, como la gravedad de las afecciones diagnosticadas. En pruebas utilizadas en contextos laborales, la condición de empleo (p. ej., solicitantes frente a actuales ocupantes de puestos), el nivel general de experiencia y antecedentes educativos, y la composición de género y étnica de la muestra pueden ser información relevante. En las pruebas utilizadas en acreditación, la condición de quienes brindan información (p. ej., candidatos para una credencial frente a personas ya acreditadas) es importante para interpretar los datos resultantes. En las pruebas utilizadas en contextos educativos, la información relevante puede incluir antecedentes educativos, nivel de desarrollo, características de la comunidad, o políticas de admisión escolar, como así también la composición de género y étnica de la muestra. En ocasiones, las restricciones legales sobre privacidad impiden obtener o divulgar esa información de la población o limitan el nivel de particularidad al que pueden divulgarse esos datos. Deben considerarse las leyes específicas sobre privacidad, si las hubiera, que rigen el tipo de datos, a fin de asegurar que cualquier descripción de una población no tenga el potencial de identificar a un individuo de una manera que no sea coherente con dichos estándares. Deben describirse el alcance de datos faltantes, si los hubiera, y los métodos para tratar los datos faltantes (p. ej., uso de procedimientos de imputación de datos).

Estándar 1.9

Cuando una validación se basa en parte en las opiniones o decisiones de jueces, observadores o calificadores expertos, se deben describir completamente los procedimientos para seleccionar a dichos expertos y para obtener los juicios o calificaciones. Deben presentarse las calificaciones y la experiencia de los jueces. La descripción de procedimientos debe incluir cualquier

capacitación e instrucciones proporcionadas, debe indicar si los participantes llegaron a sus decisiones de manera independiente y debe reportar el nivel de acuerdo alcanzado. Si los participantes interactuaron entre sí o intercambiaron información, deben establecerse los procedimientos mediante los cuales pueden haber ejercido influencia entre ellos.

Comentario: La recopilación sistemática de juicios u opiniones puede darse en muchos momentos en la construcción de la prueba (p. ej., obteniendo juicios expertos de lo adecuado del contenido o representación adecuada del contenido), en la formulación de reglas o estándares para la interpretación de los puntajes (p. ej., en el establecimiento de puntajes de corte), o en la calificación de la prueba (p. ej., calificación de respuestas de un ensayo). Cada vez que se empleen esos procedimientos, la calidad de los juicios resultantes es importante para la validación. El nivel de acuerdo debe especificarse claramente (p. ej., si el acuerdo de porcentaje se refiere al acuerdo anterior o posterior a una discusión de consenso, y si el criterio para el acuerdo es el acuerdo exacto de calificaciones o el acuerdo dentro una cierta cantidad de puntos de la escala). La base para especificar ciertos tipos de individuos (p. ej., profesores experimentados, titulares de puestos experimentados, supervisores) como expertos adecuados para la tarea de emitir un juicio o calificación debe articularse. Es posible que sea completamente adecuado que los expertos trabajen juntos para alcanzar el consenso, pero no sería apropiado tratar sus respectivos juicios como estadísticamente independientes. Pueden utilizarse diferentes jueces para diferentes fines (p. ej., un grupo puede calificar ítems para sensibilidad cultural mientras que otro puede calificar el nivel de lectura) o para diferentes partes de una prueba.

Estándar 1.10

Cuando la evidencia de validación incluye análisis estadísticos de los resultados de la prueba, ya sean solos o junto con datos u otras variables, las condiciones en que se recopilaron los datos deben describirse con detalle suficiente para que los usuarios puedan juzgar la relevancia de las conclusiones estadísticas para las condiciones locales. Se debe prestar atención a cualquier característica de una recopilación de datos de validación que probablemente difiera de las condiciones de prueba operativas típicas y que podría plausiblemente influir en el desempeño en la prueba.

Comentario: Esas condiciones podrían incluir (a modo de ejemplo) las siguientes: motivación o preparación previa de los examinandos, el rango de los puntajes de la prueba sobre los examinandos, el tiempo dado a los examinandos para responder u otras condiciones administrativas, el modo de administración de la prueba (p. ej., prueba en línea sin supervisión frente a prueba in situ), capacitación del examinador u otras características del examinador, los intervalos de tiempo que separan la recopilación de datos sobre diferentes medidas o las condiciones que puedan haber cambiado desde que se obtuvo la evidencia de validación.

Unidad 3. Formas específicas de evidencia de validación

(a) Evidencia orientada al contenido

Estándar 1.11

Cuando la razón fundamental para la interpretación de los puntajes de la prueba para un uso dado se basa en parte en lo apropiado del contenido de la prueba, los procedimientos seguidos en la especificación y generación del contenido de la prueba deben describirse y justificarse con referencia a la población que se prevé evaluar y al constructo que la prueba tiene por objeto medir o el dominio que tiene por objeto representar. Si la definición del contenido muestreado incorpora criterios como la importancia, frecuencia o criticidad, estos criterios también deben explicarse y justificarse con claridad.

Comentario: Por ejemplo, los desarrolladores de la prueba podrían proporcionar una estructura

lógica que mapee los ítems en la prueba al dominio de contenido, ilustrando la relevancia de cada ítem y la adecuación con la que el conjunto de ítems representa el dominio de contenido. También podrían indicarse áreas del dominio de contenido que no están incluidas entre los ítems de la prueba. El emparejamiento del contenido de la prueba con el dominio objetivo en términos de complejidad cognitiva y la accesibilidad del contenido de la prueba a todos los miembros de la población prevista también son consideraciones importantes.

(b) Evidencia respecto de los procesos cognitivos

Estándar 1.12

Si la razón fundamental para la interpretación de los puntajes para un uso dado depende de premisas sobre los procesos psicológicos u operaciones cognitivas de los examinandos, debe proporcionarse la evidencia teórica o empírica que respalde esas premisas. Cuando enunciados sobre los procesos empleados por observadores o calificadores sean parte del argumento de validez, debe proporcionarse información similar.

Comentario: Si la especificación de la prueba delinea los procesos a evaluar, entonces se necesita evidencia de que los ítems de la prueba, efectivamente, utilizan los procesos previstos.

(c) Evidencia respecto de la estructura interna

Estándar 1.13

Si la razón fundamental de la interpretación de los puntajes de una prueba para un uso dado depende de premisas sobre las relaciones entre ítems de la prueba o entre partes de la prueba, debe proporcionarse evidencia sobre la estructura interna de la prueba.

Comentario: Podría decirse, por ejemplo, que una prueba es esencialmente unidimensional. Tal afirmación podría estar respaldada por análisis estadístico multivariado, como un análisis factorial, que muestre que la variabilidad de los puntajes atribuible a una dimensión principal fue mucho mayor que la variabilidad de los puntajes atribuible a cualquier otra dimensión identificada, o que muestre que un solo factor representa adecuadamente la covarianza entre ítems de la prueba. Cuando una prueba proporciona más de un puntaje, debe mostrarse que las interrelaciones de esos puntajes son coherentes con el/los constructo(s) que se evalúan.

Estándar 1.14

Cuando se sugiere la interpretación de subpuntajes, diferencias de puntajes o perfiles, debe proporcionarse la razón fundamental y la evidencia relevante que respalde dicha interpretación. Cuando se desarrollan puntajes compuestos, se deben dar la base y la razón fundamental para llegar a los valores compuestos.

Comentario: Cuando una prueba proporciona más de un puntaje, debe demostrarse el carácter distintivo y la confiabilidad de los puntajes separados, y debe mostrarse que las interrelaciones de esos puntajes son coherentes con el/los constructo(s) que se evalúan. Asimismo, la evidencia para la validez de interpretaciones de dos o más puntajes separados no necesariamente justificaría una interpretación estadística o de contenido de la diferencia entre ellas. En cambio, la razón fundamental y la evidencia de respaldo deben concernir directamente al puntaje específico, la combinación de puntajes o el patrón de puntajes que se interpretarán para un uso dado. Cuando se combinan subpuntajes de una prueba o puntajes de diferentes pruebas en un valor compuesto, debe especificarse la base para combinar puntajes y cómo se combinan los puntajes (p. ej., ponderación diferencial frente a suma simple).

Estándar 1.15

Cuando se sugiere la interpretación del desempeño en ítems específicos, o pequeños subconjuntos de ítems, debe proporcionarse la razón

fundamental que respalde dicha interpretación. Cuando la interpretación de respuestas a ítems individuales es probable pero no recomendada por el desarrollador, se debe advertir al usuario de no hacer dichas interpretaciones.

Comentario: Se debe dar suficiente orientación a los usuarios para permitirles juzgar el grado de confianza justificado para cualquier interpretación para un uso recomendado por el desarrollador de la prueba. Los manuales de pruebas y los reportes de puntajes deben desalentar la sobreinterpretación de información que puede estar sujeta a error considerable. Esto es especialmente importante si se sugiere la interpretación del desempeño en ítems aislados, pequeños subconjuntos de ítems o puntajes de subpruebas.

(d) Evidencia respecto de las relaciones con constructos relacionados conceptualmente

Estándar 1.16

Cuando la evidencia de validación incluye análisis empíricos de respuestas a ítems de la prueba junto con datos sobre otras variables, debe proporcionarse la razón fundamental para seleccionar las variables adicionales. Cuando sea apropiado y viable, debe presentarse o citarse la evidencia concerniente a constructos representados por otras variables, así como sus propiedades técnicas. Debe prestarse atención a cualquier fuente probable de dependencia (o falta de independencia) entre variables distintas de las dependencias entres los constructos que representan.

Comentario: Los patrones de asociación entre puntajes en la prueba en estudio y otras variables deben ser coherentes con las expectativas teóricas. Las variables adicionales podrían ser características demográficas, indicadores de condiciones de tratamiento o puntajes sobre otras medidas. Podrían incluir medidas previstas del mismo constructo o de constructos diferentes. La confiabilidad de los puntajes de esas otras medidas y la validez de las interpretaciones previstas de puntajes de esas medidas son una parte importante de la evidencia de validación para la prueba en estudio. Si dichas variables incluyen puntajes compuestos, se debe explicar la manera en que se construyeron los valores compuestos (p. ej., transformación o estandarización de las variables, y ponderación de las variables). Además de considerar las propiedades de cada variable en forma aislada, es importante advertir sobre interpretaciones defectuosas que surgen de fuentes espurias de dependencia entre medidas, incluidos errores correlacionados o varianza compartida debido a métodos comunes de medición o elementos comunes.

(e) Evidencia respecto de las relaciones con criterios

Estándar 1.17

Cuando la validación se basa en evidencia de que los puntajes de la prueba están relacionados con una o más variables de criterios, debe reportarse información sobre la pertinencia y la calidad técnica de los criterios.

Comentario: La descripción de cada variable de criterio debe incluir evidencia respecto de su confiabilidad, la medida en que representa el constructo previsto (p. ej., desempeño de tareas en el puesto de trabajo), y la medida en que es probable que esté influida por fuentes de varianza externas. Debe prestarse especial atención a las fuentes que la investigación previa sugiera que pueden introducir varianza externa que podría sesgar el criterio a favor o en contra de grupos identificables.

Estándar 1.18

Cuando se asevera que un determinado nivel de desempeño en la prueba predice el desempeño adecuado o inadecuado del criterio, se debe proporcionar información sobre los niveles de desempeño del criterio asociados con niveles dados de puntajes de la prueba.

Comentario: A los fines de vincular puntajes específicos de la prueba con niveles específicos de desempeño de criterios, las ecuaciones de regresión son más útiles que los coeficientes de

correlación, que por lo general son insuficientes para describir completamente patrones de asociación entre pruebas y otras variables. Se necesitan medias, desviaciones estándares y otros resúmenes estadísticos, así como información sobre la distribución de desempeños de criterios condicionales a un puntaje determinado de una prueba. En el caso de variables categóricas más que continuas, deben utilizarse las técnicas apropiadas para dichos datos (p. ej., el uso de regresión logística en el caso de un criterio dicotómico). La evidencia sobre la asociación general entre variables debe complementarse con información sobre la forma de esa asociación y sobre la variabilidad de esa asociación en diferentes rangos de puntajes de la prueba. Obsérvese que las recopilaciones de datos que emplean examinandos seleccionados por sus puntajes extremos en una o más medidas (grupos extremos) por lo general no pueden proporcionar información adecuada sobre la asociación.

Estándar 1.19

Si se usan puntajes de la prueba junto con otras variables para predecir algún resultado o criterio, los análisis basados en modelos estadísticos de la relación predictor-criterio deben incluir esas variables relevantes adicionales junto con los puntajes de la prueba.

Comentario: En general, si varios predictores de algún criterio están disponibles, la combinación óptima de predictores no puede determinarse exclusivamente a partir de exámenes por pares, de la variable de criterio con cada predictor separado a su vez, debido a la intercorrelación entre predictores. Suele ser informativo estimar el incremento en la exactitud predictiva que puede esperarse cuando cada variable, incluyendo el puntaje de la prueba, se introduce además de todas las demás variables disponibles. Como las ponderaciones derivadas empíricamente para combinar predictores pueden aprovechar factores aleatorios en una muestra dada, los análisis que involucran múltiples predictores deben verificarse mediante validación cruzada o análisis equivalente siempre que sea viable, y debe reportarse la precisión de los coeficientes de regresión u otros índices. Los procedimientos de validación cruzada incluyen estimaciones de validez de fórmulas en muestras posteriores y enfoques empíricos como derivar ponderaciones en una parte de una muestra y aplicarlas a una submuestra independiente.

Estándar 1.20

Cuando las medidas del tamaño del efecto (p. ej., correlaciones entre puntajes de la prueba y medidas de criterios, diferencias de puntajes medios estandarizados de la prueba entre subgrupos) se usan para obtener inferencias que van más allá de describir la muestra o las muestras sobre las que se han recopilado datos, deben reportarse índices del grado de incertidumbre asociado con estas medidas (p. ej., errores estándares, intervalos de confianza o pruebas de significación).

Comentario: Las medidas del tamaño del efecto se emparejan de manera útil con índices que reflejan su error de muestreo para hacer que sea posible la evaluación significativa. Hay varias medidas posibles del tamaño del efecto, cada una aplicable a diferentes contextos. En la presentación de índices de incertidumbre, los errores estándares o intervalos de confianza proporcionan más información y en consecuencia se prefieren en lugar de las pruebas de significación o como complemento de estas.

Estándar 1.21

Cuando se realizan ajustes estadísticos, como aquellos para restricción de rango o atenuación, se deben reportar tanto los coeficientes ajustados como los no ajustados, así como el procedimiento específico utilizado y todas las estadísticas utilizadas en el ajuste. Las estimaciones de la relación constructo-criterio que eliminan los efectos del error de medición en la prueba deben reportarse claramente como estimaciones ajustadas.

Comentario: La correlación entre dos variables, como los puntajes de la prueba y las medidas de criterio, depende del rango de valores de cada variable. Por ejemplo, los puntajes de la prueba y

los valores de criterio de un subconjunto seleccionado de examinandos (p. ej., solicitantes para un puesto de trabajo que han sido seleccionados para contratación) por lo general tendrán un rango menor que los puntajes de todos los examinandos (p. ej., todo el grupo de solicitantes). Hay métodos estadísticos disponibles para ajustar la correlación para reflejar la población de interés en lugar de la muestra disponible. Esos ajustes suelen ser apropiados, como cuando los resultados se comparan entre varias situaciones. La correlación entre dos variables también está afectada por error de medición, y hay métodos disponibles para ajustar la correlación para estimar la fortaleza de la correlación neta de los efectos del error de medición en cualquiera de las variables o en ambas. La presentación de reportes de una correlación ajustada debe estar acompañada por un enunciado del método y las estadísticas utilizados para hacer el ajuste.

Estándar 1.22

Cuando se utiliza un metaanálisis como evidencia de la fortaleza de una relación prueba-criterio, las variables de prueba y criterio en la situación local deben ser comparables con las de los estudios resumidos. Si la investigación relevante incluye evidencia creíble de que cualquier otra característica específica de la aplicación de la prueba puede influir en la fortaleza de la relación prueba-criterio, debe reportarse la correspondencia entre esas características en la situación local y en el metaanálisis. Deben observarse explícitamente cualquier disparidad significativa que pudiera limitar la aplicabilidad de las conclusiones del metaanálisis a la situación local.

Comentario: El metaanálisis debe incorporar todos los estudios disponibles que reúnan explícitamente los criterios de inclusión indicados. La evidencia metaanalítica utilizada en la validación de la prueba suele basarse en una serie de pruebas que miden los mismos constructos o constructos muy similares y medidas de criterio que del mismo modo miden los mismos o similares constructos. Un estudio metaanalítico también puede limitarse a múltiples estudios de una sola prueba y un solo criterio. Para cada estudio incluido en el análisis, la relación prueba-criterio se expresa en alguna métrica común, a menudo como un tamaño del efecto. La fortaleza de la relación prueba-criterio puede ser moderada por características de la situación en la que se obtuvieron las medias de la prueba y el criterio (p. ej., tipos de puestos de trabajo, características de los examinandos, intervalo de tiempo entre la recolección de medidas de la prueba y del criterio, año o década en la que se recopilaron los datos). Si las relaciones prueba-criterio varían de acuerdo con esas variables moderadoras, el metaanálisis debe reportar distribuciones efecto-tamaño estimadas separadas condicionales a los niveles de esas variables moderadoras cuando la cantidad de estudios disponibles para análisis permita hacerlo. Esto puede lograrse, por ejemplo, reportando distribuciones separadas para subconjuntos de estudios o estimando las magnitudes de las influencias de características situacionales sobre los tamaños del efecto.

Este estándar aborda las responsabilidades del individuo que está recurriendo a evidencia metaanalítica para respaldar una interpretación de los puntajes de la prueba para un uso dado. En algunos casos, ese individuo puede también ser la que realiza el metaanálisis; en otros casos, se basa en metaanálisis existentes. En el último caso, el individuo que recurre a evidencia metaanalítica no tiene control sobre cómo se realizó o informó el metaanálisis, y debe evaluar la solidez del metaanálisis para el contexto en cuestión.

Estándar 1.23

Cualquier evidencia metaanalítica utilizada para respaldar una interpretación prevista de los puntajes de la prueba debe describirse claramente, incluidas las elecciones metodológicas en la identificación y codificación de estudios, corrección de artefactos y examen de potenciales variables moderadoras. Deben presentarse las suposiciones hechas en la corrección de artefactos como falta de confiabilidad del criterio y restricción

de rango, y deben aclararse las consecuencias de esas suposiciones.

Comentario: La descripción debe incluir información documentada sobre cada estudio utilizado como dato de entrada en el metaanálisis, permitiendo así la evaluación por una parte independiente. Obsérvese también que el metaanálisis involucra inevitablemente una serie de opciones metodológicas. Las bases para estos juicios deben articularse. En el caso de elecciones que involucran algún grado de incertidumbre, como correcciones de artefactos basadas en valores supuestos, la incertidumbre debe reconocerse y debe examinarse y reportarse el grado en que las conclusiones sobre validez dependen de estas suposiciones.

Como en el caso del Estándar 1.22, el individuo que recurre a evidencia metaanalítica para respaldar la interpretación de puntajes de una prueba para un uso dado puede ser o no también el que realiza el metaanálisis. Como el Estándar 1.22 aborda el reporte de evidencia metaanalítica, el individuo que recurre a evidencia metaanalítica existente debe evaluar la solidez del análisis metaanalítico para el contexto en cuestión.

Estándar 1.24

Si se recomienda una prueba para usar en la asignación de personas a tratamientos alternativos, y si los resultados de esos tratamientos pueden compararse razonablemente sobre un criterio en común, entonces, cuando sea viable, debe proporcionarse evidencia de respaldo de los resultados diferenciales.

Comentario: Si una prueba se utiliza para clasificación en programas ocupacionales, terapéuticos o educativos alternativos, no es suficiente solo mostrar que la prueba predice resultados de tratamiento. El respaldo de la validez del procedimiento de clasificación se proporciona mostrando que la prueba es útil para determinar qué personas probablemente se beneficien de manera diferencial con un tratamiento u otro. Es posible que deban combinarse categorías de tratamiento para reunir suficientes casos para análisis estadísticos. Se reconoce, no obstante, que es posible que esa investigación no sea viable, porque las restricciones éticas y legales sobre asignaciones diferenciales pueden prohibir los grupos de control.

(f) Evidencia basada en consecuencias de las pruebas

Estándar 1.25

Cuando surgen consecuencias imprevistas del uso de la prueba, debe intentarse investigar si dichas consecuencias surgen de la sensibilidad de la prueba a características distintas de las que tiene previsto evaluar o de que la prueba no logra representar completamente el constructo previsto.

Comentario: La validez de las interpretaciones de los puntajes de la prueba puede estar limitada por componentes irrelevantes de constructo o infrarrepresentación de constructo. Cuando las consecuencias imprevistas parecen provenir, al menos en parte, del uso de una o más pruebas, es especialmente importante comprobar que estas consecuencias no surjan de componentes irrelevantes de constructo o infrarrepresentación de constructo. Por ejemplo, si bien las diferencias del grupo, de por sí, no cuestionan la validez de una interpretación propuesta, pueden aumentar la prominencia de hipótesis rivales plausibles que deben evaluarse como parte del esfuerzo de validación. Encontrar consecuencias imprevistas también puede llevar a reconsiderar lo adecuado del constructo en cuestión. Asegurar que las consecuencias imprevistas se evalúen es responsabilidad de quienes toman la decisión de usar o no una prueba en particular, aunque las restricciones legales puedan limitar la discreción del usuario de la prueba para descartar los resultados de una prueba administrada previamente, cuando esa decisión se base en diferencias en puntajes para subgrupos de diferentes razas, orígenes étnicos o géneros. Estas cuestiones se analizan en mayor detalle en el capítulo 3.

2. CONFIABILIDAD/PRECISIÓN Y ERRORES DE MEDIDA

ANTECEDENTES

Una prueba, definida en términos generales, es un conjunto de tareas o estímulos diseñado para suscitar respuestas que proporcionen una muestra del comportamiento o desempeño de un individuo examinado en un dominio especificado. La prueba está acompañada por un procedimiento de calificación que permite al evaluador evaluar las muestras de comportamiento o trabajo y generar un puntaje. Al interpretar y utilizar puntajes de prueba es importante tener alguna indicación de su confiabilidad.

El término *confiabilidad* se ha utilizado de dos maneras en la bibliografía de medición. En primer lugar, el término se ha utilizado para hacer referencia a los coeficientes de confiabilidad de la teoría clásica de los tests, definidos como la correlación entre puntajes en dos formularios equivalentes de la prueba, suponiendo que completar un formulario no tiene efecto sobre el desempeño en el segundo formulario. En segundo lugar, el término se ha utilizado en un sentido más general para hacer referencia a la coherencia de puntajes entre replicaciones de un procedimiento de evaluación, independientemente de cómo se estime o reporte esta coherencia (p. ej., en términos de errores estándares, coeficientes de confiabilidad per se, coeficientes de generabilidad, relaciones error/tolerancia, funciones de información de la teoría de respuesta al ítem (TRI), o diversos índices de coherencia de clasificación). Para mantener un vínculo con las nociones tradicionales de confiabilidad y evitar al mismo tiempo la ambigüedad inherente en el uso de un único término conocido para hacer referencia a una amplia variedad de conceptos e índices, utilizamos el término *confiabilidad/precisión* para indicar la noción más general de coherencia de los puntajes entre instancias del procedimiento de evaluación, y el término *coeficiente de confiabilidad* para hacer referencia a los coeficientes de confiabilidad de la teoría clásica de los tests.

La confiabilidad/precisión de medida es siempre importante. Sin embargo, la necesidad de precisión aumenta a medida que las consecuencias de las decisiones e interpretaciones crecen en importancia. Si el puntaje de una prueba conduce a una decisión que no se revierte fácilmente, como la denegación o admisión de un candidato a una escuela de formación, o un juicio clínico basado en el puntaje (p. ej., en un contexto legal) respecto de que se ha sufrido una lesión cognitiva grave, se justifica un mayor grado de confiabilidad/precisión. Si una decisión puede corroborarse y será corroborada por información de otras fuentes o si una decisión inicial errónea puede corregirse fácilmente, los puntajes con confiabilidad/precisión más modesta pueden ser suficientes.

Las interpretaciones de los puntajes de una prueba por lo general dependen de suposiciones de que los individuos y grupos exhiben cierto grado de coherencia en sus puntajes entre administraciones independientes del procedimiento de evaluación. Sin embargo, diferentes muestras de desempeño de la misma persona rara vez son idénticas. Los desempeños, productos y respuestas de un individuo a conjuntos de tareas o preguntas de una prueba varían en calidad o carácter de una muestra de tareas a otra y de una ocasión a otra, incluso en condiciones estrictamente controladas. Diferentes evaluadores pueden asignar diferentes puntajes a un desempeño específico. Todas estas fuentes de variación se reflejan en los puntajes de los individuos examinados, que variarán entre instancias de un procedimiento de medición.

La confiabilidad/precisión de los puntajes depende de cuánto varíen los puntajes entre replicaciones del procedimiento de evaluación, y los análisis de confiabilidad/precisión dependen

de las clases de variabilidad permitidas en el procedimiento de evaluación (p. ej., entre tareas, contextos, evaluadores) y la interpretación propuesta de los puntajes de la prueba. Por ejemplo, si la interpretación de los puntajes supone que el constructo que se evalúa no varía entre ocasiones, la variabilidad entre ocasiones es una posible fuente de error de medida. Si las tareas de la prueba varían entre formularios alternativos de la prueba, y los desempeños observados se tratan como una muestra de un dominio de tareas similares, la variabilidad aleatoria en los puntajes de un formulario a otro se consideraría un error. Si se utilizan evaluadores para asignar puntajes a respuestas, la variabilidad en los puntajes entre evaluadores cualificados es una fuente de error. Las variaciones en los puntajes de un examinando que no son coherentes con la definición del constructo que se evalúa se atribuyen a errores de medida.

Una manera muy básica de evaluar la coherencia de puntajes involucra un análisis de la variación en los puntajes de cada examinando entre replicaciones del procedimiento de evaluación. La prueba se administra y luego, tras un período breve durante el cual no se prevería que cambie la situación del individuo examinado respecto de la variable sometida a medición, la prueba (o un formulario distinto pero equivalente de la prueba) se administra por segunda vez; se supone que la primera administración no tiene influencia sobre la segunda administración. Dado que se supone que el atributo sometido a medición permanece igual para cada examinado durante las dos administraciones y que las administraciones de la prueba son independientes una de otra, más variación entre las dos administraciones indica más error en los puntajes de la prueba y, por lo tanto, menor confiabilidad/precisión.

El impacto de dichos errores de medida puede resumirse de varias maneras, pero generalmente, en la medición educativa y psicológica, se conceptualiza en términos de la desviación estándar en los puntajes para una persona durante replicaciones del procedimiento de evaluación. En la mayoría de los contextos de evaluación, no es posible replicar el procedimiento de evaluación repetidas veces y, por lo tanto, no es posible estimar el error estándar para el puntaje de cada persona mediante medición repetida. En cambio, utilizando suposiciones basadas en modelos, el error promedio de medida se estima respecto de alguna población, y este promedio se denomina *error estándar de medida* (SEM, por sus siglas en inglés). El SEM es un indicador de una falta de coherencia en los puntajes generados por el procedimiento de evaluación para alguna población. Un SEM relativamente grande indica confiabilidad/precisión relativamente baja. El *error estándar de medida condicional* para un nivel de puntaje es el error estándar de medida a ese nivel de puntaje.

Decir que un puntaje incluye error implica que existe un valor hipotético sin error que caracteriza la variable que se evalúa. En la teoría clásica de los tests, este valor sin error se denomina *puntaje verdadero* de la persona para el procedimiento de la prueba. Se conceptualiza como el puntaje promedio hipotético en un conjunto infinito de replicaciones del procedimiento de evaluación. En términos estadísticos, el puntaje verdadero de una persona es un parámetro desconocido, o constante, y el puntaje observado para la persona es una variable aleatoria que fluctúa en torno al puntaje verdadero para la persona.

La *teoría de generabilidad* proporciona un marco diferente para estimar la confiabilidad/precisión. Si bien la teoría clásica de los tests supone una sola distribución para los errores en los puntajes de un examinando, la teoría de generabilidad busca evaluar las contribuciones de diferentes fuentes de error (p. ej., ítems, ocasiones, evaluadores) al error general. El *puntaje de universo* para una persona se define como el valor esperado sobre un universo de todas las replicaciones posibles de un procedimiento de evaluación para el examinando. El puntaje de universo de la teoría de generabilidad cumple un rol que es similar al rol de los puntajes verdaderos en la teoría clásica de los tests.

La teoría de respuesta al ítem (TRI) aborda la cuestión básica de la confiabilidad/precisión utilizando funciones de información, que indican la

precisión con la que los desempeños en las tareas/ítems observados pueden utilizarse para estimar el valor de un rasgo latente para cada examinando. Utilizando TRI, los índices análogos a los coeficientes de confiabilidad tradicionales pueden estimarse a partir de las funciones de información del ítem y distribuciones del rasgo latente en alguna población.

En la práctica, la confiabilidad/precisión de los puntajes suele evaluarse en términos de varios coeficientes, incluyendo coeficientes de confiabilidad, coeficientes de generabilidad, y funciones de información de TRI, dependiendo del enfoque del análisis y del modelo de medición que se utilice. Los coeficientes tienden a tener valores altos cuando la variabilidad asociada con el error es pequeña en comparación con la variación observada en los puntajes (o diferencias de puntajes) a estimar.

Implicaciones para la validez

Si bien en este caso se analiza la confiabilidad/precisión como una característica independiente de los puntajes de prueba, debe reconocerse que el nivel de confiabilidad/precisión de puntajes tiene implicaciones para la validez. La confiabilidad/precisión de datos en última instancia incide en la generabilidad o fiabilidad de los puntajes y/o la coherencia de clasificaciones de individuos derivadas de los puntajes. En la medida en que los puntajes no sean coherentes entre replicaciones del procedimiento de evaluación (es decir, en la medida en que reflejen errores de medida aleatorios), su potencial de predicción exacta de criterios, para diagnóstico beneficioso del individuo examinado, y para toma de decisiones inteligentes es limitado.

Especificaciones para replicaciones del procedimiento de evaluación

Como se indicó anteriormente, la noción general de confiabilidad/precisión se define en términos de coherencia entre replicaciones del procedimiento de evaluación. La confiabilidad/precisión es alta si los puntajes para cada persona son coherentes entre replicaciones del procedimiento de evaluación y es baja si los puntajes no son coherentes entre replicaciones. Por lo tanto, al evaluar la confiabilidad/precisión, es importante ser claros respecto de qué constituye una replicación del procedimiento de evaluación.

Las replicaciones involucran administraciones independientes del procedimiento de evaluación, tal que no se esperaría que el atributo sometido a medición cambie. Por ejemplo, al evaluar un atributo que no se espera que cambie durante un período de tiempo prolongado (p. ej., en la medición de un rasgo), los puntajes generados en dos días consecutivos (utilizando diferentes formularios de prueba si corresponde) se considerarían replicaciones. Para una variable de estado (p. ej., estado de ánimo o hambre), donde los cambios bastante rápidos son comunes, los puntajes generados en dos días consecutivos no se considerarían replicaciones; los puntajes obtenidos en cada ocasión se interpretarían en términos del valor de la variable de estado en esa ocasión. En muchas pruebas de conocimiento o habilidad, la administración de formularios alternativos de una prueba con diferentes muestras de ítems se considerarían replicaciones de la prueba; para instrumentos de sondeo y algunas medidas de personalidad, se espera que las mismas preguntas se utilicen cada vez que se administre la prueba, y cualquier cambio sustancial en la redacción constituiría un formulario de prueba diferente.

Las pruebas estandarizadas presentan los mismos materiales de la prueba o materiales muy similares a todos los examinandos, mantienen una rigurosa adhesión a procedimientos estipulados para la administración de pruebas y emplean reglas de calificación prescriptas que pueden aplicarse con un alto grado de coherencia. Administrar las mismas preguntas o preguntas puestas en una escala común a todos los examinandos en las mismas condiciones promueve la imparcialidad y facilita las comparaciones de puntajes entre individuos. Las condiciones de observación que se fijan o estandarizan para el procedimiento de evaluación permanecen iguales entre replicaciones. Sin embargo, se permitirá variar algunos aspectos de cualquier procedimiento de evaluación

estandarizado. Por lo general se permite que el momento y el lugar de evaluación, así como las personas que administran la prueba, varíen en cierta medida. Es posible que se permita variar las tareas en particular incluidas en la prueba (como muestras de un dominio de contenido común), y las personas que califican los resultados pueden variar en algún conjunto de evaluadores cualificados.

Los *formularios alternativos* (o *formularios paralelos*) de una prueba estandarizada se diseñan para que tengan la misma distribución general de contenido y formatos de ítems (según lo descripto, por ejemplo, en especificaciones de la prueba detalladas), los mismos procedimientos administrativos y al menos aproximadamente las mismas medias de puntaje y desviaciones estándares en alguna población o poblaciones especificadas. Los formularios alternativos de una prueba se consideran intercambiables, en el sentido de que se elaboran según las mismas especificaciones, y se interpretan como medidas del mismo constructo.

En la teoría clásica de los tests, se supone que las pruebas estrictamente paralelas miden el mismo constructo y arrojan puntajes que tienen las mismas medias y desviaciones estándares en las poblaciones de interés y tienen las mismas correlaciones con todas las demás variables. Un coeficiente de confiabilidad clásico se define en términos de la correlación entre puntajes de formularios estrictamente paralelos de la prueba, pero se estima en términos de la correlación entre formularios alternativos de la prueba que pueden no ser tan estrictamente paralelos.

Pueden implementarse diferentes enfoques a la estimación de confiabilidad/precisión para ajustarse a diferentes diseños de recopilación de datos y diferentes interpretaciones y usos de puntajes. En algunos casos, es posible que sea viable estimar la variabilidad entre replicaciones directamente (p. ej., teniendo una serie de evaluadores cualificados que evalúen una muestra de desempeños en la prueba para cada examinando). En otros casos, es posible que sea necesario usar estimaciones menos directas del coeficiente de confiabilidad. Por ejemplo, las estimaciones de confiabilidad de coherencia interna (p. ej., coeficiente dividido, KR-20, coeficiente alfa) utilizan la medida de concordancia observada entre diferentes partes de una prueba para estimar la confiabilidad asociada con variabilidad entre formularios. Para el método dividido, se correlacionan los puntajes en dos mitades más o menos paralelas de la prueba (p. ej. ítems con números impares e ítems con números pares), y el coeficiente de confiabilidad de la mitad de la prueba que se obtiene se ajusta estadísticamente para estimar la confiabilidad de la prueba completa. Sin embargo, cuando una prueba se diseña para reflejar la tasa de trabajo, es probable que las estimaciones de confiabilidad de coherencia interna (en particular por el método par-impar) arrojen estimaciones infladas de confiabilidad para pruebas de aceleración alta.

En algunos casos, es posible que sea razonable suponer que es probable que una posible fuente de variabilidad sea insignificante o que el usuario podrá inferir confiabilidad adecuada de otros tipos de evidencia. Por ejemplo, si los puntajes de una prueba se utilizan principalmente para predecir algunos puntajes de criterio y la prueba hace un trabajo aceptable en la predicción del criterio, puede inferirse que los puntajes de la prueba son suficientemente confiables/precisos para su uso previsto.

La definición de lo que constituye una prueba o procedimiento de medición estandarizado se ha ampliado significativamente en las últimas décadas. Se han desarrollado varias clases de evaluaciones de desempeño, simulaciones y evaluaciones basadas en porfolios para brindar medidas de constructos que de otro modo podrían ser difíciles de evaluar. Cada paso hacia una mayor flexibilidad en los procedimientos de evaluación amplía el alcance de las variaciones permitidas en replicaciones del procedimiento de evaluación, y por lo tanto tiende a aumentar el error de medida. Sin embargo, algunos de estos sacrificios en la confiabilidad/precisión pueden reducir la irrelevancia de constructo o infrarrepresentación de constructo y, por consiguiente, mejorar la validez de las interpretaciones previstas de los puntajes. Por ejemplo, las evaluaciones de desempeño que

dependen de calificaciones de respuestas extendidas tienden a tener menor confiabilidad que las evaluaciones más estructuradas (p. ej., pruebas de opciones múltiples o de respuestas cortas), pero a veces pueden proporcionar medidas más directas del atributo de interés.

Los *errores de medida aleatorios* se ven como fluctuaciones impredecibles en los puntajes. Se distinguen conceptualmente de los errores sistemáticos, que también pueden afectar los desempeños de individuos o grupos, pero de una manera coherente más que aleatoria. Por ejemplo, una hoja de respuestas incorrecta contribuiría a un *error sistemático*, como lo harían las diferencias en la dificultad de los formularios de prueba que no se hayan equiparado o vinculado adecuadamente; los individuos examinados que completen un formulario pueden recibir puntajes más altos en promedio que si hubieran completado el otro formulario. Esos errores sistemáticos por lo general no se incluirían en el error estándar de medida, y no se considera que contribuyan a una falta de confiabilidad/precisión. En cambio, los errores sistemáticos constituyen factores irrelevantes de constructo que reducen la validez, pero no la confiabilidad/precisión.

Las fuentes importantes de error aleatorio pueden agruparse en dos categorías amplias: las que tienen su origen en los examinandos y las externas a ellos. Las fluctuaciones en el nivel de motivación, interés o atención de un individuo examinado y la aplicación incoherente de habilidades son claramente fuentes internas que pueden conducir a error aleatorio. Las variaciones en las condiciones de evaluación (p. ej., momento del día, nivel de distracciones) y las variaciones en la calificación debido a subjetividad del evaluador son ejemplos de fuentes externas que pueden conducir a error aleatorio. La importancia de cualquier fuente de variación en particular depende de las condiciones específicas en las que se tomen las medidas, cómo se califican los desempeños y las interpretaciones derivadas de los puntajes.

Algunos cambios en los puntajes de una ocasión a otra no se consideran error (aleatorio o sistemático), porque surgen, en parte, de cambios en el constructo sometido a medición (p. ej., debido a aprendizaje o maduración que ha ocurrido entre las medidas iniciales y finales). En esos casos, los cambios en el desempeño constituirían el fenómeno de interés y no se considerarían errores de medida.

El error de medida reduce la utilidad de los puntajes de prueba. Limita la medida en que los resultados de la prueba pueden generalizarse más allá de los detalles de una replicación dada del procedimiento de evaluación. Reduce la confianza que puede depositarse en los resultados de una sola medición y por lo tanto la confiabilidad/precisión de los puntajes. Dado que los errores de medida aleatorios son impredecibles, no pueden eliminarse de los puntajes observados. Sin embargo, su magnitud agregada puede resumirse de varias maneras, como se analiza a continuación, y pueden controlarse hasta cierto punto (p. ej., mediante estandarización o promediando múltiples puntajes).

El error estándar de medida, como tal, proporciona una indicación del nivel esperado de error aleatorio entre puntos de puntaje y replicaciones para una población específica. En muchos casos, es útil tener estimaciones de los errores estándares para cada individuo examinado (o para individuos examinados con puntajes en determinados rangos de puntaje). Estos errores estándares condicionales son difíciles de estimar en forma directa, pero pueden estimarse indirectamente. Por ejemplo, las funciones de información de prueba basadas en modelos de TRI pueden usarse para estimar errores estándares para diferentes valores de un parámetro de capacidad latente y/o para diferentes puntajes observados. Al usar cualquiera de estas estimaciones de errores estándares condicionales basadas en modelos, es importante que las suposiciones del modelo sean coherentes con los datos.

Evaluación de la confiabilidad/precisión

El enfoque ideal de la evaluación de confiabilidad/precisión requeriría muchas replicaciones independientes del procedimiento de evaluación en una muestra grande de examinandos. El rango de diferencias permitido en replicaciones del procedimiento de evaluación y la interpretación

propuesta de los puntajes proporcionan un marco para investigar la confiabilidad/precisión.

En la mayoría de los programas de evaluación, se espera que los puntajes se generalicen entre formularios alternativos de la prueba, ocasiones (dentro del mismo período), contextos de evaluación y evaluadores (si se requiere juicio en la calificación). En la medida en que se prevea que el impacto de cualquiera de estas fuentes de variabilidad sea sustancial, la variabilidad debería estimarse de alguna manera. No es necesario que las diferentes fuentes de varianza se estimen por separado. La confiabilidad/precisión general, dada la varianza de error debido al muestreo de formularios, ocasiones y evaluadores, puede estimarse a través de un estudio test-retest que involucre diferentes formularios administrados en diferentes ocasiones y calificados por diferentes evaluadores.

La interpretación de los análisis confiabilidad/precisión depende de la población que se evalúa. Por ejemplo, los coeficientes de confiabilidad o generabilidad derivados de puntajes de una muestra representativa a nivel nacional pueden diferir significativamente de los obtenidos de una muestra más homogénea tomada de un género, un grupo étnico o una comunidad. Por lo tanto, en la medida en que sea viable (es decir, si los tamaños de la muestra son lo suficientemente grandes), la confiabilidad/ precisión debe estimarse por separado para todos los subgrupos relevantes (p. ej., definidos en términos de raza/origen étnico, género, competencia en un idioma) en la población. (Véase también el cap. 3, "Imparcialidad en las pruebas").

Coeficientes de confiabilidad/ generabilidad

En la teoría clásica de los tests, la coherencia de los puntajes de una prueba se evalúa principalmente en términos de coeficientes de confiabilidad, definidos en términos de la correlación entre puntajes derivados de replicaciones del procedimiento de evaluación en una muestra de examinandos. Se reconocen tres amplias categorías de coeficientes de confiabilidad: (a) coeficientes derivados de la administración de formularios alternativos en sesiones de evaluación independientes (*coeficientes de formularios alternativos*); (b) coeficientes obtenidos mediante la administración del mismo formulario en ocasiones separadas (*coeficientes test-retest*); y (c) coeficientes basados en las relaciones/interacciones entre puntajes derivados de ítems individuales o subconjuntos de los ítems dentro de una prueba, donde todos los datos se acumulan de una sola administración (*coeficientes de coherencia interna*). Además, cuando la calificación de la prueba involucra un alto nivel de juicio, se obtienen comúnmente índices de coherencia entre evaluadores. En tratamientos formales de la teoría clásica de los tests, la confiabilidad puede definirse como la relación de la varianza de puntaje verdadero respecto de la varianza de puntaje observado, pero se estima en términos de coeficientes de confiabilidad de las clases mencionadas arriba.

En la teoría de generabilidad, estos análisis de confiabilidad diferentes se tratan como casos especiales de un marco más general para estimar la varianza de error en términos de los componentes de varianza asociados con diferentes fuentes de error. Un *coeficiente de generabilidad* se define como la relación de la varianza del puntaje de universo con respecto a la varianza del puntaje observado. A diferencia de los enfoques tradicionales al estudio de la confiabilidad, la teoría de generabilidad alienta al investigador a especificar y estimar componentes de varianza de puntaje verdadero, varianza de puntaje de error y varianza de puntaje observado, y a calcular coeficientes basados en estas estimaciones. La estimación suele realizarse mediante la aplicación de técnicas de análisis de varianza. Las estimaciones numéricas separadas de los componentes de varianza (p. ej., componentes de varianza para ítems, ocasiones y evaluadores, y para las interacciones entre estas posibles fuentes de error) pueden utilizarse para evaluar la contribución de cada fuente de error al error de medida general; las estimaciones del componente de varianza pueden ser útiles en la identificación de una estrategia efectiva para controlar la varianza de error general.

Diferentes coeficientes de confiabilidad (y generabilidad) pueden parecer intercambiables,

pero los diferentes coeficientes transmiten información diferente. Un coeficiente puede abarcar una o más fuentes de error. Por ejemplo, un coeficiente puede reflejar error debido a incoherencias del evaluador, pero no reflejar la variación en los desempeños o productos de un individuo examinado. Un coeficiente puede reflejar solo la coherencia interna de repuestas al ítem dentro de un instrumento y no reflejar el error de medida asociado con los cambios diarios en el desempeño del individuo examinado.

No debe inferirse, sin embargo, que los coeficientes de formularios alternativos o test-retest basados en administraciones de la prueba con varios días o semanas de diferencia son siempre preferibles a los coeficientes de coherencia interna. En casos en que podemos suponer que no es probable que los puntajes cambien, en función de experiencia pasada y/o consideraciones teóricas, es posible que sea razonable suponer invariancia entre ocasiones (sin realizar un estudio test-retest). Otra limitación de los coeficientes test-retest es que, cuando se utiliza el mismo formulario de la prueba, la correlación entre los primeros y segundos puntajes podría inflarse por el recuerdo del examinando de las respuestas iniciales.

La función de información de prueba, un resultado importante de TRI, resume qué tan bien la prueba discrimina entre individuos en varios niveles de capacidad en el rasgo que se evalúa. En la conceptualización de TRI para ítems calificados de manera dicotómica, *la curva característica de ítem* o *función de respuesta al ítem* se utiliza como un modelo para representar la proporción creciente de respuestas correctas a un ítem en niveles crecientes de la capacidad o rasgo sometido a medición. Dados los datos apropiados, pueden estimarse los parámetros de la curva característica para cada ítem en una prueba. La función de información de prueba puede entonces calcularse a partir de estimaciones de parámetros para el conjunto de ítems en la prueba y puede usarse para derivar coeficientes con interpretaciones similares a los coeficientes de confiabilidad.

La función de información puede verse como un enunciado matemático de la precisión de medida en cada nivel del rasgo dado. La función de información de TRI se basa en los resultados obtenidos en una ocasión específica o en un contexto específico, y por lo tanto no proporciona una indicación de generabilidad entre ocasiones o contextos.

Los coeficientes (p. ej., coeficientes de confiabilidad, generabilidad y basados en TRI) tienen dos ventajas principales sobre los errores estándares. En primer lugar, como se indicó anteriormente, pueden usarse para estimar errores estándares (generales y/o condicionales) en casos en que no sería posible hacerlo directamente. En segundo lugar, los coeficientes (p. ej., coeficientes de confiabilidad y generabilidad), que se definen en términos de relaciones de varianzas para puntajes en la misma escala, son invariantes en transformaciones lineales de la escala de puntajes y pueden ser útiles para comparar diferentes procedimientos de evaluación sobre la base de escalas diferentes. Sin embargo, esas comparaciones rara vez son directas, porque pueden depender de la variabilidad de los grupos en que se basan los coeficientes, las técnicas usadas para obtener los coeficientes, las fuentes de error reflejadas en los coeficientes, y las extensiones y contenidos de los instrumentos que se comparan.

Factores que afectan la confiabilidad/precisión

Varios factores pueden tener efectos significativos en la confiabilidad/precisión, y en algunos casos, esos factores pueden conducir a interpretaciones erróneas de los resultados, si no se tienen en cuenta.

En primer lugar, cualquier evaluación de confiabilidad/precisión se aplica a un procedimiento de evaluación en particular y es probable que cambie si el procedimiento cambia de cualquier manera sustancial. En general, si la evaluación es acortada (p. ej., reduciendo la cantidad de ítems o tareas), es probable que la confiabilidad disminuya; y si la evaluación se extiende con tareas o ítems comparables, es probable que la confiabilidad aumente. De hecho, extender la evaluación, y por consiguiente aumentar el tamaño de la muestra de tareas/ítems (o evaluadores u ocasiones) que

se utilizan, es un método efectivo y comúnmente utilizado para mejorar la confiabilidad/precisión.

En segundo lugar, si la variabilidad asociada con evaluadores se estima para un grupo selecto de evaluadores que han sido especialmente bien capacitados (y tal vez participaron en el desarrollo de los procedimientos), pero los evaluadores no están tan bien capacitados en algunos contextos operativos, el error asociado con la variabilidad de evaluadores en estos contextos operativos puede ser mucho más alta que la indicada por los coeficientes de confiabilidad entre los evaluadores reportados. De manera similar, si los evaluadores aún están perfeccionando su desempeño en los primeros días de una ventana de calificación extendida, el error asociado con la variabilidad entre evaluadores puede ser mayor para individuos examinados que realizan la prueba antes en la ventana que para los que la realizan más adelante.

La confiabilidad/precisión también puede depender de la población para la que se utiliza el procedimiento. En particular, si la variabilidad en el constructo de interés en la población para la que se generan los puntajes es sustancialmente diferente de lo que es en la población para la que se evaluó la confiabilidad/precisión, la confiabilidad/precisión puede ser bastante diferente en las dos poblaciones. Cuando la variabilidad en el constructo sometido a medición es baja, los coeficientes de confiabilidad y generabilidad tienden a ser pequeños, y cuando la variabilidad en el constructo sometido a medición es más alta, los coeficientes tienden a ser más grandes. Los errores estándares de medida dependen menos de la variabilidad en la muestra de examinandos que los coeficientes de confiabilidad y generabilidad.

Además, la confiabilidad/precisión puede variar de una población a otra, incluso si la variabilidad en el constructo de interés en las dos poblaciones es la misma. La confiabilidad puede variar de una población a otra porque fuentes de error en particular (efectos del evaluador, familiaridad con formatos e instrucciones, etc.) tienen más impacto en una población que en la otra. En general, si algunos aspectos de los procedimientos de evaluación o de la población que se evalúa se cambian en un contexto operativo, la confiabilidad/precisión puede cambiar.

Errores estándares de medida

El error estándar de medida puede utilizarse para generar intervalos de confianza en torno a puntajes reportados. Por lo tanto, es generalmente más informativo que un coeficiente de confiabilidad o generabilidad, una vez que se ha adoptado un procedimiento de medición y la interpretación de puntajes se ha vuelto el principal interés del usuario.

Las estimaciones de los errores estándares en diferentes niveles de puntaje (es decir, errores estándares condicionales) por lo general son un complemento valioso para la estadística única para todos los niveles de puntaje combinados. Los errores estándares de medida condicionales pueden ser mucho más informativos que un solo error estándar promedio para una población. Si las decisiones se basan en puntajes de la prueba y esas decisiones se concentran en un área o algunas áreas de la escala de puntajes, los errores condicionales en esas áreas son de especial interés.

Al igual que los coeficientes de confiabilidad y generalidad, los errores estándares pueden reflejar variación de muchas fuentes de error o de solo algunas. Un error estándar más completo (es decir, uno que incluya las fuentes de error más relevantes, dada la definición del procedimiento de evaluación y la interpretación propuesta) tiende a ser más informativo que un error estándar menos completo. Sin embargo, las restricciones prácticas suelen impedir estas clases de estudios que arrojarían información sobre todas las posibles fuentes de error, y en esos casos, es más informativo evaluar las fuentes de error que probablemente tengan el mayor impacto.

Las interpretaciones de los puntajes de una prueba pueden clasificarse ampliamente como *relativas* o *absolutas*. Las interpretaciones relativas transmiten la situación de un individuo o grupo dentro de una población de referencia. Las interpretaciones absolutas relacionan el estado de un individuo o grupo respecto de estándares de desempeño definidos. El error estándar no es

el mismo para los dos tipos de interpretaciones. Cualquier fuente de error que sea la misma para todos los individuos no contribuye al error relativo, pero puede contribuir al error absoluto.

Los coeficientes de confiabilidad conformes a normas tradicionales se desarrollaron para evaluar la precisión con la que los puntajes de la prueba estiman la situación relativa de individuos examinados en la misma escala, y evalúan la confiabilidad/precisión en términos de la relación de la varianza de puntaje verdadero respecto de la varianza de puntaje observado. A medida que se ha expandido la variedad de usos de los puntajes de prueba y se han extendido los contextos de uso (p. ej., categorización de diagnóstico, la evaluación de programas educativos), el rango de índices que se usan para evaluar la confiabilidad/precisión también ha aumentado para incluir índices para diversas clases de puntajes de cambio y puntajes de diferencia, índices de coherencia de decisiones, e índices apropiados para evaluar la precisión de las medias de grupos.

Algunos índices de precisión, especialmente errores estándares y errores estándares condicionales, también dependen de la escala en la que se reportan. Un índice expresado en términos de puntajes brutos o de estimaciones de TRI del nivel de rasgo puede transmitir una percepción muy diferente del error si se vuelve a expresar en términos de puntajes de escala. Por ejemplo, para la escala de puntajes brutos, el error estándar condicional puede parecer alto en un nivel de puntaje y bajo en otro, pero cuando los errores estándares condicionales se reexpresan en unidades de puntajes de escala, pueden surgir tendencias bastante diferentes en precisión comparativa.

Coherencia de decisiones

Cuando la finalidad de la medición es la clasificación, algunos errores de medida son más graves que otros. Los examinandos que están muy por encima o muy por debajo del puntaje de corte establecido para aprobar/reprobar o para elegibilidad para un programa especial pueden tener error considerable en sus puntajes observados sin ningún efecto en sus decisiones de clasificación. Es más probable que los errores de medida para individuos examinados cuyos puntajes verdaderos se acercan al puntaje de corte conduzcan a errores de clasificación. La elección de las técnicas utilizadas para cuantificar la confiabilidad/precisión debería tener en cuenta estas circunstancias. Esto puede hacerse reportando el error estándar condicional en la proximidad del puntaje de corte o los índices de coherencia/exactitud de decisiones (p. ej., porcentaje de decisiones correctas, kappa de Cohen), que varían como funciones tanto de la confiabilidad/precisión del puntaje como de la ubicación del puntaje de corte.

La *coherencia de decisiones* se refiere a la medida en que las clasificaciones observadas de individuos examinados sería la misma entre replicaciones del procedimiento de evaluación. La *exactitud de decisiones* se refiere a la medida en que las clasificaciones observadas de individuos examinados basadas en los resultados de una sola replicación concordarían con su estado de clasificación verdadero. Hay métodos estadísticos disponibles para calcular índices tanto para coherencia de decisiones como para exactitud de decisiones. Estos métodos evalúan la coherencia o exactitud de clasificaciones más que la coherencia en los puntajes per se. Obsérvese que el grado de coherencia o concordancia en la clasificación del individuo examinado es específico del puntaje de corte empleado y su ubicación dentro de la distribución de puntajes.

Confiabilidad/precisión de medias de grupos

Las estimaciones de puntajes medios (o promedio) de grupos (o proporciones en ciertas categorías) involucran fuentes de error que son diferentes de las que operan a nivel individual. Dichas estimaciones suelen utilizarse como medidas de efectividad de programas (y, en algunos sistemas de rendición de cuentas en materia educativa, pueden usarse para evaluar la efectividad de escuelas y profesores).

Al evaluar el desempeño grupal estimando el desempeño medio o mejora media en el

desempeño para muestras del grupo, la variación debida al muestreo de personas puede ser una fuente de error importante, en especial si los tamaños de la muestra son pequeños. En la medida en que diferentes muestras del grupo de interés (p. ej., todos los estudiantes que usan determinados materiales educativos) arrojen resultados diferentes, las conclusiones sobre el resultado esperado entre todos los estudiantes en el grupo (incluyendo los que podrían unirse al grupo en el futuro) son inciertas. Para muestras grandes, la variabilidad debida al muestreo de personas en las estimaciones de las medias del grupo puede ser bastante pequeña. Sin embargo, en casos en que las muestras de personas no son muy grandes (p. ej., en la evaluación del rendimiento medio de estudiantes en una sola aula o la satisfacción expresada promedio de muestras de clientes en un programa clínico), el error asociado con el muestreo de personas puede ser un componente importante del error general. Puede ser una fuente de error significativa en inferencias sobre programas incluso si existe un alto grado de precisión en los puntajes individuales de la prueba.

Los errores estándares para puntajes individuales no son medidas apropiadas de la precisión de los promedios del grupo. Una estadística más apropiada es el error estándar para las estimaciones de las medias del grupo.

Documentación de la confiabilidad/precisión

Por lo general, los desarrolladores y distribuidores de pruebas tienen la responsabilidad principal de obtener y reportar evidencia de confiabilidad/precisión (p. ej., errores estándares apropiados, coeficientes de confiabilidad o generabilidad, o funciones de información de la prueba). El usuario de la prueba debe tener dichos datos para hacer una elección informada entre enfoques de medición alternativos y por lo general podrá realizar estudios de confiabilidad/precisión adecuados antes del uso operativo de un instrumento.

En algunos casos, no obstante, los usuarios locales de un procedimiento de prueba o evaluación deben aceptar al menos responsabilidad parcial de documentar la precisión de medida. Esta obligación se mantiene cuando una de las finalidades principales de la medición es clasificar estudiantes usando estándares de desempeño desarrollados localmente, o clasificar a los individuos examinados dentro de la población local. También se mantiene cuando los usuarios deben basarse en evaluadores locales que están capacitados para usar las rúbricas de puntajes proporcionadas por el desarrollador de la prueba. En esos contextos, los factores locales pueden afectar sustancialmente la magnitud de la varianza de error y la varianza de puntajes observados. Por lo tanto, la confiabilidad/precisión de puntajes puede diferir apreciablemente de la reportada por el desarrollador.

Las evaluaciones de confiabilidad/precisión reportadas deben identificar las posibles fuentes de error para el programa de evaluación, dados los usos propuestos de los puntajes. Estas posibles fuentes de error pueden luego evaluarse en términos de investigación reportada previamente, nuevos estudios empíricos o análisis de los motivos para suponer que es probable que una posible fuente de error sea insignificante y, por lo tanto, pueda ignorarse.

El reporte de índices de confiabilidad/precisión solo —con escaso detalle respecto de los métodos usados para estimar los índices reportados, la naturaleza del grupo del que se derivaron los datos, y las condiciones en las que se obtuvieron los datos— constituye documentación inadecuada. Las declaraciones generales al efecto de que una prueba sea "confiable" o de que sea "suficientemente confiable para permitir interpretaciones de puntajes individuales" casi nunca, o nunca, son aceptables. Es el usuario quien debe asumir la responsabilidad de determinar si los puntajes son suficientemente fiables para justificar usos e interpretaciones previstos para usos particulares. No obstante, los constructores y editores de pruebas están obligados a proporcionar datos suficientes para que los juicios informados sean posibles.

Si los puntajes deben usarse para clasificación, son útiles los índices de coherencia de decisiones además de las estimaciones de la confiabilidad/precisión de los puntajes. Si es probable que las

medias del grupo tengan un rol sustancial en el uso de los puntajes, la confiabilidad/precisión de estos puntajes medios debe reportarse.

Como se destaca en los comentarios anteriores, no existe un único enfoque preferido para la cuantificación de la confiabilidad/precisión. Ningún índice solo transmite adecuadamente toda la información relevante. Ningún método de investigación es óptimo en todas las situaciones, ni el desarrollador de la prueba se limita a un único enfoque para cualquier instrumento. La elección de técnicas de estimación y el nivel mínimo aceptable de cualquier índice continúan siendo un asunto de juicio profesional.

ESTÁNDARES DE CONFIABILIDAD/PRECISIÓN

Los estándares en este capítulo comienzan con un estándar global (numerado 2.0), que se ha diseñado para transmitir la intención central o enfoque principal del capítulo. El estándar global también puede verse como el principio rector del capítulo, y es aplicable a todas las pruebas y usuarios de pruebas. Todos los estándares posteriores se han separado en ocho unidades temáticas denominadas de la siguiente manera:

1. Especificaciones para replicaciones del procedimiento de evaluación
2. Evaluación de la confiabilidad/precisión
3. Coeficientes de confiabilidad/generabilidad
4. Factores que afectan la confiabilidad/precisión
5. Errores estándares de medida
6. Coherencia de decisiones
7. Confiabilidad/precisión de medias de grupos
8. Documentación de la confiabilidad/precisión

Estándar 2.0

Se debe proporcionar evidencia apropiada de confiabilidad/precisión para la interpretación de cada uso previsto de los puntajes.

Comentario: La forma de la evidencia (coeficiente de confiabilidad o generabilidad, función de información, error estándar condicional, índice de coherencia de decisiones) para la confiabilidad/precisión debe ser apropiada para los usos previstos de los puntajes, la población involucrada y los modelos psicométricos utilizados para derivar los puntajes. Se requiere un grado de confiabilidad/precisión más alto para usos de puntajes que tienen consecuencias más significativas para los examinandos. Al contrario, un grado más bajo puede ser aceptable cuando una decisión basada en el puntaje de una prueba es reversible o depende de la corroboración de otras fuentes de información.

Unidad 1. Especificaciones para replicaciones del procedimiento de evaluación

Estándar 2.1

El rango de replicaciones sobre el que se evalúa la confiabilidad/precisión debe indicarse claramente, junto con una justificación para la elección de esta definición, dada la situación de evaluación.

Comentario: Para cualquier programa de evaluación, es probable que algunos aspectos del procedimiento de evaluación (p. ej., límites de tiempo y disponibilidad de recursos como libros, calculadoras y computadoras) sean fijos, y se permitirá que algunos aspectos varíen de una administración a otra (p. ej., tareas o estímulos específicos, contextos de evaluación, evaluadores y, posiblemente, ocasiones). Cualquier administración de la prueba que mantenga condiciones fijas e involucre muestras aceptables de las condiciones que se permita variar se consideraría una replicación legítima del procedimiento de evaluación. Como primer paso en la evaluación de la confiabilidad/precisión de los puntajes obtenidos con un procedimiento de evaluación, es importante identificar el rango de condiciones de varias clases que se permitan variar, y sobre qué puntajes deben generalizarse.

Estándar 2.2

La evidencia proporcionada para la confiabilidad/precisión de los puntajes debe ser coherente con el dominio de replicaciones asociadas con los procedimientos de evaluación, y con las interpretaciones previstas para uso de los puntajes de la prueba.

Comentario: La evidencia de confiabilidad/ precisión debe ser coherente con el diseño de los procedimientos de evaluación y con las interpretaciones propuestas para uso de los puntajes de la prueba. Por ejemplo, si la prueba puede tomarse en cualquiera de una serie de ocasiones, y la interpretación supone que los puntajes son invariantes en estas ocasiones, entonces cualquier variabilidad en los puntajes en esas ocasiones es una posible fuente de error. Si se permite que las tareas o estímulos varíen entre formularios alternativos de la prueba, y los desempeños observados son tratados como una muestra de un dominio de tareas similares, la variabilidad en los puntajes de un formulario a otro se consideraría un error. Si se utilizan evaluadores para asignar puntajes a respuestas, la variabilidad en los puntajes entre evaluadores cualificados es una fuente de error. Diferentes fuentes de error pueden evaluarse en un solo coeficiente o error estándar, o pueden evaluarse por separado, pero todas deben abordarse de alguna manera. Los reportes de confiabilidad/precisión deben especificar las posibles fuentes de error incluidas en los análisis.

Unidad 2. Evaluación de la confiabilidad/ precisión

Estándar 2.3

Para cada puntaje total, subpuntaje o combinación de puntajes que deba interpretarse, deben reportarse estimaciones de índices relevantes de confiabilidad/ precisión.

Comentario: No es suficiente reportar estimaciones de confiabilidades y errores estándares de medida solo para puntajes totales cuando también se interpretan subpuntajes. La coherencia entre formularios y día a día de los puntajes totales en una prueba puede ser aceptablemente alta, aunque los subpuntajes pueden tener confiabilidad inaceptablemente baja, dependiendo de cómo se definan y utilicen. Se debe suministrar a los usuarios datos de confiabilidad para todos los puntajes a interpretarse, y esos datos deben ser lo suficientemente detallados para permitir que los usuarios juzguen si los puntajes son lo suficientemente precisos para las interpretaciones previstas para su uso. Los puntajes compuestos formados a partir de subpruebas seleccionadas dentro de una batería de pruebas suelen proponerse para fines predictivos y de diagnóstico. Los usuarios necesitan información sobre la confiabilidad de esos puntajes compuestos.

Estándar 2.4

Cuando la interpretación de puntajes de una prueba destaca diferencias entre dos puntajes observados de un individuo o dos promedios de un grupo, deben proporcionarse datos de confiabilidad/precisión, incluyendo errores estándares, para dichas diferencias.

Comentario: Las diferencias de puntajes observados se utilizan para diversos fines. Los logros de rendimiento suelen ser de interés para grupos y para individuos. En algunos casos, la confiabilidad/precisión de puntajes de cambio puede ser mucho más baja que las confiabilidades de los puntajes separados involucrados. Las diferencias entre puntajes verbales y de desempeño en pruebas de inteligencia o capacidad académica suelen emplearse en el diagnóstico de deterioro cognitivo y problemas de aprendizaje. Las inferencias psicodiagnósticas suelen hacerse a partir de diferencias entre puntajes de subpruebas. Las baterías de aptitud y rendimiento, inventarios de interés y evaluaciones de personalidad se utilizan comúnmente para identificar y cuantificar las fortalezas y debilidades relativas, o el patrón de niveles de rasgos, de un examinando. Cuando la interpretación de los puntajes de la prueba se centra en los valores altos y bajos en el perfil de puntajes de la prueba del individuo examinado, la confiabilidad de las diferencias de puntajes es crítica.

Estándar 2.5

Los procedimientos de estimación de confiabilidad deben ser coherentes con la estructura de la prueba.

Comentario: Un solo puntaje total puede calcularse en pruebas que son multidimensionales. El puntaje total de una prueba que es sustancialmente multidimensional debe tratarse como un puntaje compuesto. Si una estimación de coherencia interna de la confiabilidad del puntaje total se obtiene mediante el procedimiento dividido, las mitades deben ser comparables en contenido y características estadísticas.

En procedimientos de pruebas adaptables, el conjunto de tareas incluidas en la prueba y el secuenciamiento de tareas se hacen a medida del examinando, utilizando algoritmos basados en modelos. En este contexto, la confiabilidad/precisión puede estimarse utilizando simulaciones basadas en el modelo. Para las pruebas adaptables, los errores estándares condicionales basados en modelos pueden ser particularmente útiles y apropiados para evaluar la adecuación técnica del procedimiento.

Unidad 3. Coeficientes de confiabilidad/generabilidad

Estándar 2.6

Un coeficiente de confiabilidad o generabilidad (o error estándar) que aborda un tipo de variabilidad no debe interpretarse como intercambiable con índices que abordan otros tipos de variabilidad, a menos que sus definiciones de error de medida puedan considerarse equivalentes.

Comentario: Los coeficientes de coherencia interna, formularios alternativos y test-retest no deben considerarse equivalentes, dado que cada uno incorpora una definición única de error de medida. Las varianzas de error derivadas mediante teoría de respuesta al ítem por lo general no son equivalentes a las varianzas de error estimadas mediante otros enfoques. Los desarrolladores de la prueba deben indicar las fuentes de error que se reflejan en los coeficientes de confiabilidad o generabilidad reportados, y las que son ignoradas por estos.

Estándar 2.7

Cuando el juicio subjetivo entre en la calificación de la prueba, debe proporcionarse evidencia tanto de coherencia entre los evaluadores en la calificación como de coherencia dentro del individuo examinado en mediciones repetidas. Debe hacerse una distinción clara entre datos de confiabilidad basados en (a) paneles independientes de evaluadores que califican los mismos desempeños o productos, (b) un solo panel que califica desempeños sucesivos o nuevos productos, y (c) paneles independientes que califican desempeños sucesivos o nuevos productos.

Comentario: Las variaciones entre tareas en la calidad del desempeño de un individuo examinado y las incoherencias entre los evaluadores en la calificación representan fuentes independientes de error de medida. Los reportes de estudios de confiabilidad/precisión deben aclarar cuáles de esas fuentes se reflejan en los datos. Los estudios de generabilidad y los análisis de componentes de varianza pueden ser útiles para estimar las varianzas de error que surgen de cada fuente de error. Estos análisis pueden proporcionar estimaciones de varianza de error separadas para tareas, para jueces, y para ocasiones dentro del período de tiempo de estabilidad de rasgos. Debe proporcionarse información sobre las cualificaciones y capacitación de los jueces utilizados en los estudios de confiabilidad. La concordancia entre los evaluadores o entre los observadores puede ser particularmente importante para calificaciones y datos observacionales que involucran discriminaciones sutiles. Debe observarse, no obstante, que cuando los evaluadores evalúan positivamente características correlacionadas, una evaluación favorable o desfavorable de un rasgo puede influir en sus opiniones de otros rasgos. Además, la alta coherencia entre los evaluadores no implica alta coherencia del individuo examinado de una tarea a otra. Por

lo tanto, la concordancia entre los evaluadores no garantiza alta confiabilidad de los puntajes del individuo examinado.

Unidad 4. Factores que afectan la confiabilidad/precisión

Estándar 2.8

Cuando las pruebas de respuesta construida se califican localmente, los datos de confiabilidad/precisión deben reunirse y reportarse para la calificación local cuando hay disponibles muestras de tamaño adecuado.

Comentario: Por ejemplo, muchos programas de evaluación a nivel estatal dependen de calificaciones locales de ensayos, ejercicios de respuesta construida, y tareas de desempeño. Los análisis de confiabilidad/precisión pueden indicar que se necesita capacitación adicional de los calificadores y, por consiguiente, deben ser una parte integral de la supervisión del programa. Los datos de confiabilidad/precisión deben comunicarse solo cuando son suficientes para arrojar resultados sólidos desde el punto de vista estadístico y son coherentes con las obligaciones de privacidad aplicables.

Estándar 2.9

Cuando una prueba está disponible en versiones largas y cortas, la evidencia de confiabilidad/precisión debe reportarse para puntajes en cada versión, preferentemente basada en administraciones independientes de cada versión con muestras independientes de examinandos.

Comentario: La confiabilidad/precisión de puntajes en cada versión se evalúa mejor a través de una administración independiente de cada una, utilizando los límites de tiempo designados. Pueden utilizarse modelos psicométricos para estimar la confiabilidad/precisión de una versión más corta (o más larga) de una prueba existente, basados en datos de una administración de la prueba existente. Sin embargo, estos modelos por lo general hacen suposiciones que es posible que no se cumplan (p. ej., que los ítems en la prueba existente y los ítems que se agregarán o quitarán son todos muestreados de manera aleatoria de un solo dominio). Los efectos del contexto son corrientes en las pruebas de desempeño máximo, y la versión corta de una prueba estandarizada a menudo comprende una muestra no aleatoria de ítems de la versión completa. Como resultado, es posible que el valor predicho de la confiabilidad/precisión no proporcione una estimación muy buena del valor real, y por lo tanto, cuando sea viable, la confiabilidad/precisión de ambos formularios debería evaluarse directa e independientemente.

Estándar 2.10

Cuando se permitan variaciones significativas en las pruebas o procedimientos de administración de pruebas, deben proporcionarse análisis de confiabilidad/precisión separados para puntajes producidos en cada variación importante si hay disponibles tamaños de la muestra adecuados.

Comentario: Para hacer que una prueba sea accesible para todos los individuos examinados, los editores o usuarios de la prueba podrían autorizar, o podría requerirse legalmente que se autoricen, adecuaciones o modificaciones en los procedimientos que se especifican para la administración de una prueba. Por ejemplo, pueden usarse versiones en audio o en letra grande para los examinandos que tienen problemas de la vista. Cualquier alteración en los materiales o procedimientos de evaluación estándares puede tener un impacto en la confiabilidad/precisión de los puntajes resultantes y por lo tanto, en la medida en que sea viable, la confiabilidad/precisión debe examinarse para todas las versiones de la prueba y procedimientos de evaluación.

Estándar 2.11

Los editores de la prueba deben proporcionar estimaciones de confiabilidad/precisión tan pronto como sea viable para cada subgrupo relevante para el que se recomienda la prueba.

Comentario: Reportar estimaciones de confiabilidad/precisión para subgrupos relevantes es útil en muchos contextos, pero es especialmente importante si la interpretación de puntajes involucra inferencias dentro del grupo (p. ej., en términos de normas del subgrupo). Por ejemplo, los usuarios de la prueba que trabajan con un subgrupo lingüístico y cultural específico o con individuos que tienen una discapacidad en particular se beneficiarían con una estimación del error estándar para el subgrupo. Del mismo modo, la evidencia de que los niños de preescolar tienden a responder a estímulos de la prueba de una manera menos coherente que los niños mayores sería útil para los usuarios de la prueba que interpretan puntajes entre grupos etarios.

Al considerar la confiabilidad/precisión de puntajes de la prueba para subgrupos relevantes, es útil evaluar y reportar el error estándar de medida, así como cualquier coeficiente que se estime. Los coeficientes de confiabilidad y generabilidad pueden diferir sustancialmente cuando los subgrupos tienen varianzas diferentes en el constructo que se evalúa. Las diferencias en la variabilidad dentro del grupo tienden a tener menos impacto en el error estándar de medida.

Estándar 2.12

Si una prueba se propone para utilizarse en varios grados o en un rango de edades, y si se proporcionan normas separadas para cada grado o rango de edades, deben proporcionarse los datos de confiabilidad/precisión para cada edad o subgrupo de nivel de grado, no solo para todos los grados o edades combinados.

Comentario: Un coeficiente de confiabilidad o generabilidad basado en una muestra de individuos examinados que abarca varios grados o un rango amplio de edades en que los puntajes promedio aumentan en forma constante por lo general dará una impresión de confiabilidad/precisión falsamente inflada. Cuando una prueba tiene por objeto discriminar dentro de poblaciones de edades o grados, los coeficientes de confiabilidad o generabilidad y los errores estándares deben reportarse por separado para cada subgrupo.

Unidad 5. Errores estándares de medida

Estándar 2.13

El error estándar de medida, tanto general como condicional (si se reporta), debe proporcionarse en unidades de cada puntaje reportado.

Comentario: El error estándar de medida (general o condicional) que se reporta debe ser coherente con las escalas que se utilizan en el reporte de puntajes. Los errores estándares en unidades de puntajes de escala para las escalas utilizadas para reportar puntajes y/o para tomar decisiones son particularmente útiles para el usuario de la prueba típico. Los datos sobre desempeño del individuo examinado deben ser coherentes con las suposiciones incorporadas en cualquier modelo estadístico utilizado para generar puntajes de escala y estimar los errores estándares para esos puntajes.

Estándar 2.14

Cuando sea posible y corresponda, los errores estándares de medida condicionales deben reportarse en varios niveles de puntajes a menos que exista evidencia de que el error estándar es constante entre los niveles de puntajes. Cuando se especifican puntajes de corte para selección o clasificación, los errores estándares de medida deben reportarse en la proximidad de cada puntaje de corte.

Comentario: La estimación de errores estándares condicionales por lo general es viable con los tamaños de la muestra que se usan para análisis de confiabilidad/precisión. Si se supone que el error estándar es constante en un amplio rango de niveles de puntaje, debe presentarse la justificación para esta suposición. El modelo en el que se basa el cálculo de los errores estándares condicionales debe especificarse.

Estándar 2.15

Cuando existe evidencia creíble para esperar que los errores estándares de medida condicionales o funciones de información de prueba difieran sustancialmente para varios subgrupos, debe realizarse una investigación del alcance y el impacto de esas diferencias y reportarse tan pronto como sea viable.

Comentario: Si se encuentran diferencias, deben indicarse claramente en la documentación correspondiente. Además, si efectivamente existen diferencias sustanciales, el contenido de la prueba y los modelos de calificación deben examinarse para ver si hay alternativas legalmente aceptables que no den por resultado dichas diferencias.

Unidad 6. Coherencia de decisiones

Estándar 2.16

Cuando una prueba o combinación de medidas se utiliza para tomar decisiones de clasificación, deben proporcionarse estimaciones del porcentaje de examinandos que se clasificarían de la misma manera en dos replicaciones del procedimiento.

Comentario: Cuando un puntaje de prueba o puntaje compuesto se utiliza para tomar decisiones de clasificación (p. ej., aprobar/reprobar, niveles de rendimiento), el error estándar de medida en o cerca de los puntajes de corte tiene importantes implicaciones para la fiabilidad de esas decisiones. Sin embargo, el error estándar no puede traducirse en el porcentaje esperado de decisiones coherentes o exactas sin suposiciones sólidas sobre las distribuciones de errores de medida y puntajes verdaderos. Si bien la coherencia de decisiones suele estimarse a partir de la administración de un solo formulario, puede y debería estimarse directamente a través del uso de un enfoque de test-retest, si es coherente con los requisitos de seguridad de la prueba, y si se cumple la suposición de ausencia de cambio en el constructo y hay muestras adecuadas disponibles.

Unidad 7. Confiabilidad/precisión de medias de grupos

Estándar 2.17

Cuando los puntajes promedio de la prueba para grupos son el centro de la interpretación propuesta de los resultados de la prueba, los grupos evaluados por lo general deben considerarse como una muestra de una población más grande, incluso si se evalúan todos los individuos examinados disponibles en el momento de la medición. En esos casos, debe reportarse el error estándar de la media de los grupos, porque refleja variabilidad debida al muestreo de individuos examinados, así como variabilidad debida a error de medida individual.

Comentario: Los niveles generales de desempeño en varios grupos tienden a ser el centro en la evaluación de programas y sistemas de rendición de cuentas, y los grupos que son de interés incluyen a todos los estudiantes/clientes que podrían participar en el programa en algún período. Por lo tanto, los estudiantes en una clase o escuela en particular en el momento actual, los clientes actuales de un organismo de servicios sociales, y grupos análogos expuestos a un programa de interés por lo general constituyen una muestra en un sentido longitudinal. Presuntamente, grupos comparables de la misma población se repetirán en años futuros, dadas condiciones estáticas. Los factores que conducen a incertidumbre en las conclusiones sobre la efectividad del programa surgen del muestreo de personas así como del error de medida individual.

Estándar 2.18

Cuando la finalidad de la evaluación es medir el desempeño de grupos en lugar del de individuos, pueden asignarse aleatoriamente subconjuntos de ítems a diferentes submuestras de individuos examinados. Los datos se agregan entre submuestras y subconjuntos de ítems para obtener una medida del desempeño del grupo. Cuando se usan estos procedimientos para la

evaluación de programas o descripciones de poblaciones, los análisis de confiabilidad/precisión deben tener en cuenta el esquema de muestreo.

Comentario: Este tipo de programa de medición recibe el nombre de *muestreo de matriz*. Se ha diseñado para reducir el tiempo requerido de cada individuo examinado y aun así aumentar la cantidad total de ítems sobre los que pueden obtenerse datos. Este enfoque de evaluación proporciona el mismo tipo de información sobre desempeños de grupos que se obtendría si todos los individuos examinados hubieran realizado todos los ítems. Las estadísticas de confiabilidad/precisión deben reflejar el plan de muestreo utilizado con respecto a los individuos examinados e ítems.

Unidad 8. Documentación de la confiabilidad/precisión

Estándar 2.19

Cada método de cuantificación de la confiabilidad/precisión de puntajes debe describirse claramente y expresarse en términos de estadísticas apropiadas para el método. Deben reportarse los procedimientos de muestreo utilizados para seleccionar examinandos para análisis de confiabilidad/precisión y las estadísticas descriptivas sobre estas muestras, con sujeción a las obligaciones de privacidad cuando corresponda.

Comentario: La información sobre el método de recopilación de datos, tamaños de las muestras, medias, desviaciones estándares y características demográficas de los grupos evaluados ayuda a los usuarios a juzgar en qué medida los datos reportados se aplican a sus propias poblaciones de individuos examinados. Si se utiliza el enfoque de test-retest o de formularios alternativos, debe indicarse el intervalo entre administraciones.

Dado que hay muchas maneras de estimar la confiabilidad/precisión, y cada una está influenciada por diferentes fuentes de error de medida, es inaceptable decir simplemente: "La confiabilidad/precisión de puntajes en la prueba X es 0,90". Un enunciado mejor sería: "El coeficiente de confiabilidad de 0,90 reportado para puntajes en la prueba X se obtuvo correlacionando puntajes de los formularios A y B administrados en días consecutivos. Los datos se basaron en una muestra de 400 estudiantes de 10.° grado de cinco escuelas suburbanas de clase media en el estado de Nueva York. El desglose demográfico de este grupo fue el siguiente:...". En algunos casos, por ejemplo, cuando se involucran tamaños pequeños de la muestra o datos especialmente confidenciales, las restricciones legales aplicables que rigen la privacidad pueden limitar el nivel de información que debería divulgarse.

Estándar 2.20

Si los coeficientes de confiabilidad se ajustan para restricción de rango o variabilidad, deben informarse el procedimiento de ajuste y los coeficientes tanto ajustados como no ajustados. Deben presentarse las desviaciones estándares del grupo efectivamente evaluado y de la población de destino, así como la justificación del ajuste.

Comentario: La aplicación de una corrección para restricción en la variabilidad supone que la muestra disponible no es representativa (en términos de variabilidad) de la población de examinandos a la que podría esperarse que los usuarios generalicen. La justificación para la corrección debe considerar lo apropiado de esa generalización. Las fórmulas de ajuste que suponen constancia en el error estándar entre niveles de puntajes no deben usarse a menos que la constancia pueda defenderse.

3. IMPARCIALIDAD EN LAS PRUEBAS

ANTECEDENTES

Este capítulo aborda la importancia de la imparcialidad como cuestión fundamental en la protección de los examinandos y usuarios de pruebas en todos los aspectos de evaluación. El término *imparcialidad* no tiene un solo significado técnico y se utiliza de muchas maneras diferentes en el debate público. Es posible que individuos avalen la imparcialidad en las pruebas como una meta social deseable, y aun así lleguen a conclusiones bastante diferentes sobre la imparcialidad de un programa de evaluación determinado. Una consideración completa del tema exploraría las múltiples funciones de las pruebas en relación con sus numerosas metas, incluyendo la meta amplia de lograr igualdad de oportunidades en nuestra sociedad. Consideraría las propiedades técnicas de las pruebas, las maneras en que se reportan y utilizan los resultados de las pruebas, los factores que afectan la validez de las interpretaciones de puntajes y las consecuencias del uso de las pruebas. Un análisis completo de imparcialidad en las pruebas también examinaría las regulaciones, leyes y la jurisprudencia que rigen el uso de pruebas y las reparaciones para prácticas de evaluación perjudiciales. Los *Estándares* no pueden esperar tratar adecuadamente todas estas amplias cuestiones, algunas de las cuales han suscitado fuerte desacuerdo entre especialistas en evaluación y otras partes interesadas en la evaluación. Nuestro enfoque debe limitarse en este caso a delinear los aspectos de las pruebas, la evaluación y el uso de pruebas que se relacionan con la imparcialidad según se describe en este capítulo, que son la responsabilidad de quienes desarrollan, usan e interpretan los resultados de las pruebas, y sobre los cuales existe acuerdo profesional y técnico general.

La imparcialidad es una cuestión de validez fundamental y requiere atención en todas las etapas del desarrollo y uso de las pruebas. En versiones anteriores de los *Estándares*, la imparcialidad y la evaluación de individuos de subgrupos específicos de examinandos, como individuos con discapacidades e individuos con características lingüísticas y culturales diversas, se presentaron en capítulos separados. En la versión actual de los *Estándares*, estas cuestiones se presentan en un solo capítulo para hacer hincapié en que la imparcialidad para todos los individuos en la población prevista de examinandos es un interés primordial y fundamental, y que se aplican principios comunes en la respuesta a características de los examinandos que podrían interferir con la validez de la interpretación de los puntajes de la prueba. Esto no quiere decir que la respuesta a características de los examinandos sea la misma para individuos de subgrupos diversos como los definidos por raza, origen étnico, género, cultura, idioma, edad, discapacidad o nivel socioeconómico, sino que esas respuestas deberían ser sensibles a características individuales que de otro modo comprometerían la validez. No obstante, como se analizó en la introducción, es importante tener presente, al usar los *Estándares*, que la aplicabilidad depende del contexto. Por ejemplo, posibles amenazas a la validez de la prueba para individuos examinados con competencia limitada en inglés son diferentes de las correspondientes a individuos examinados con discapacidades. Además, las amenazas a la validez pueden diferir incluso para individuos dentro del mismo subgrupo. Por ejemplo, individuos con discapacidades específicas diversas constituyen el subgrupo de "individuos con discapacidades" e individuos examinados clasificados como con "competencia limitada en inglés" representan un rango de niveles de competencia en un idioma, nivel educativo y características culturales y experiencias previas. Además, la equivalencia del constructo que se evalúa es un tema central en la imparcialidad, tanto si el contexto es, por ejemplo, individuos con discapacidades especiales diversas, individuos con competencia limitada en inglés o individuos de diversos países y culturas.

Al igual que en versiones anteriores de los *Estándares*, el capítulo actual aborda el sesgo de medición como una amenaza central a la imparcialidad en las pruebas. Sin embargo, también incorpora dos conceptos importantes que han surgido en la bibliografía, en especial en la bibliografía relacionada con educación, para minimizar el sesgo y por consiguiente aumentar la imparcialidad. El primer concepto es la *accesibilidad*, la noción de que todos los examinandos deben tener la oportunidad sin obstáculos de demostrar su situación respecto de los constructos sometidos a medición. Por ejemplo, es posible que los individuos con competencia limitada en inglés no se diagnostiquen adecuadamente en el constructo de destino de un examen clínico si la evaluación requiere un nivel de competencia en inglés que no poseen. De manera similar, la letra estándar y algunos formatos electrónicos pueden constituir desventajas para los individuos examinados con problemas de la vista y algunos adultos mayores que necesitan aumento para leer, y la desventaja se considera injusta si la agudeza visual es relevante para el constructo sometido a medición. Estos ejemplos muestran cómo el acceso al constructo que mide la prueba puede verse impedido por características y/o habilidades que no se relacionan con el constructo previsto y que, por ende, pueden limitar la validez de las interpretaciones de los puntajes para los usos previstos para determinados individuos y/o subgrupos en la población prevista de examinandos. La accesibilidad es un requisito legal en algunos contextos de evaluación.

El segundo nuevo concepto contenido en este capítulo es el de *diseño universal*. El diseño universal es un enfoque hacia el diseño de pruebas que busca maximizar la accesibilidad para todos los examinandos previstos. El diseño universal, según se describe con mayor profundidad más adelante en este capítulo, requiere que los desarrolladores de la prueba sean claros sobre los constructos sometidos a medición, incluyendo el objetivo de la evaluación, el fin para el que se usarán los puntajes, las inferencias que se harán a partir de los puntajes, y las características de los individuos examinados y los subgrupos de la población prevista de la prueba que podrían influir en el acceso. Los ítems y tareas de la prueba pueden entonces diseñarse y desarrollarse intencionalmente desde el comienzo para reflejar el constructo previsto, minimizar las características irrelevantes del constructo que de otro modo podrían impedir el desempeño de los grupos previstos de individuos examinados, y para maximizar, en la medida posible, el acceso para tantos individuos examinados como sea posible en la población prevista, independientemente de la raza, origen étnico, edad, género, nivel socioeconómico, discapacidad o características de idioma o culturales.

Aun así, para algunos individuos en algunos contextos de prueba y para algunos fines —como se describe más adelante— es posible que exista la necesidad de adaptaciones adicionales de la prueba para responder a características individuales que de otro modo limitarían el acceso al constructo tal como se mide. Algunos ejemplos son la creación de una versión de la prueba en sistema braille, permitir tiempo adicional de evaluación, y proporcionar traducciones o simplificación del lenguaje de la prueba. Cualquier adaptación de la prueba debe considerarse atentamente, ya que algunas adaptaciones pueden alterar el constructo previsto de la prueba. Responder a características individuales que de otro modo impedirían el acceso y mejorar la validez de las interpretaciones de los puntajes de la prueba para los usos previstos son dos consideraciones para respaldar la imparcialidad.

En resumen, este capítulo interpreta la imparcialidad como la capacidad de respuesta a características individuales y contextos de evaluación de modo que los puntajes de la prueba arrojen interpretaciones válidas para los usos previstos. La definición de imparcialidad de los *Estándares* es a menudo más amplia de lo que se requiere legalmente. Una prueba que es imparcial dentro del significado de los *Estándares* refleja los mismos constructos para todos los examinandos, y los puntajes de esta tienen el mismo significado para todos los individuos en la población prevista; una prueba imparcial no favorece ni desfavorece a algunos individuos debido a características irrelevantes para el constructo previsto. En la medida posible, deben considerarse las características de todos los individuos en la población prevista de la prueba,

incluyendo las asociadas con raza, origen étnico, género, edad, nivel socioeconómico, o características lingüísticas o culturales, a lo largo de todas las etapas de desarrollo, administración, calificación, interpretación y uso, de modo que puedan reducirse los obstáculos a la evaluación imparcial. Al mismo tiempo, los puntajes de la prueba deben arrojar interpretaciones válidas para los usos previstos, y es posible que diferentes contextos y usos de la prueba requieran diferentes enfoques hacia la imparcialidad. Por ejemplo, en las pruebas utilizadas para fines de selección, las adaptaciones a procedimientos estandarizados que aumentan la accesibilidad para algunos individuos, pero cambian el constructo sometido a medición podrían reducir la validez de las inferencias de los puntajes para los fines previstos y favorecer injustamente a quienes reúnen los requisitos para adaptación en relación con los que no lo hacen. Por el contrario, para fines de diagnóstico en medicina y educación, adaptar una prueba para aumentar la accesibilidad para algunos individuos podría aumentar la exactitud del diagnóstico.

Estas cuestiones se analizan en las secciones a continuación y se representan en los estándares que siguen a la introducción del capítulo.

Puntos de vista generales de la imparcialidad

El primer punto de vista de la imparcialidad en las pruebas que se describe en este capítulo establece el principio de trato justo y equitativo para todos los examinandos durante el proceso de evaluación. El segundo, tercer y cuarto punto de vista presentados aquí hacen hincapié en cuestiones de imparcialidad en la calidad de la medición: imparcialidad como falta o ausencia de sesgo de medición, imparcialidad como acceso a los constructos medidos, e imparcialidad como validez de las interpretaciones de los puntajes individuales de la prueba para el uso o los usos previstos.

Imparcialidad en el trato durante el proceso de evaluación

Independientemente de la finalidad de la prueba, la meta de la imparcialidad es maximizar, en la medida posible, la oportunidad para que los examinandos demuestren su situación respecto del o de los constructos que la prueba tiene por objeto medir. Tradicionalmente, la estandarización cuidadosa de las pruebas, las condiciones de administración y los procedimientos de calificación han ayudado a asegurar que los examinandos tengan contextos comparables en los que demostrar sus capacidades o atributos sometidos a medición. Por ejemplo, se implementan instrucciones uniformes, límites de tiempo especificados, arreglos especificados en las salas, uso de monitores, y uso de procedimientos de seguridad coherentes de modo que las diferencias en las condiciones de administración no influyan involuntariamente en el desempeño de algunos examinandos respecto de otros. De manera similar, las cuestiones sobre imparcialidad en el trato pueden requerir, para algunas pruebas, que todos los examinandos tengan administradores de pruebas cualificados con quienes puedan comunicarse y sentirse cómodos en la medida posible. En los casos que involucren tecnología, es importante que los individuos examinados hayan tenido exposición previa similar a la tecnología y que los equipos proporcionados a todos los examinandos tengan una velocidad de procesamiento similar y proporcionen claridad y tamaño similares para las imágenes y otros medios. Los procedimientos para la administración estandarizada de una prueba deben ser documentados con detenimiento por el desarrollador de la prueba y el administrador de la prueba debe seguirlos cuidadosamente.

Si bien la estandarización ha sido un principio fundamental para asegurar que todos los individuos examinados tengan la misma oportunidad de demostrar su situación respecto del constructo que la prueba tiene por objeto medir, a veces se necesita flexibilidad para proporcionar oportunidades esencialmente equivalentes para algunos examinandos. En esos casos, es posible que aspectos de un proceso de evaluación estandarizado que no plantean un desafío en particular para la mayoría de los examinandos eviten que grupos o individuos específicos demuestren con exactitud su situación con respecto al constructo de interés. Por ejemplo, pueden surgir desafíos debido a la discapacidad, origen cultural, característica

lingüística, raza, origen étnico, nivel socioeconómico de un individuo examinado, limitaciones que pueden venir con la edad, o alguna combinación de estos u otros factores. En algunos casos, puede alcanzarse mayor comparabilidad de puntajes si los procedimientos estandarizados se cambian para abordar las necesidades de grupos o individuos específicos sin ningún efecto adverso en la validez o confiabilidad de los resultados obtenidos. Por ejemplo, pueden proporcionarse un formulario de prueba en sistema braille, una hoja de respuestas en letra grande o un lector de pantalla para permitir que quienes tienen problemas de la vista obtengan acceso más equitativo al contenido de la prueba. Las consideraciones legales también pueden influir en cómo abordar necesidades individualizadas.

Imparcialidad como falta de sesgo de medición

Las características de la prueba propiamente dicha que no se relacionen con el constructo sometido a medición, o la manera en que se utiliza la prueba, pueden en ocasiones dar por resultado diferentes significados para los puntajes obtenidos por los miembros de subgrupos identificables. Por ejemplo, se dice que ocurre *funcionamiento diferencial de los ítems* (DIF, por sus siglas en inglés) cuando examinandos con iguales capacidades difieren en sus probabilidades de responder a un ítem de la prueba correctamente como una función de pertenencia a un grupo. El DIF puede evaluarse de diversas maneras. La detección de DIF no siempre indica sesgo en un ítem; es necesario que haya una explicación adecuada sustancial para que el DIF justifique la conclusión de que el ítem está sesgado. El *funcionamiento diferencial de la prueba* (DTF, por sus siglas en inglés) se refiere a diferencias en el funcionamiento de las pruebas (o conjuntos de ítems) para diferentes grupos especialmente definidos. Cuando ocurre DTF, los individuos de diferentes grupos que tienen la misma situación respecto de la característica evaluada por la prueba no tienen el mismo puntaje de la prueba esperado.

El término *sesgo predictivo* puede usarse cuando se encuentra evidencia de que existen diferencias en los patrones de asociaciones entre puntajes de la prueba y otras variables para diferentes grupos, lo que trae consigo preocupaciones sobre sesgo en las inferencias extraídas del uso de los puntajes de la prueba. La predicción diferencial se examina utilizando análisis de regresión. Un enfoque examina las diferencias de pendiente e intersección entre dos grupos de destino (p. ej., individuos examinados afroamericanos e individuos examinados caucásicos), mientras que otro examina desviaciones sistemáticas de una línea de regresión común para cualquier número de grupos de interés. Ambos enfoques proporcionan información valiosa al examinar predicción diferencial. Los coeficientes de correlación proporcionan evidencia inadecuada a favor o en contra de una hipótesis de predicción diferencial si se determina que los grupos tienen medias y varianzas desiguales en la prueba y en el criterio.

Cuando evidencia creíble indica posible sesgo en la medición (es decir, falta de significado coherente del constructo entre grupos, DIF, DTF) o sesgo en relaciones predictivas, estas posibles fuentes de sesgo deben investigarse de manera independiente porque la presencia o ausencia de una forma de dicho sesgo puede no tener relación con otras formas de sesgo. Por ejemplo, es posible que una prueba predictora no muestre niveles significativos de DIF, pero muestre diferencias de grupos en líneas de regresión en la predicción de un criterio. Si bien es importante advertir sobre la posibilidad de sesgo de medición para los subgrupos que se han definido como relevantes en la población prevista de la prueba, es posible que no sea viable investigar completamente todas las posibilidades, en especial en el contexto laboral. Por ejemplo, el número de miembros del subgrupo en la prueba de campo o población de normalización puede limitar la posibilidad de análisis empíricos estándares. En estos casos, la investigación previa, una justificación basada en el constructo y/o datos de pruebas similares pueden abordar las inquietudes relacionadas con posible sesgo en la medición. Además, y especialmente cuando existe evidencia creíble de posible sesgo, deben considerarse metodologías para muestras pequeñas. Por ejemplo, se puede examinar el posible sesgo para subgrupos relevantes mediante ensayos a pequeña

escala que utilizan laboratorios cognitivos y/o entrevistas o grupos focales para solicitar evidencia sobre la validez de interpretaciones hechas a partir de puntajes de la prueba.

Una cuestión relacionada es la medida en que el constructo que se evalúa tiene un significado equivalente entre los individuos y grupos dentro de la población prevista de examinandos. Esto es especialmente importante cuando la evaluación se realiza a nivel internacional y de diferentes culturas. La evaluación del constructo subyacente y propiedades de la prueba dentro de un país o cultura no puede generalizarse a nivel internacional o de otras culturas. Esto puede llevar a interpretaciones inválidas de los puntajes de la prueba. En esos contextos se debe prestar mucha atención al sesgo en las interpretaciones de los puntajes.

Imparcialidad en el acceso a los constructos tal como se miden

La meta de que todos los examinandos previstos tengan una oportunidad plena de demostrar su situación respecto del constructo sometido a medición ha generado inquietudes sobre la accesibilidad en las pruebas. Las situaciones de evaluación accesibles son aquellas que permiten que todos los examinandos en la población prevista, en la medida en que sea viable, muestren su estado respecto de los constructos de destino sin ser indebidamente favorecidos o desfavorecidos por características individuales (p. ej., características relacionadas con la edad, discapacidad, raza/origen étnico, género o idioma) que son irrelevantes para el constructo que la prueba tiene por objeto medir. La accesibilidad es en realidad una cuestión de sesgo de la prueba porque los obstáculos a la accesibilidad pueden dar lugar a diferentes interpretaciones de los puntajes de la prueba para los individuos de diferentes grupos. La accesibilidad tiene también importantes ramificaciones éticas y legales.

La accesibilidad puede entenderse mejor comparando el conocimiento, las habilidades y las capacidades que reflejan los constructos que la prueba tiene por objeto medir con el conocimiento, las habilidades y las capacidades que no son el objeto de la prueba pero que se requieren para responder a las tareas de la prueba o a los ítems de la prueba. Para algunos examinandos, los factores relacionados con características individuales como edad, raza, origen étnico, nivel socioeconómico, antecedentes culturales, discapacidad o competencia en lengua inglesa pueden restringir la accesibilidad y por consiguiente interferir con la medición de los constructos de interés. Por ejemplo, es posible que un examinando con problemas de la vista no pueda acceder al texto impreso de una prueba de personalidad. Si el texto se proporcionara en letra grande, las preguntas de la prueba podrían ser más accesibles para el examinando y sería más probable que llevaran a una medición válida de las características de personalidad del examinando. Es importante ser consciente de las características de la prueba que pueden hacer involuntariamente que las preguntas de la prueba sean menos accesibles para algunos subgrupos de la población prevista de la prueba. Por ejemplo, una pregunta de una prueba que emplee frases idiomáticas no relacionadas al constructo sometido a medición podría tener el efecto de hacer que la prueba sea menos accesible para examinandos que no son hablantes nativos de inglés. La accesibilidad de una prueba también podría verse reducida por preguntas que utilizan vocabulario regional no relacionado con el constructo de destino o que utilizan contextos de estímulo que son menos conocidos para los individuos de algunos subgrupos culturales que de otros.

Como se analiza más adelante en este capítulo, algunas características de los examinandos que impiden el acceso se relacionan con el constructo sometido a medición, por ejemplo, dislexia en el contexto de pruebas de lectura. En estos casos, proporcionar a los individuos acceso al constructo y obtener alguna medida de este puede requerir alguna adaptación del constructo también. En situaciones como esta, es posible que no se pueda desarrollar una medición que sea comparable entre versiones adaptadas y no adaptadas de la prueba; sin embargo, la medida obtenida por la prueba adaptada muy probablemente proporcione una evaluación más exacta de las habilidades y/o capacidades del individuo (aunque tal vez

no de todo el constructo previsto) que la obtenida sin usar la adaptación.

Proporcionar acceso al constructo de una prueba se vuelve particularmente difícil para los individuos con más de una característica que podría interferir con el desempeño en la prueba; por ejemplo, adultos mayores que no tienen un buen nivel de inglés o estudiantes de inglés con discapacidades cognitivas moderadas.

Imparcialidad como validez de las interpretaciones de los puntajes individuales de la prueba para los usos previstos

Es importante tener presente que la imparcialidad se relaciona con la validez de las interpretaciones de los puntajes individuales para los usos previstos. Al intentar asegurar la imparcialidad, a menudo generalizamos entre grupos de examinandos como individuos con discapacidades, adultos mayores, individuos que están aprendiendo inglés y los de diferentes grupos raciales o étnicos o diferentes características culturales y/o socioeconómicas; sin embargo, esto se hace por cuestiones prácticas y no tiene por objeto dejar implícito que estos grupos son homogéneos o que, en consecuencia, todos los miembros de un grupo deben tratarse de manera similar cuando se hacen interpretaciones de los puntajes de la prueba para individuos (a menos que exista evidencia de validación para respaldar esas generalizaciones). Es especialmente importante, cuando se hacen inferencias sobre las habilidades o capacidades de un individuo examinado, tener en cuenta las características individuales del examinando y cómo estas características pueden interactuar con las características contextuales de la situación de evaluación.

La compleja interacción de competencia en un idioma y contexto brinda un ejemplo de los desafíos para una interpretación válida de los puntajes de la prueba para algunos fines de evaluación. La competencia en inglés no solo afecta la interpretación de los puntajes de la prueba de un estudiante de lengua inglesa en pruebas administradas en inglés, sino, lo que es más importante, también afecta el progreso de desarrollo y académico del individuo. Los individuos con diferencias culturales y lingüísticas respecto de la mayoría de los examinandos se exponen al riesgo de interpretaciones de puntajes inexactas debido a múltiples factores asociados con la suposición de que, en ausencia de cuestiones de competencia en un idioma, estas personas tienen trayectorias de desarrollo comparables con los individuos que han crecido en un entorno mediado por un solo idioma y cultura. Por ejemplo, consideremos dos niños de sexto grado que ingresaron en la escuela con competencia limitada en inglés. El primer niño ingresó en la escuela en jardín de infancia y ha recibido instrucción en cursos académicos en inglés; el segundo también ingresó en la escuela en jardín de infancia, pero recibió instrucción en su lengua nativa. Los dos tendrán un patrón de desarrollo diferente. En el primer caso, el desarrollo interrumpido en lengua nativa tiene un efecto atenuante en el aprendizaje y el desempeño académico, pero es posible que la competencia en inglés del individuo no sea un obstáculo significativo para la prueba. Por el contrario, el individuo examinado que ha recibido instrucción en su lengua nativa hasta sexto grado ha tenido la oportunidad de un desarrollo cognitivo, académico y de la lengua completamente apropiado para la edad; pero, si se lo evalúa en inglés, el individuo examinado necesitará que la prueba se administre de tal manera que minimice el obstáculo de la lengua si la competencia en inglés no es parte del constructo sometido a medición.

Como muestran los ejemplos anteriores, la adaptación a las características individuales y el reconocimiento de la heterogeneidad dentro de subgrupos pueden ser importantes para la validez de las interpretaciones individuales de los resultados de la prueba en situaciones donde la intención es comprender y responder al desempeño individual. Se puede justificar que los profesionales se aparten de los procedimientos estandarizados para obtener una medida más exacta del constructo previsto y proporcionar decisiones individuales más apropiadas. Sin embargo, para otros contextos y usos, las desviaciones de los procedimientos estandarizados pueden ser inapropiadas porque cambian el constructo sometido a medición, comprometen la comparabilidad de

puntajes o uso de normas y/o favorecen injustamente a algunos individuos.

Al cerrar esta sección sobre los significados de la imparcialidad, obsérvese que la perspectiva de medición de los *Estándares* excluye explícitamente un punto de vista común de la imparcialidad en el debate público: la imparcialidad como la igualdad de resultados de evaluación para subgrupos de examinandos relevantes. Desde luego, la mayoría de los profesionales dedicados a la evaluación coinciden en que las diferencias de grupos en los resultados de evaluación deberían dar lugar a mayor escrutinio sobre posibles fuentes de sesgo en las pruebas. El examen de diferencias de grupos también puede ser importante en la generación de nuevas hipótesis sobre sesgo, trato imparcial, y la accesibilidad del constructo tal como se mide; y de hecho, es posible que existan requisitos legales para investigar ciertas diferencias en los resultados de evaluación entre subgrupos. Sin embargo, las diferencias de grupos en los resultados en sí mismas no indican que una aplicación de evaluación esté sesgada o sea imparcial.

En muchos casos, no está claro si las diferencias se deben a diferencias reales entre grupos en el constructo sometido a medición o a alguna fuente de sesgo (p. ej., varianza irrelevante de constructo o infrarrepresentación de constructo). En la mayoría de los casos, puede ser alguna combinación de diferencias reales y sesgo. Una búsqueda seria de posibles fuentes de sesgo que no arroje resultados proporciona la confirmación de que el potencial de sesgo es limitado, pero incluso un programa de investigación muy extensivo no puede descartar la posibilidad. Siempre es posible que algo se pase por alto, y por consiguiente, la prudencia sugeriría que se intente minimizar las diferencias. Por ejemplo, algunos subgrupos raciales y étnicos tienen puntajes medios más bajos en algunas pruebas estandarizadas que otros subgrupos. Algunos de los factores que contribuyen a estas diferencias se entienden (p. ej., grandes diferencias en el ingreso familiar y otros recursos, diferencias en la calidad escolar y la oportunidad de aprendizaje de los estudiantes en cuanto al material que se evaluará), pero incluso cuando se han hecho esfuerzos serios para eliminar posibles fuentes de sesgo en el contenido y los formatos de la prueba, el potencial de algún sesgo de puntaje no puede descartarse por completo. Por lo tanto, se justifican los esfuerzos continuos en el diseño y desarrollo de pruebas para eliminar posibles fuentes de sesgo sin comprometer la validez, y que sean compatibles con los estándares legales y regulatorios.

Amenazas a las interpretaciones imparciales y válidas de los puntajes de una prueba

Una amenaza principal a la interpretación imparcial y válida de los puntajes de una prueba proviene de aspectos de la prueba o del proceso de evaluación que pueden producir varianza irrelevante de constructo en los puntajes que sistemáticamente reduce o aumenta los puntajes para grupos identificables de examinados y da por resultado interpretaciones inapropiadas de los puntajes para los usos previstos. Dichos componentes irrelevantes del constructo de los puntajes pueden ser introducidos por muestreo inapropiado del contenido de la prueba, aspectos del contexto de la prueba como falta de claridad en las instrucciones de la prueba, complejidades de los ítems que no se relacionan con el constructo sometido a medición, y/o expectativas de respuestas a la prueba o criterios de calificación que pueden favorecer a un grupo por sobre otro. Además, la oportunidad de aprendizaje (es decir, la medida en que un individuo examinado ha estado expuesto a instrucción o experiencias que han sido supuestas por el desarrollador y/o usuario de la prueba) puede influir en las interpretaciones imparciales y válidas de los puntajes de una prueba para sus usos previstos.

Contenido de la prueba

Una posible fuente de varianza irrelevante de constructo en los puntajes de la prueba surge de contenido inapropiado de la prueba, es decir, contenido de la prueba que confunde la medición del constructo de destino y favorece en forma diferencial a individuos de algunos subgrupos por sobre otros. Una prueba que tiene por objeto medir lectura crítica, por ejemplo, no debe incluir

palabras y expresiones especialmente asociadas con ocupaciones, disciplinas o características culturales, nivel socioeconómico, grupos raciales/étnicos o ubicaciones geográficas en particular, de modo que se maximice la medición del constructo (la capacidad para leer críticamente) y se minimice la confusión de esta medición con conocimientos y experiencias previos que probablemente favorezcan o desfavorezcan a examinandos de subgrupos en particular.

El compromiso y el valor motivacional diferenciales también pueden ser factores en la exacerbación de los componentes del contenido irrelevantes del constructo. El material que probablemente sea interesante de manera diferencial debe equilibrarse para atraer en general a todo el alcance de la población de destino de la evaluación (excepto cuando el nivel de interés sea parte del constructo sometido a medición). En las pruebas, ese equilibrio se extiende a la representación de individuos de una variedad de subgrupos dentro del contenido de la prueba propiamente dicho. Por ejemplo, problemas aplicados pueden presentar a niños y familias de diferentes grupos raciales/étnicos, socioeconómicos y de idioma. Además, el contenido de la prueba o situaciones que sean ofensivos o perturbadores desde el punto de vista emocional para algunos examinandos y que puedan impedir su capacidad para comprometerse con la prueba no deben aparecer en la prueba a menos que el uso del contenido ofensivo o perturbador sea necesario para medir el constructo previsto. Ejemplos de este tipo de contenido son las descripciones gráficas de esclavitud o del Holocausto, cuando dichas descripciones son específicamente requeridas por el constructo.

Dependiendo del contexto y de la finalidad de las pruebas, es tanto común como aconsejable que los desarrolladores de la prueba contraten a un panel independiente y diverso de expertos para que revisen el contenido de la prueba en cuanto a representaciones del lenguaje, ilustraciones, gráficos y otras que podrían ser diferencialmente conocidas o interpretadas de manera diferente por miembros de diferentes grupos y en cuanto a materiales que podrían ser ofensivos o perturbadores desde el punto de vista emocional para algunos examinandos.

Contexto de la prueba

El término *contexto de la prueba*, tal como se usa en el presente, se refiere a múltiples aspectos de la prueba y del entorno de evaluación que pueden afectar el desempeño de un individuo examinado y en consecuencia dar lugar a varianza irrelevante de constructo en los puntajes de la prueba. Dado que la investigación de factores contextuales (p. ej., amenaza de estereotipo) es continua, los desarrolladores de la prueba y usuarios de la prueba deben prestar atención a la bibliografía empírica que surja sobre estos temas de modo que puedan usar esta información cuando la preponderancia de evidencia indique que es apropiado hacerlo. La varianza irrelevante de constructo puede surgir de una falta de claridad en las instrucciones de la prueba, de complejidad no relacionada o de exigencias de lenguaje en las tareas de la prueba, y/o de otras características de ítems de la prueba que no se relacionan con el constructo pero que pueden llevar a algunos individuos a responder de ciertas maneras. Por ejemplo, los individuos examinados de diversos orígenes raciales/étnicos, lingüísticos o culturales o que difieren por género pueden ser mal evaluados por un inventario de interés vocacional cuyas preguntas se refieren de manera desproporcionada a competencias, actividades e intereses que están típicamente asociadas con subgrupos en particular.

Cuando los ámbitos de prueba tienen un contexto interpersonal, la interacción del examinador con el examinando puede ser una fuente de varianza irrelevante de constructo o sesgo. Los usuarios de pruebas deben estar alertas ante la posibilidad de que dichas interacciones puedan en ocasiones afectar la imparcialidad de la prueba. Los profesionales que administran la prueba deben ser conscientes de la posibilidad de interacciones complejas con los examinandos y otras variables situacionales. Los factores que pueden afectar el desempeño del examinando incluyen la raza, origen étnico, género y características lingüísticas y culturales tanto del examinador como del examinando, la experiencia del examinador

con la educación formal, el estilo de evaluación del examinador, el nivel de aculturación del examinado y del examinador, el idioma principal del examinando, el idioma utilizado para la administración de la prueba (si no es el idioma principal del examinando), y el uso de un intérprete bilingüe o bicultural.

La evaluación de individuos que son bilingües o multilingües plantea desafíos especiales. Es posible que una persona que sabe dos o más idiomas no salga bien en la prueba en uno o más de los idiomas. Por ejemplo, es posible que los niños de hogares cuyas familias hablan español puedan comprender el español pero se expresen mejor en inglés o viceversa. Además, algunas personas que son bilingües utilizan su lengua nativa en la mayoría de las situaciones sociales y utilizan el inglés principalmente para actividades académicas y relacionadas con el trabajo; el uso de una o ambas lenguas depende de la naturaleza de la situación. Los hablantes de inglés no nativos que dan la impresión de tener buen nivel en inglés conversacional pueden ser más lentos o no completamente competentes para realizar pruebas que requieren habilidades de comprensión y lectoescritura en inglés. Por lo tanto, en algunos contextos, un entendimiento del tipo y grado de bilingüismo o multilingüismo de un individuo es importante para evaluar al individuo de manera apropiada. Obsérvese que esta cuestión puede no aplicarse cuando el constructo de interés se define como una clase particular de competencia en lenguaje (p. ej., lenguaje académico del tipo que se encuentra en libros, lenguaje y vocabulario específico de las pruebas de centro de trabajo y empleo).

Respuesta a la prueba

En algunos casos, la varianza irrelevante de constructo puede surgir porque los ítems de la prueba suscitan variedades de respuestas distintas de las previstas o porque los ítems pueden resolverse de maneras que no fueron previstas. En la medida en que dichas respuestas sean más típicas de algunos subgrupos de que otros, pueden surgir interpretaciones de puntajes sesgadas. Por ejemplo, algunos clientes que responden a una prueba neuropsicológica pueden intentar proporcionar las respuestas que consideren que espera el administrador de la prueba, en lugar de las respuestas que mejor los describen.

Los componentes irrelevantes del constructo en los puntajes de las pruebas también pueden asociarse con formatos de respuesta a la prueba que plantean dificultades particulares o que son valorados de manera diferencial por individuos en particular. Por ejemplo, el desempeño en la prueba puede depender de alguna capacidad (p. ej., competencia en lengua inglesa o coordinación de motricidad fina) que es irrelevante para los constructos de destino, pero que no obstante implica impedimentos a las respuestas de la prueba para algunos examinandos que no tienen la capacidad. De manera similar, diferentes valores asociados con la naturaleza y el grado de producción verbal pueden influir en las respuestas del examinando. Algunos individuos pueden juzgar la verbosidad o el discurso rápido como algo grosero, mientras que otros pueden considerar esos patrones del habla como indicaciones de alta capacidad mental o cordialidad. Un individuo del primer tipo que es evaluado con valores apropiados para el segundo puede considerarse taciturno, introvertido o de baja capacidad mental. Otro ejemplo es la persona con problemas de memoria o de lenguaje o depresión; la capacidad de esa persona para comunicarse o mostrar interés en comunicarse verbalmente puede estar restringida, lo cual puede dar lugar a interpretaciones de los resultados de la evaluación que sean inválidos y posiblemente perjudiciales para la persona que se evalúa.

En el desarrollo y uso de rúbricas de puntajes, es especialmente importante que el crédito se otorgue por características de respuesta centrales para el constructo sometido a medición y no por características de respuesta que sean irrelevantes o tangenciales al constructo. Las rúbricas de puntajes pueden favorecer involuntariamente a algunos individuos por sobre otros. Por ejemplo, una rúbrica de puntajes para un ítem de respuesta construida podría reservar el nivel de puntaje más alto para los examinandos que proporcionan más información o elaboración que la que efectivamente se solicitó. En esta situación, los

examinandos que simplemente siguen instrucciones, o los examinandos que valoran la concisión en las respuestas, obtendrán menores puntajes; por consiguiente, las características de los individuos se convierten en componentes irrelevantes del constructo de los puntajes de la prueba. De manera similar, la calificación de repuestas abiertas puede introducir varianza irrelevante de constructo para algunos examinandos si los evaluadores y/o rutinas de puntaje automático no son sensibles a toda la diversidad de modos en que los individuos expresan sus ideas. Con el advenimiento del puntaje automático para tareas de desempeño complejas, por ejemplo, es importante examinar la validez de los resultados del puntaje automático para subgrupos relevantes en la población de examinandos.

Oportunidad de aprendizaje

Por último, la *oportunidad de aprendizaje* —el grado en que los individuos han estado expuestos a instrucción o conocimientos que les ofrezcan la oportunidad de aprender el contenido y las habilidades objeto de la prueba— tiene varias implicaciones para la interpretación imparcial y válida de los puntajes de la prueba para sus usos previstos. La oportunidad previa de aprendizaje de los individuos puede ser un importante factor contextual a considerar al interpretar y hacer inferencias de los puntajes de la prueba. Por ejemplo, es posible que un inmigrante reciente que ha tenido escasa exposición previa a la escuela no haya tenido la oportunidad de aprender conceptos que un inventario de personalidad o medida de capacidad suponen como conocimientos comunes, incluso si la medida es administrada en la lengua nativa del examinando. De manera similar, como otro ejemplo, ha habido considerable debate público sobre las posibles desigualdades en los recursos escolares disponibles para estudiantes de grupos tradicionalmente desfavorecidos, por ejemplo, minorías raciales, étnicas, de lenguas y culturales y estudiantes rurales. Dichas desigualdades afectan la calidad de educación recibida. En la medida en que exista desigualdad, la validez de las inferencias sobre la capacidad de los estudiantes extraídas de puntajes de pruebas de rendimiento puede verse comprometida. No tener en cuenta la oportunidad previa de aprendizaje podría dar lugar a un diagnóstico equivocado, colocación inapropiada y/o asignación inapropiada de servicios, lo que podría tener consecuencias significativas para un individuo.

Más allá de su impacto en la validez de las interpretaciones de puntajes de la prueba para usos previstos, la oportunidad de aprendizaje tiene importantes ramificaciones legales y en materia de políticas en educación. La oportunidad de aprendizaje es una cuestión de imparcialidad cuando una autoridad proporciona acceso diferencial a la oportunidad de aprendizaje para algunos individuos y responsabiliza de su desempeño en la prueba a los individuos a quienes no se les proporcionó esa oportunidad. Este problema puede afectar a las pruebas de competencia de alto riesgo en educación, por ejemplo, cuando las autoridades educativas requieren un cierto nivel de desempeño en la prueba para la graduación de la escuela secundaria. En este caso, existe una cuestión de imparcialidad en cuanto a que los estudiantes no sean responsabilizados de sus resultados en la prueba, o enfrenten consecuencias negativas permanentes graves por ellos, cuando sus experiencias escolares no les hayan dado la oportunidad de aprender la asignatura cubierta por la prueba. En esos casos, los puntajes bajos de los estudiantes pueden reflejar exactamente qué saben y pueden hacer, de modo que, técnicamente, la interpretación de los resultados de la prueba para el fin de medir cuánto han aprendido los estudiantes no puede estar sesgada. Sin embargo, puede considerarse injusto penalizar severamente a los estudiantes por circunstancias ajenas a su control, es decir, por no aprender contenido que sus escuelas no han enseñado. Se encuentra generalmente aceptado que antes de que puedan imponerse consecuencias de alto riesgo por reprobar un examen en contextos educativos, debe haber evidencia de que los estudiantes han recibido un plan de estudios e instrucción que incorporan los constructos abordados por la prueba.

Varias cuestiones importantes surgen cuando la oportunidad de aprendizaje se considera como un componente de imparcialidad. En primer

lugar, es difícil definir la oportunidad de aprendizaje en la práctica educativa, particularmente a nivel de individuo. La oportunidad es generalmente un asunto de grado y es difícil de cuantificar; además, la medición de algunos resultados de aprendizaje importantes puede requerir que los estudiantes trabajen con materiales que han visto antes. En segundo lugar, incluso si es posible documentar los temas incluidos en el plan de estudios para un grupo de estudiantes, la cobertura de contenido específico para cualquier estudiante puede ser imposible de determinar. En tercer lugar, otorgar un diploma a un individuo examinado con bajo puntaje basándose en que el estudiante no ha tenido suficiente oportunidad de aprender el material evaluado significa certificar a alguien que no alcanzado el grado de competencia que el diploma tiene por objeto representar.

Debe observarse que las inquietudes sobre la oportunidad de aprendizaje no necesariamente se aplican a situaciones en las que la misma autoridad no es responsable tanto de impartir instrucción como de evaluar y/o interpretar los resultados. Por ejemplo, en las decisiones sobre admisión universitaria, la oportunidad de aprendizaje puede escapar al control de los usuarios de la prueba y puede no influir en la validez de las interpretaciones de la prueba para su uso previsto (p. ej., decisiones de selección y/o admisiones). El capítulo 12, "Pruebas y evaluación educativas", proporciona una perspectiva adicional sobre la oportunidad de aprendizaje.

Minimizar los componentes irrelevantes del constructo mediante el diseño de la prueba y adaptaciones de la prueba

Las pruebas estandarizadas deben diseñarse para facilitar la accesibilidad y minimizar los obstáculos irrelevantes del constructo para todos los examinandos en la población de destino, siempre que sea posible. Antes de considerar la necesidad de cualquier adaptación de evaluación para los examinandos que puedan tener necesidades especiales, el desarrollador de la evaluación primero debe intentar mejorar la accesibilidad dentro de la propia prueba. Algunos de estos principios básicos se incluyen en el proceso de diseño de pruebas denominado diseño universal. Al utilizar el diseño universal, los desarrolladores de la prueba comienzan el proceso de desarrollo de la prueba con vistas a maximizar la imparcialidad. El diseño universal destaca la necesidad de desarrollar pruebas que sean tan utilizables como sea posible para todos los examinandos en la población prevista de la prueba, independientemente de características tales como género, edad, características lingüísticas, cultura, nivel socioeconómico o discapacidad.

Los principios del diseño universal incluyen definir constructos de manera precisa, de modo que lo que se mida pueda diferenciarse claramente de las características del examinando que sean irrelevantes para el constructo pero que podrían de otro modo interferir con la capacidad de responder de algunos examinandos. El diseño universal evita, cuando es posible, características y formatos de los ítems, o características de la prueba (por ejemplo, aceleración de la prueba inapropiada), que puedan sesgar los puntajes para individuos o subgrupos debido a características irrelevantes del constructo que sean específicas de estos examinandos.

Los procesos del diseño universal se esfuerzan por minimizar las dificultades de acceso teniendo en cuenta características de la prueba que pueden impedir el acceso al constructo para determinados examinandos, como la elección de contenido, las tareas de la prueba, los procedimientos de respuesta y los procedimientos de evaluación. Por ejemplo, el contenido de pruebas puede hacerse más accesible proporcionando tamaños de fuente seleccionadas por los usuarios en una prueba basada en tecnología, evitando contextos de ítems que probablemente no serían conocidos para los individuos debido a su contexto cultural, proporcionando tiempo de administración extendido cuando la velocidad no es relevante para el constructo sometido a medición, o minimizando la carga lingüística de los ítems de la prueba previstos para medir constructos distintos de competencias en el idioma en que se administra la prueba.

Si bien los principios del diseño universal para evaluación proporcionan una guía útil para desarrollar evaluaciones que reducen la varianza irrelevante de constructo, los investigadores aún están reuniendo evidencia empírica para respaldar algunos de estos principios. Es importante observar que no todas las pruebas pueden hacerse accesibles para todos mediante atención a cambios de diseño como los mencionados arriba. Incluso cuando las pruebas se desarrollan para maximizar la imparcialidad a través del uso de diseño universal y otras prácticas para aumentar el acceso, aún existirán situaciones en las que la prueba no es apropiada para todos los examinandos en la población prevista. Por lo tanto, es posible que se necesiten algunas adaptaciones de la prueba para los individuos cuyas características de otro modo impedirían su acceso al examen.

Las *adaptaciones* son cambios al diseño o administración originales de la prueba para aumentar el acceso a la prueba para dichos individuos. Por ejemplo, una persona que es ciega puede leer solo en formato braille, y es posible que un individuo con hemiplejia no pueda sostener un lápiz y por lo tanto tenga dificultad para completar un examen escrito estándar. Los estudiantes con competencia limitada en inglés pueden ser competentes en física, pero es posible que no puedan demostrar su conocimiento si la prueba de física se administra en inglés. Dependiendo de las circunstancias de evaluación y los fines de la prueba, así como de las características individuales, esas adaptaciones podrían incluir cambiar el contenido o presentación de los ítems de la prueba, cambiar las condiciones de administración y/o cambiar los procesos de respuesta. El término adaptación se utiliza para hacer referencia a cualquiera de estos cambios. Es importante, no obstante, diferenciar entre cambios que dan lugar a puntajes comparables y cambios que pueden no producir puntajes que sean comparables a los de la prueba original. Si bien los términos pueden tener significados diferentes en virtud de las leyes aplicables, tal como se utiliza en los *Estándares* el término *adecuación* se utiliza para indicar cambios con los que se conserva la comparabilidad de puntajes, y el término modificación se utiliza para indicar cambios que afectan el constructo medido por la prueba. Con una *modificación*, los cambios afectan el constructo sometido a medición y en consecuencia llevan a puntajes que difieren en significado de los de la prueba original.[1]

Es importante tener presente que la atención al diseño y la provisión de pruebas alteradas no siempre garantiza que los resultados de la prueba serán imparciales y válidos para todos los individuos examinados. Quienes administran pruebas e interpretan los puntajes de la prueba necesitan desarrollar una comprensión cabal de la utilidad y las limitaciones de los procedimientos de diseño de pruebas para accesibilidad y cualquier alteración que se ofrezca.

Variedad de adaptaciones de prueba

En lugar de una simple dicotomía, las posibles adaptaciones de prueba reflejan una amplia variedad de cambios en las pruebas. En un extremo de la variedad se encuentran las adecuaciones de la prueba. Tal como el término se utiliza en los Estándares, las adecuaciones consisten en cambios relativamente menores en la presentación y/o el formato de la prueba, la administración de la prueba, o los procedimientos de respuesta que mantienen el constructo original y dan por resultado puntajes comparables a los de la prueba original. Por ejemplo, el aumento del tamaño del texto podría ser una adecuación para un examinando con un problema de la vista que de otro modo tendría dificultad para descifrar las instrucciones o ítems de la prueba. Los glosarios de lengua inglesa nativa son un ejemplo de una adecuación que podría proporcionarse para

[1]La Ley sobre Estadounidenses con Discapacidades (ADA, por sus siglas en inglés) utiliza los términos *adecuación y modificación* de manera diferente que los Estándares. El Título I de la ADA utiliza el término *adecuación razonable* para referirse a cambios que permiten que individuos cualificados con discapacidades obtengan empleo para realizar sus trabajos. Los Títulos II y III utilizan el término *modificación razonable* gran en parte de la misma manera. En virtud de la ADA, una adecuación o modificación a una prueba que fundamentalmente altera el constructo sometido a medición no se llamaría de manera diferente; sino que probablemente se consideraría no "razonable".

examinandos con competencia limitada en inglés en una prueba de seguridad en construcción para ayudarles a comprender lo que se pregunta. Los glosarios contendrían palabras que, si bien no se relacionan directamente al constructo sometido a medición, ayudarían a examinandos con competencia limitada en inglés a comprender el contexto de la pregunta o tarea planteada.

En el otro extremo de la variedad se encuentran las adaptaciones que transforman el constructo sometido a medición, incluyendo el contenido de la prueba y/o las condiciones de evaluación, para obtener una medida razonable de un constructo algo diferente pero apropiado para los examinandos designados. Por ejemplo, en evaluación educativa, se diseñan diferentes pruebas que abordan los estándares de rendimiento alternativos para estudiantes con discapacidades cognitivas graves correspondientes a los mismos temas en los que se evalúa a los estudiantes sin discapacidades. Claramente, los puntajes de estas pruebas diferentes no pueden considerarse comparables a los que surgen de la evaluación general, pero en cambio representan puntajes de una nueva prueba que requiere los mismos procesos rigurosos de desarrollo y validación que se llevarían a cabo para cualquier nueva evaluación. (En el capítulo 12 se incluye un debate ampliado del uso de dichas evaluaciones alternativas; las evaluaciones alternativas no se seguirán tratando en el presente capítulo). Otras adaptaciones cambian el constructo previsto para hacer que sea accesible para los estudiantes designados mientras conservan tanto como sea posible del constructo original. Por ejemplo, una adaptación de una prueba de lectura podría proporcionar a un estudiante disléxico un lector de pantalla que lea en voz alta los pasajes y las preguntas de la prueba que miden la comprensión de lectura. Si el constructo está intencionalmente definido como que requiere tanto la capacidad de decodificar como la capacidad de comprender lenguaje escrito, la adaptación requeriría una interpretación diferente de los puntajes de la prueba como una medida de la comprensión de lectura. Claramente, esta adaptación cambia el constructo sometido a medición, porque el estudiante no tiene que decodificar el texto impreso; pero sin la adaptación, es posible que el estudiante no pueda demostrar ninguna situación con respecto al constructo de comprensión de lectura. Por otra parte, si la finalidad de la prueba de lectura es evaluar la comprensión sin importar la capacidad de decodificación, podría juzgarse que la adaptación respalda interpretaciones más válidas de la comprensión de lectura de algunos estudiantes y la esencia de las partes relevantes del constructo podría juzgarse intacta. El desafío para quienes reportan, interpretan y/o utilizan puntajes de pruebas de pruebas adaptadas es reconocer qué adaptaciones proporcionan puntajes que son comparables con los puntajes de la evaluación original sin adaptar y qué adaptaciones no. Este desafío se vuelve aún más difícil cuando la evidencia para respaldar la comparabilidad de puntajes no está disponible.

Adecuaciones de la prueba: medidas comparables que mantienen el constructo previsto

La *comparabilidad de puntajes* permite a los usuarios de las pruebas hacer inferencias comparables basadas en los puntajes para todos los examinandos. La comparabilidad también es la característica definitoria para que una adaptación de prueba se considere una adecuación. Los puntajes de la versión adaptada de la prueba deben arrojar inferencias comparables a los de la versión estándar; hacer que esto ocurra es una proposición que plantea desafíos. Por un lado, los procedimientos comunes, uniformes son un apoyo básico para la validez y comparabilidad de puntajes. Por otra parte, las adecuaciones por su propia naturaleza significan que algo en las circunstancias de evaluación ha sido cambiado porque adherir a los procedimientos estandarizados originales interferiría con la medición válida de los constructos previstos para algunos individuos.

La comparabilidad de inferencias hechas a partir de puntajes de prueba adaptados se basa en gran parte en que los puntajes representen o no los mismos constructos que los de la prueba original. Esta determinación requiere

una definición muy clara de los constructos previstos. Por ejemplo, cuando hablantes no nativos del idioma de la prueba completan una encuesta de sus conocimientos sobre salud y nutrición, uno puede no saber si el puntaje de la prueba es, total o parcialmente, una medida de la capacidad para leer en el idioma de la prueba más que una medida del constructo previsto. Si la prueba no tiene por objeto también ser una medida de la capacidad para leer en inglés, los puntajes de la prueba no representan los mismos constructos para los individuos examinados que pueden tener habilidades de lectura deficientes, como examinandos con competencia limitada en inglés, que para los que son completamente competentes para leer en inglés. Una adaptación que mejora la accesibilidad de la prueba para hablantes no nativos de inglés proporcionando apoyos lingüísticos directos o indirectos puede arrojar un puntaje no contaminado por la capacidad de comprender inglés.

Al mismo tiempo, la infrarrepresentación de constructo es una amenaza primaria a la validez de las adecuaciones de la prueba. Por ejemplo, el tiempo extra es una adecuación común, pero si la velocidad es parte del constructo previsto, no es apropiado permitir tiempo extra en la administración de la prueba. Los puntajes obtenidos en la prueba con tiempo de administración extendido pueden infrarrepresentar el constructo medido por la prueba estrictamente cronometrada porque la velocidad no será parte del constructo medido por la prueba de tiempo extendido. De manera similar, traducir una prueba de comprensión de lectura utilizada para la selección para un programa de capacitación de una organización es inapropiado si la comprensión de lectura en inglés es importante para la participación exitosa en el programa.

Las afirmaciones de que las versiones adaptadas de una prueba arrojan interpretaciones comparables a las basadas en puntajes de la prueba original y de que el constructo sometido a medición no se ha cambiado deben evaluarse y sustentarse con evidencia. Si bien la comparabilidad de puntajes es más fácil de establecer cuando diferentes formularios de prueba se construyen siguiendo procedimientos idénticos y luego se equiparan estadísticamente, esos procedimientos por lo general no son posibles para versiones adaptadas y no adaptadas de las pruebas. En cambio, la evidencia relevante puede adoptar diversas formas, desde estudios experimentales para determinar la equivalencia de constructo hasta estudios cualitativos, más pequeños, y/o el uso de juicio profesional y revisión de expertos. Cualquiera sea el caso, los desarrolladores y/o usuarios de la prueba deben buscar evidencia de la comparabilidad de las evaluaciones adaptada y original.

Se ha implementado una variedad de estrategias para adecuar las pruebas y procedimientos de evaluación para responder a las necesidades de los examinandos con discapacidades y aquellos con características lingüísticas y culturales diversas. Similares enfoques pueden adaptarse para otros subgrupos. Las estrategias específicas dependen de la finalidad de la prueba y de los constructos que la prueba tiene por objeto medir. Algunas estrategias requieren cambiar los procedimientos de administración de la prueba (p. ej., instrucciones, formato de respuesta), mientras que otras alteran el medio, el momento, los contextos o el formato de evaluación. Dependiendo del contexto lingüístico o de la naturaleza y grado de la discapacidad, uno o más cambios en la evaluación pueden ser apropiados para un individuo en particular.

Independientemente de las características del individuo que hacen que las adecuaciones sean necesarias, es importante que las adecuaciones de la prueba aborden las cuestiones de acceso específicas que de otro modo sesgarían los resultados de la prueba de un individuo. Por ejemplo, las adecuaciones provistas a examinandos con competencia limitada en inglés deben diseñarse para abordar necesidades de apoyo lingüístico apropiado; las proporcionadas a examinandos con problemas de la vista deben abordar la incapacidad de ver el material de la prueba. Las adecuaciones deben ser efectivas en la eliminación de los obstáculos irrelevantes del constructo al desempeño en la prueba de un individuo sin proporcionar una ventaja injusta sobre individuos

que no reciben la adecuación. Verdaderamente, alcanzar ambos objetivos puede ser un desafío.

Las adaptaciones que involucran traducciones de la prueba merecen consideración especial. Simplemente traducir una prueba de un idioma a otro no asegura que la traducción produzca una versión de la prueba que sea comparable en contenido y nivel de dificultad con la versión original de la prueba, o que la prueba traducida produzca puntajes que sean igualmente confiables/precisos y válidos que los de la prueba original. Además, no se puede suponer que la aculturación relevante, las experiencias clínicas o educativas sean similares para los examinandos que realizan la versión traducida y para el grupo de destino utilizado para desarrollar la versión original. Asimismo, no se puede suponer que la traducción a la lengua nativa sea siempre una adecuación preferida. La investigación en evaluaciones educativas, por ejemplo, muestra que las pruebas con contenido traducido no son efectivas a menos que a los examinandos se los haya instruido utilizando el idioma de la prueba traducida. Cuando las pruebas se traducen de un idioma a un segundo idioma, debe reunirse y reportarse evidencia de la validez, confiabilidad/precisión y comparabilidad de puntajes en las diferentes versiones de las pruebas.

Cuando la adecuación de la prueba emplea el uso de un intérprete, es aconsejable, cuando sea viable, obtener a alguien que tenga una comprensión básica del proceso de evaluación psicológica y educativa, tenga buen nivel en el idioma de la prueba y la lengua nativa del examinando y esté familiarizado con el contexto cultural del examinando. El intérprete idealmente debe comprender la importancia de seguir procedimientos estandarizados, la importancia de transmitir exactamente al examinador las respuestas reales del examinando, y el rol y las responsabilidades del intérprete en la evaluación. El intérprete debe ser cuidadoso de no proporcionar asistencia alguna al candidato que pudiera comprometer la validez de la interpretación para los usos previstos de los resultados de la evaluación.

Por último, es importante estandarizar procedimientos para implementar adecuaciones, siempre que sea posible, de modo que se mantenga la comparabilidad de puntajes. Los procedimientos estandarizados para las adecuaciones de las pruebas deben incluir reglas para determinar quién es elegible para una adecuación, y precisamente cómo debe administrarse la adecuación. Los usuarios de la prueba deben supervisar la adhesión a las reglas de elegibilidad y administración apropiada de la prueba adaptada.

Modificaciones de la prueba: medidas no comparables que cambian el constructo previsto

Es posible que haya ocasiones en que se requiera flexibilidad adicional para obtener incluso una medida parcial del constructo; es decir, es posible que sea necesario considerar una modificación a una prueba que dará por resultado cambios en el constructo previsto para proporcionar incluso acceso limitado al constructo sometido a medición. Por ejemplo, un individuo con discalculia puede tener capacidad limitada para hacer cálculos sin una calculadora; sin embargo, si se le proporciona una calculadora, es posible que el individuo pueda hacer los cálculos requeridos en la evaluación. Si el constructo que se evalúa involucra una habilidad matemática más amplia, el individuo puede tener acceso limitado al constructo que se mide sin el uso de una calculadora; con la modificación, no obstante, el individuo puede demostrar habilidades de resolución de problemas matemáticos, incluso si no puede demostrar habilidades de cálculo. Puesto que las evaluaciones modificadas miden un constructo diferente del medido por la evaluación estandarizada, es importante interpretar los puntajes de la evaluación como puntajes resultantes de una nueva prueba y reunir toda evidencia que sea necesaria para evaluar la validez de las interpretaciones para los usos previstos de los puntajes. Para interpretaciones de puntajes basadas en normas, cualquier modificación que cambie el constructo invalidará las normas para las interpretaciones de puntajes. Del mismo modo, si se cambia el constructo, las interpretaciones de puntajes basadas en criterios de la evaluación modificada (por ejemplo, tomar decisiones de clasificación como “aprobado/

reprobado" o asignar categorías de dominio como "básico," "competente" o "avanzado" utilizando puntajes de corte determinados sobre la evaluación original) no serán válidas.

Reporte de puntajes de pruebas adaptadas y modificadas

Por lo general, los administradores de pruebas y los profesionales de evaluación documentan pasos utilizados al hacer adecuaciones o modificaciones de las pruebas en el reporte de la prueba; los médicos también pueden incluir una discusión de la validez de las interpretaciones de los puntajes resultantes para los usos previstos. Esta práctica de reportar la naturaleza de las adecuaciones y modificaciones es coherente con los requisitos implícitos para comunicar información en cuanto a la naturaleza del proceso de evaluación si estos cambios pueden afectar la confiabilidad/precisión de los puntajes de la prueba o la validez de interpretaciones derivadas de los puntajes de la prueba.

La indicación de reportes de puntajes de la prueba puede ser una cuestión controvertida y sujeta a requisitos legales. Cuando existe evidencia clara de que los puntajes de pruebas o administraciones de pruebas regulares y alteradas no son comparables, debe considerarse informar a los usuarios de los puntajes, posiblemente indicando los resultados de la prueba para señalar su naturaleza especial, en la medida permitida por ley. Cuando existe evidencia creíble de que los puntajes de pruebas regulares y alteradas son comparables, la indicación por lo general no es apropiada. Existe escaso acuerdo en el campo en cuanto a cómo proceder cuando no existe evidencia creíble sobre comparabilidad. En la medida posible, los desarrolladores y/o usuarios de la prueba deben reunir evidencia para examinar la comparabilidad de pruebas o procedimientos de administración regulares y alterados para los fines previstos de la prueba.

Uso apropiado de adecuaciones o modificaciones

Dependiendo del constructo sometido a medición y de la finalidad de la prueba, existen algunas situaciones de evaluación en las que las adecuaciones tal como las definen los *Estándares* no son necesarias o en que las modificaciones tal como las definen los *Estándares* no son apropiadas. En primer lugar, el motivo para la posible alteración, como habilidades en lengua inglesa o una discapacidad, puede de hecho ser directamente relevante para el constructo principal. En las pruebas de empleo, sería inapropiado hacer cambios en la prueba si la prueba se ha diseñado para evaluar habilidades esenciales requeridas para el puesto y los cambios en la prueba alterarían fundamentalmente el constructo sometido a medición. Por ejemplo, a pesar del aumento de la automatización y el uso de dispositivos de grabación, algunos puestos de escribientes judiciales requieren individuos que puedan trabajar rápidamente y con precisión. La velocidad es un aspecto importante del constructo y no puede adaptarse. En otro ejemplo, una muestra de trabajo para un puesto de servicio al cliente que requiere comunicación fluida en inglés no sería traducida a otro idioma.

En segundo lugar, una adaptación para una discapacidad en particular es inapropiada cuando la finalidad de una prueba es diagnosticar la presencia o el grado de esa discapacidad. Por ejemplo, dar tiempo extra en una prueba de tiempo para determinar el nivel de distracción y las dificultades en la velocidad de procesamiento asociadas con trastorno por déficit de atención haría imposible determinar el grado en que realmente existen las dificultades de atención y de velocidad de procesamiento.

En tercer lugar, es importante destacar que no todos los individuos dentro de una clase general de individuos examinados, como los de características lingüísticas y culturales diversas o con discapacidades, pueden requerir disposiciones especiales cuando realizan pruebas. Las habilidades de idioma, el conocimiento cultural o discapacidades específicas que poseen estos individuos, por ejemplo, podrían no influir en su desempeño en un tipo particular de prueba. Por consiguiente, para estos individuos, no se necesitan cambios.

La efectividad de una adecuación dada también desempeña un papel en determinaciones de

uso apropiado. Si una determinada adecuación o modificación no aumenta el acceso al constructo tal como se mide, no tiene mucho sentido utilizarla. La evidencia de efectividad puede reunirse a través de estudios cuantitativos o cualitativos. El juicio profesional necesariamente desempeña un papel sustancial en las decisiones sobre cambios en la prueba o situación de evaluación.

En resumen, la imparcialidad es una cuestión fundamental para la interpretación válida de los puntajes de la prueba, y por lo tanto debe ser la meta para todas las aplicaciones de evaluación. La imparcialidad es responsabilidad de todas las partes involucradas en el desarrollo, la administración y la interpretación de puntajes de la prueba para los fines previstos de la prueba.

ESTÁNDARES DE IMPARCIALIDAD

Los estándares en este capítulo comienzan con un estándar global (numerado 3.0), que se ha diseñado para transmitir la intención central o enfoque principal del capítulo. El estándar global también puede verse como el principio rector del capítulo, y es aplicable a todas las pruebas y usuarios de pruebas. Todos los estándares posteriores se han separado en cuatro unidades temáticas denominadas de la siguiente manera:

1. Diseño, desarrollo, administración y procedimientos de calificación de las pruebas que minimizan los obstáculos a interpretaciones válidas de los puntajes para la variedad más amplia posible de individuos y subgrupos relevantes
2. Validez de las interpretaciones de los puntajes de las pruebas para los usos previstos para la población prevista de individuos examinados
3. Adecuaciones para eliminar obstáculos irrelevantes del constructo y respaldar interpretaciones válidas de puntajes para sus usos previstos
4. Protecciones contra las interpretaciones inapropiadas de los puntajes para los usos previstos

Estándar 3.0

Todos los pasos en el proceso de evaluación, incluyendo diseño, validación, desarrollo, administración y procedimientos de calificación de la prueba, deben diseñarse de tal manera que minimicen la varianza irrelevante de constructo y promuevan las interpretaciones válidas de los puntajes para los usos previstos para todos los individuos examinados en la población prevista.

Comentario: La idea central de la imparcialidad en las pruebas es identificar y eliminar obstáculos irrelevantes del constructo al desempeño máximo para cualquier individuo examinado. Eliminar estos obstáculos permite la interpretación comparable y válida de los puntajes de la prueba para todos los individuos examinados. La imparcialidad es por lo tanto central para la validez y comparabilidad de la interpretación de puntajes de la prueba para usos previstos.

Unidad 1. Diseño, desarrollo, administración y procedimientos de calificación de las pruebas que minimizan los obstáculos a interpretaciones válidas de los puntajes para la variedad más amplia de individuos y subgrupos relevantes

Estándar 3.1

Los responsables del desarrollo, la revisión y la administración de la prueba deben diseñar todos los pasos del proceso de evaluación para promover interpretaciones válidas de los puntajes para los usos previstos de los puntajes para la variedad más amplia posible de individuos y subgrupos relevantes en la población prevista.

Comentario: Los desarrolladores de la prueba deben delinear claramente tanto los constructos que ha de medir la prueba como las características de los individuos y subgrupos en la población prevista de examinandos. Las tareas e ítems de la prueba deben diseñarse para maximizar el acceso y estar libres de obstáculos irrelevantes del constructo siempre que sea posible para todos los individuos y subgrupos relevantes en la población prevista de examinandos. Una manera de lograr estas metas es crear la prueba utilizando principios de diseño universal, que tienen en cuenta las características de todos los individuos para los que está prevista la prueba e incluyen elementos tales como definir precisamente los constructos y evitar, cuando es posible, características y formatos de ítems y pruebas (por ejemplo, aceleración de la prueba) que pueden comprometer las interpretaciones válidas de los puntajes para individuos o subgrupos relevantes. Otro principio del diseño universal es proporcionar procedimientos e instrucciones de

evaluación simples, claros e intuitivos. En última instancia, la meta es diseñar un proceso de evaluación que, en la medida de lo posible, eliminará los potenciales obstáculos a la medición del constructo previsto para todos los individuos, incluyendo los individuos que requieren adecuaciones. Los desarrolladores de la prueba deben ser conocedores de las diferencias del grupo que pueden interferir con la precesión de puntajes y la validez de inferencias de puntajes de la prueba y deben poder tomar medidas para reducir el sesgo.

Estándar 3.2

Los desarrolladores de la prueba son responsables de desarrollar pruebas que midan el constructo previsto y de minimizar el potencial de que las pruebas se vean afectadas por características irrelevantes del constructo, como características lingüísticas, comunicativas, cognitivas, culturales, físicas y otras.

Comentario: Las características lingüísticas, comunicativas, cognitivas, culturales, físicas y/u otras innecesarias en el estímulo del ítem de la prueba y/o los requisitos de respuesta pueden impedir a algunos individuos la demostración de su situación respecto de los constructos previstos. Los desarrolladores de pruebas deben utilizar lenguaje en las pruebas que sea coherente con los fines de las pruebas y que sea familiar para la variedad más amplia posible de examinandos. Evitar el uso de lenguaje que tenga diferentes significados o diferentes connotaciones para subgrupos relevantes de examinandos ayudará a garantizar que los examinandos que tienen las habilidades que se evalúan puedan comprender qué se les está pidiendo y respondan adecuadamente. El nivel de competencia en idioma, la respuesta física u otras exigencias requeridas por la prueba deben mantenerse al mínimo requerido para satisfacer los requisitos de trabajo y acreditación y/o para representar los constructos de destino. En situaciones laborales, la modalidad en la que se evalúa la competencia en idioma debe ser comparable con la requerida en el puesto, por ejemplo, oral y/o escrita, comprensión y/o producción. De manera similar, las exigencias físicas y verbales de los requisitos de respuesta deben ser coherentes con el constructo previsto.

Estándar 3.3

Los responsables del desarrollo de la prueba deben incluir subgrupos relevantes en estudios de validez, confiabilidad/precisión y otros estudios preliminares utilizados cuando se construye la prueba.

Comentario: Los desarrolladores de la prueba deben incluir a individuos de subgrupos relevantes de la población prevista de la prueba en muestras de pruebas piloto o de campo utilizadas para evaluar lo adecuado de un ítem y una prueba para las interpretaciones del constructo. Los análisis que se llevan a cabo utilizando datos de pruebas piloto y de campo deben procurar detectar aspectos del diseño, contenido y formato de la prueba que podrían distorsionar las interpretaciones de los puntajes de la prueba para los usos previstos de los puntajes de la prueba para grupos e individuos en particular. Dichos análisis podrían emplear una variedad de metodologías, incluyendo las apropiadas para tamaños de la muestra pequeños, como el juicio de expertos, grupos focales y laboratorios cognitivos. Las fuentes de evidencia tanto cualitativas como cuantitativas son importantes al evaluar si los ítems son sólidos y apropiados desde el punto de vista psicométrico para todos los subgrupos relevantes.

Si los tamaños de la muestra lo permiten, a menudo es valioso llevar adelante análisis separados para subgrupos relevantes de la población. Cuando no es posible incluir cantidades suficientes en las muestras de las pruebas piloto y/o de campo a fin de hacer análisis separados, los resultados de la prueba operativa pueden acumularse y utilizarse para llevar a cabo análisis cuando los tamaños de la muestra se vuelven lo suficientemente grandes para respaldar los análisis.

Si los resultados de las pruebas piloto o de campo indican que los ítems o pruebas funcionan de manera diferencial para individuos de, por ejemplo grupos etarios, culturales, de

discapacidad, género, lingüísticos y/o raciales/étnicos relevantes en la población de examinandos, los desarrolladores de la prueba deben investigar aspectos del diseño, contenido y formato de la prueba (incluyendo formatos de respuesta) que podrían contribuir al desempeño diferencial de miembros de estos grupos y, si se justifica, eliminar estos aspectos de prácticas de desarrollo de pruebas futuras.

Las revisiones de expertos y de sensibilidad pueden servir para proteger contra lenguaje e imágenes irrelevantes del constructo, incluyendo los que pueden ofender a algunos individuos o subgrupos, y contra contexto irrelevante del constructo que puede ser más conocido para algunos que para otros. Los editores de la prueba suelen realizar revisiones de sensibilidad de todo el material de la prueba para detectar y eliminar material sensible de las pruebas (p. ej., texto, gráficos y otras representaciones visuales dentro de la prueba que podrían percibirse como ofensivas para algunos grupos y posiblemente afectar los puntajes de individuos de estos grupos). Esas revisiones deben llevarse a cabo antes de que una prueba se vuelva operativa.

Estándar 3.4

Los examinandos deben recibir un trato comparable durante la administración y el proceso de calificación de la prueba.

Comentario: Los responsables de evaluar deben adherir a administración, calificación y protocolos de seguridad de la prueba estandarizados de modo que los puntajes de la prueba reflejen los constructos que se evalúan y no estén indebidamente influidos por idiosincrasias en el proceso de evaluación. Los responsables de la administración de la prueba deben mitigar la posibilidad de predisposiciones personales que podrían afectar la administración de la prueba o la interpretación de puntajes.

Las pruebas computarizadas y otras formas de evaluación basadas en tecnología suman cuestiones extras para la estandarización en la administración y calificación. Los individuos examinados deben tener acceso a tecnología de modo que los aspectos de la tecnología propiamente dichos no influyan en los puntajes. Los individuos examinados que trabajan en equipos más viejos y más lentos pueden verse injustamente desfavorecidos en relación con los que trabajan en equipos más nuevos. Si las computadoras u otros dispositivos difieren en velocidad de procesamiento o movimiento de una pantalla a la otra, en la fidelidad de los objetos visuales, o en otras maneras importantes, es posible que factores irrelevantes del constructo puedan influir en el desempeño en la prueba.

Cuestiones relacionadas con la seguridad de la prueba y la fidelidad de la administración también pueden amenazar la comparabilidad del trato de individuos y la validez e imparcialidad de las interpretaciones de puntajes de la prueba. Por ejemplo, la distribución no autorizada de ítems a algunos individuos examinados, pero no a otros, o administraciones de pruebas sin supervisión en las que la estandarización no puede garantizarse, podrían proporcionar una ventaja a algunos examinandos por sobre otros. En estas situaciones, los resultados de la prueba deben interpretarse con cautela.

Estándar 3.5

Los desarrolladores de la prueba deben especificar y documentar disposiciones que se hayan hecho para la administración de la prueba y los procedimientos de calificación para eliminar obstáculos irrelevantes del constructo para todos los subgrupos relevantes en la población de examinandos.

Comentario: Los desarrolladores de la prueba deben especificar cómo se minimizaron los obstáculos irrelevantes del constructo en el proceso de desarrollo de la prueba para individuos de todos los subgrupos relevantes en la población prevista de la prueba. Los desarrolladores y/o usuarios de la prueba también deben documentar cualquier estudio llevado a cabo para examinar la confiabilidad/precisión de los puntajes y la validez de las interpretaciones de los evaluadores para subgrupos

relevantes de la población prevista de examinandos para los usos previstos de los puntajes de la prueba. Los procedimientos especiales de administración, calificación y presentación de reportes de la prueba deben documentarse y ponerse a disposición de los usuarios de la prueba.

Unidad 2. Validez de las interpretaciones de los puntajes de la prueba para los usos previstos para la población prevista de individuos examinados

Estándar 3.6

Cuando evidencia creíble indique que los puntajes de la prueba pueden diferir en significado para subgrupos relevantes de la población prevista de individuos examinados, los desarrolladores y/o usuarios de la prueba son responsables de examinar la evidencia de validación de las interpretaciones de los puntajes para los usos previstos para individuos de esos subgrupos. Las leyes aplicables pueden definir lo que constituye una diferencia significativa en los puntajes de los subgrupos y qué acciones se llevan a cabo en respuesta a dichas diferencias.

Comentario: Las diferencias medias de los subgrupos no indican de por sí falta de imparcialidad, pero esas diferencias deberían dar lugar a estudios de seguimiento, cuando sean viables, para identificar las posibles causas de esas diferencias. Dependiendo de que las diferencias de subgrupos se descubran durante la fase de desarrollo o de uso, ya sea el desarrollador de la prueba o el usuario de la prueba es responsable de iniciar las averiguaciones de seguimiento y, según corresponda, los estudios relevantes. La averiguación debe investigar infrarrepresentación de constructo y fuentes de varianza irrelevante de constructo como posibles causas de diferencias de los subgrupos, investigadas según sea viable, mediante estudios cuantitativos y/o cualitativos. Las clases de evidencia de validación consideradas pueden incluir análisis de contenido de la prueba, estructura interna de respuestas de la prueba, la relación de los puntajes de la prueba con otras variables, o los procesos de respuesta empleados por cada individuo examinado. Cuando los tamaños de la muestra sean suficientes, también deben realizarse estudios de precisión y exactitud de puntajes para subgrupos relevantes. Cuando los tamaños de la muestra sean pequeños, a veces pueden acumularse datos con las administraciones operativas de la prueba de modo que puedan realizarse análisis cuantitativos adecuados por subgrupo después de que la prueba haya estado en uso durante un período de tiempo. Los estudios cualitativos también son relevantes para los argumentos de validez de respaldo (p. ej., revisiones de expertos, grupos focales, laboratorios cognitivos). Los desarrolladores de la prueba deben considerar detenidamente las conclusiones de los análisis cuantitativos y/o cualitativos al documentar las interpretaciones para los usos previstos de los puntajes, así como también en las revisiones de pruebas posteriores.

Los análisis, cuando sea posible, pueden necesitar tener en cuenta el nivel de heterogeneidad dentro de subgrupos relevantes, por ejemplo, individuos con diferentes discapacidades, o individuos examinados de minorías lingüísticas en diferentes niveles de competencia en inglés. Las diferencias dentro de estos subgrupos pueden influir en lo adecuado que resulte el contenido de la prueba, la estructura interna de las repuestas a la prueba, la relación de los puntajes de la prueba con otras variables, o los procesos de respuesta empleados por cada individuo examinado.

Estándar 3.7

Cuando la evidencia de validación relacionada con criterios se utiliza como base para predicciones de desempeño futuro basadas en puntajes de la prueba y los tamaños de la muestra son suficientes, los desarrolladores y/o usuarios de la prueba son responsables de evaluar la posibilidad de predicción diferencial para subgrupos relevantes para los que existe evidencia previa o teoría que sugiera predicción diferencial.

Comentario: Cuando los tamaños de la muestra son suficientes, la predicción diferencial suele examinarse utilizando análisis de regresión. Un enfoque al análisis de regresión examina las diferencias de pendiente e intersección entre dos grupos de destino (p. ej., muestras de negros y blancos), mientras que otro examina las desviaciones sistemáticas de una línea de regresión común para los grupos de interés. Ambos enfoques pueden tener en cuenta la posibilidad de sesgo predictivo y/o diferencias en heterogeneidad entre grupos y proporcionar información valiosa para el examen de predicciones diferenciales. Por el contrario, los coeficientes de correlación proporcionan evidencia inadecuada a favor o en contra de una hipótesis de predicción diferencial si se determina que los grupos tienen medias y varianzas desiguales en la prueba y en el criterio. Es particularmente importante en el contexto de evaluación para fines de alto riesgo que los desarrolladores y/o usuarios de la prueba examinen la predicción diferencial y eviten el uso de coeficientes de correlación en situaciones en las que los grupos o tratos den lugar a medias o varianzas desiguales en la prueba y el criterio.

Estándar 3.8

Cuando las pruebas requieran la calificación de respuestas construidas, los desarrolladores y/o usuarios de la prueba deben reunir y reportar evidencia de la validez de las interpretaciones de puntajes para subgrupos relevantes en la población prevista de examinandos para los usos previstos de los puntajes de la prueba.

Comentario: Las diferencias de los subgrupos en las respuestas de los individuos examinados y/o las expectativas y percepciones de los evaluadores pueden introducir varianza irrelevante de constructo en los puntajes de pruebas de respuestas construidas. Estas, a su vez, podrían afectar seriamente la confiabilidad/precisión, validez y comparabilidad de las interpretaciones de los puntajes para los usos previstos para algunos individuos. Diferentes métodos de calificación podrían influir de manera diferencial en la representación de puntajes del constructo para individuos de algunos subgrupos.

Para la calificación realizada por seres humanos, los procedimientos de calificación deben diseñarse con la intención de que los puntajes reflejen la situación del individuo examinado en relación con los constructos evaluados y no estén influenciados por las percepciones y predisposiciones personales de los evaluadores. Es esencial que se realice y supervise la capacitación y calibración adecuadas de los evaluadores en todo el proceso de calificación para respaldar la coherencia de calificaciones de los evaluadores para individuos de subgrupos relevantes. Cuando los tamaños de la muestra lo permitan, la precisión y exactitud de puntajes para subgrupos relevantes también debería calcularse.

Se pueden usar algoritmos de puntaje automático para calificar respuestas construidas complejas, como ensayos, ya sea como único determinador del puntaje o en conjunto con un puntaje proporcionado por un evaluador humano. Los algoritmos de calificación deben revisarse para detectar posibles fuentes de sesgo. La precisión de puntajes y validez de interpretaciones de puntajes resultantes de puntajes automáticos deben evaluarse para todos los subgrupos relevantes de la población prevista.

Unidad 3. Adecuaciones para eliminar obstáculos irrelevantes del constructo y respaldar interpretaciones válidas de puntajes para sus usos previstos

Estándar 3.9

Los desarrolladores de la prueba y/o los usuarios de la prueba son responsables de desarrollar y proporcionar adecuaciones de la prueba, cuando corresponda y sea viable, para eliminar obstáculos irrelevantes del constructo que de otro modo interferirían con la capacidad de los individuos examinados de demostrar su situación respecto de los constructos de destino.

Comentario: Las adecuaciones de la prueba están diseñadas para eliminar obstáculos irrelevantes del constructo relacionados con características individuales que de otro modo interferirían con la medición del constructo de destino y por lo tanto desfavorecerían injustamente a individuos con estas características. Estas adecuaciones incluyen cambios en el contexto de administración, presentación, interfaz/compromiso, y requisitos de respuesta, y pueden incluir la incorporación de individuos al proceso de administración (p. ej., lectores, copistas).

Una adecuación apropiada es aquella que responde a características individuales específicas, pero lo hace de una manera que no cambia el constructo que está midiendo la prueba ni el significado de los puntajes. Los desarrolladores de la prueba y/o los usuarios de la prueba deben documentar el fundamento para la conclusión de que la adecuación no cambia el constructo que está midiendo la prueba. Las adecuaciones deben abordar necesidades específicas de cada examinando (p. ej., cognitivas, lingüísticas, sensoriales, físicas) y pueden ser requeridas por ley. Por ejemplo, es posible que las personas que no son completamente competentes en inglés necesiten adecuaciones lingüísticas que aborden su condición en cuanto a la lengua, mientras que individuos con problemas de la vista pueden necesitar el aumento del tamaño del texto. En muchos casos, cuando se utiliza una prueba para evaluar el progreso académico de un individuo, la adecuación que mejor eliminará la irrelevancia de constructo corresponderá a la adecuación utilizada para la instrucción.

Las modificaciones de la prueba que cambian el constructo que la prueba está midiendo pueden ser necesarias para que algunos individuos examinados demuestren su situación respecto de algún aspecto del constructo previsto. Si una evaluación se modifica para mejorar el acceso al constructo previsto para individuos designados, la evaluación modificada debería tratarse como una evaluación recientemente desarrollada que necesita adherir a los estándares de la prueba para validez, confiabilidad/precisión, imparcialidad, etc.

Estándar 3.10

Cuando se permitan adecuaciones de la prueba, los desarrolladores de la prueba y/o usuarios de la prueba son responsables de documentar disposiciones estándares para usar la adecuación y para supervisar la implementación apropiada de la adecuación.

Comentario: Las adecuaciones de la prueba deben utilizarse solo cuando el examinando tenga una necesidad documentada de la adecuación, por ejemplo, un Plan Educativo Individualizado (IEP, por sus siglas en inglés) o documentación de un médico, psicólogo, u otro profesional cualificado. La documentación debe prepararse con antelación a la experiencia de realización de la prueba y debe ser revisada por uno o más expertos cualificados para tomar una decisión sobre la relevancia de la documentación con respecto a la adecuación solicitada.

Los desarrolladores y/o usuarios de la prueba deben proporcionar a los individuos que requieren adecuaciones en una situación de evaluación información sobre la disponibilidad de adecuaciones y los procedimientos para solicitarlas antes de la administración de la prueba. En contextos en que las adecuaciones se proporcionen habitualmente para individuos con necesidades documentadas (p. ej., contextos educativos), la documentación debe describir adecuaciones aceptables e incluir protocolos y/o procedimientos estandarizados para identificar a los individuos examinados elegibles, identificar y asignar adecuaciones apropiadas para estos individuos, y administrar adecuaciones, calificar y presentar reportes de conformidad con reglas estandarizadas.

Los administradores y usuarios de la prueba deben también proporcionar a quienes cumplen un rol en la determinación y administración de adecuaciones suficiente información y conocimientos para usar apropiadamente las adecuaciones que puedan aplicarse a la evaluación. Las instrucciones para administrar cualquier cambio en la prueba o procedimientos de evaluación deben documentarse claramente y, cuando sea necesario, los administradores de la prueba deben

capacitarse para seguir estos procedimientos. El administrador de la prueba debe administrar las adecuaciones de una manera estandarizada según lo documentado por el desarrollador de la prueba. Los procedimientos de administración deben incluir procedimientos para registrar qué adecuaciones se utilizaron para individuos específicos y, cuando corresponda, para registrar cualquier desviación de procedimientos estandarizados para administrar las adecuaciones.

El administrador de la prueba o el representante correspondiente del usuario de la prueba debe documentar cualquier uso de adecuaciones. Para evaluaciones educativas a gran escala, los usuarios de la prueba deben supervisar el uso apropiado de adecuaciones.

Estándar 3.11

Cuando una prueba se cambia para eliminar obstáculos a la accesibilidad del constructo sometido a medición, los desarrolladores y/o usuarios de la prueba son responsables de obtener y documentar evidencia de la validez de las interpretaciones de los puntajes para los usos previstos de la prueba cambiada, cuando los tamaños de la muestra lo permitan.

Comentario: Es aconsejable, cuando sea viable y corresponda, hacer una prueba piloto y/o de campo de cualquier alteración en la prueba con individuos que representen a cada subgrupo relevante para el que está prevista la alteración. Los estudios de validez por lo general deben investigar tanto la eficacia de la alteración para los subgrupos previstos como la comparabilidad de las inferencias de puntaje de las pruebas alteradas y originales.

En algunas circunstancias, es posible que los desarrolladores no puedan obtener suficientes muestras de individuos, por ejemplo, aquellos con la misma discapacidad o niveles similares de una discapacidad, para realizar análisis empíricos estándares de confiabilidad/precisión y validez. En estas situaciones, deben buscarse maneras alternativas para evaluar la validez de la prueba cambiada para subgrupos relevantes, por ejemplo, a través de estudios cualitativos de muestras pequeñas o juicios profesionales que examinen la comparabilidad de las pruebas originales y alteradas y/o que investiguen explicaciones alternativas para el desempeño en las pruebas cambiadas.

Debe proporcionarse evidencia para las alteraciones recomendadas. Si el desarrollador de la prueba recomienda diferentes límites de tiempo, por ejemplo, para individuos con discapacidades o para aquellos con características lingüísticas y culturales diversas, deben utilizarse pruebas piloto o de campo, cuando sea posible, para establecer estos límites de tiempo en particular más que simplemente permitir a los examinandos un múltiplo del tiempo estándar sin examinar la utilidad de la implementación arbitraria de múltiplos del tiempo estándar. Cuando sea posible, deben investigarse la fatiga y otras cuestiones relacionadas con el tiempo como factores potencialmente importantes cuando se extienden los límites de tiempo.

Cuando las pruebas se simplifican desde el punto de vista lingüístico para eliminar la varianza irrelevante de constructo, los desarrolladores y/o usuarios de la prueba son responsables de documentar evidencia de la comparabilidad de puntajes de las pruebas lingüísticamente simplificadas con la prueba original, cuando los tamaños de la muestra lo permitan.

Estándar 3.12

Cuando una prueba se traduce y adapta de un idioma a otro, los desarrolladores de la prueba y/o usuarios de la prueba son responsables de describir los métodos utilizados al establecer la adecuación de la adaptación y documentar la evidencia empírica y lógica para la validez de las interpretaciones de los puntajes de la prueba para el uso previsto.

Comentario: El término *adaptación* se utiliza aquí para describir cambios hechos a pruebas traducidas de un idioma a otro para reducir la varianza irrelevante de constructo que puede surgir debido a características individuales o de subgrupos. En este caso, el proceso de traducción/

adaptación involucra no solo traducir el idioma de la prueba de modo que sea adecuado para el subgrupo que realiza la prueba, sino también abordar cualquier característica del subgrupo lingüística o cultural irrelevante del constructo que pueda interferir con la medición de los constructos previstos. Cuando versiones en múltiples idiomas de una prueba tienen por objeto proporcionar puntajes comparables, los desarrolladores de la prueba deben describir en detalle los métodos utilizados para la traducción y la adaptación de la prueba y deben reportar evidencia de la validez de los puntajes de la prueba pertinente a los grupos lingüísticos y culturales para los que está prevista la prueba y pertinente a los usos previstos de los puntajes. La evidencia de validación puede incluir estudios empíricos y/o juicio profesional que documente que las versiones en diferentes idiomas miden constructos comparables o similares y que las interpretaciones de los puntajes de las dos versiones tienen validez comparable para sus usos previstos. Por ejemplo, si una prueba se traduce y adapta al español para usarse con poblaciones centroamericanas, cubanas, mexicanas, portorriqueñas, sudamericanas y españolas, la validez de las interpretaciones de los puntajes de la prueba para usos específicos debe evaluarse con miembros de cada uno de estos grupos por separado, cuando sea viable. Cuando los tamaños de la muestra lo permitan, debe proporcionarse evidencia de la exactitud y precisión de los puntajes para cada grupo, y las propiedades de la prueba para cada grupo deben incluirse en los manuales de la prueba.

Estándar 3.13

Una prueba debe administrarse en el idioma que sea más relevante y apropiado para la finalidad de la prueba.

Comentario: Los usuarios de la prueba deben tener en cuenta las características lingüísticas y culturales y las competencias en idioma relativas de los individuos examinados que son bilingües o utilizan varios idiomas. Identificar el o los idiomas más apropiados para la evaluación también requiere una consideración atenta del contexto y la finalidad de la evaluación. Excepto en casos en los que la finalidad de la evaluación sea determinar el nivel de competencia de los examinandos en un idioma en particular, los examinandos deben evaluarse en el idioma en el que tienen mayor competencia. En algunos casos, el idioma en el que los examinandos tienen mayor competencia en general puede no ser el idioma en el que recibieron instrucción o capacitación en relación con los constructos evaluados, y en estos casos es posible que se sea apropiado administrar la prueba en el idioma de instrucción.

Debe emplearse el juicio profesional para determinar los procedimientos más apropiados para establecer las competencias en idioma relativas. Esos procedimientos pueden variar desde autoidentificación por parte de los individuos examinados hasta pruebas formales de competencia en idioma. La sensibilidad a características lingüísticas y culturales puede requerir el uso exclusivo de un idioma en la evaluación o el uso de múltiples idiomas para minimizar la introducción de componentes irrelevantes del constructo en el proceso de medición.

La determinación del idioma en el que el examinando tiene mayor competencia para la administración de la prueba no garantiza automáticamente la validez de las inferencias de puntajes para el uso previsto. Por ejemplo, los individuos pueden tener mayor competencia en un idioma que en otro, pero no ser necesariamente competentes desde el punto de vista del desarrollo en cualquiera de los dos; las desconexiones entre el idioma de adquisición del constructo y el de la evaluación también pueden comprometer la interpretación apropiada de los puntajes del examinando.

Estándar 3.14

Cuando la prueba requiere el uso de un intérprete, el intérprete debe seguir procedimientos estandarizados y, en la medida en que sea viable, tener un nivel suficientemente bueno en el idioma y contenido de la prueba y la lengua nativa y la cultura del individuo examinado

para traducir la prueba y los materiales de evaluación relacionados y explicar las respuestas de la prueba del individuo examinado, según sea necesario.

Comentario: Si bien los individuos con competencia limitada en el idioma de la prueba (incluyendo individuos sordos y con dificultades auditivas cuya lengua nativa puede ser la lengua de señas) idealmente deben ser evaluados por examinadores bilingües/biculturales profesionalmente capacitados, el uso de un intérprete puede ser necesario en algunas situaciones. Si se requiere un intérprete, el usuario de la prueba es responsable de seleccionar un intérprete con cualificaciones, experiencia y preparación razonables para ayudar apropiadamente en la administración de la prueba. Al igual que con otros aspectos de la evaluación estandarizada, los procedimientos para administrar una prueba cuando se utiliza un intérprete deben estandarizarse y documentarse. Es necesario que el intérprete comprenda la importancia de seguir procedimientos estandarizados para esta prueba, la importancia de transmitir exactamente al examinador las respuestas reales de un individuo examinado, y el rol y las responsabilidades del intérprete en la evaluación. Cuando la traducción de términos técnicos sea importante para evaluar con exactitud el constructo, el intérprete debe estar familiarizado con el significado de estos términos y los vocabularios correspondientes en los idiomas respectivos.

A menos que la prueba se haya estandarizado y normalizado con el uso de intérpretes, su uso puede necesitar ser visto como una alteración que podría cambiar la medición del constructo previsto, en particular debido a la introducción de un tercero durante la evaluación, así como la modificación del protocolo estandarizado. Las diferencias en el significado, familiaridad, frecuencia, connotaciones y asociaciones de las palabras hacen que sea difícil comparar directamente puntajes de cualquier traducción no estandarizada con las normas de la lengua inglesa.

Cuando es probable que la prueba requiera el uso de intérpretes, el desarrollador de la prueba debe proporcionar orientación clara sobre cómo deben seleccionarse los intérpretes y su rol en la administración.

Unidad 4. Protecciones contra interpretaciones inapropiadas de los puntajes para los usos previstos

Estándar 3.15

Los desarrolladores y editores de la prueba que afirman que una prueba puede ser usada con individuos examinados de subgrupos específicos son responsables de proporcionar la información necesaria para respaldar interpretaciones apropiadas de puntajes de la prueba para sus usos previstos para individuos de estos subgrupos.

Comentario: Los desarrolladores de la prueba deben incluir en los manuales de la prueba e instrucciones para la interpretación de puntajes declaraciones explícitas sobre la aplicabilidad de la prueba para subgrupos relevantes. Los desarrolladores de la prueba deben proporcionar evidencia de la aplicabilidad de la prueba para subgrupos relevantes y hacer advertencias explícitas contra usos indebidos previsibles (basadas en experiencia previa u otras fuentes relevantes como bibliografía de investigación) de los resultados de la prueba.

Estándar 3.16

Cuando investigación creíble indique que los puntajes de la prueba para algunos subgrupos relevantes se ven diferencialmente afectados por características irrelevantes del constructo de la prueba o de los individuos examinados, cuando sea legalmente aceptable, los usuarios de la prueba deben utilizar la prueba solo para esos subgrupos para los que existe evidencia suficiente de validez para respaldar las interpretaciones de los puntajes para los usos previstos.

Comentario: Una prueba no puede medir los mismos constructos para individuos de diferentes subgrupos relevantes porque diferentes características del contenido o formato de la prueba

influyen en los puntajes de los examinandos de un subgrupo a otro. Cualquiera de esas diferencias puede favorecer o desfavorecer involuntariamente a individuos de estos subgrupos. La decisión en cuanto usar una prueba con cualquier subgrupo relevante dado involucra necesariamente un análisis detenido de la evidencia de validación para el subgrupo, como se requiere en el Estándar 1.4. La decisión también requiere consideración de los requisitos legales aplicables y el ejercicio de juicio profesional profundo respecto de la significación de cualquier componente irrelevante del constructo. En los casos en que existe evidencia creíble de validez diferencial, los desarrolladores deben proporcionar orientación clara al usuario de la prueba sobre cuándo y si las interpretaciones válidas de los puntajes para sus usos previstos pueden o no pueden extraerse para individuos de estos subgrupos.

Es posible que existan ocasiones en que los individuos examinados soliciten o exijan que se tome una versión de la prueba distinta de la considerada más apropiada por el desarrollador o usuario. Por ejemplo, un individuo con una discapacidad puede rechazar un formato alterado y solicitar el formulario estándar. Acceder a tales solicitudes, después de informar completamente al individuo examinado sobre las características de la prueba, las adecuaciones que están disponibles, y cómo se utilizarán los puntajes de la prueba, no es una violación de este estándar y en algunos casos puede ser requerido por ley.

En algunos casos, como cuando una prueba distribuirá beneficios o cargas (como reunir requisitos para una clase para estudiantes sobresalientes o la denegación de una promoción en un empleo), la ley puede limitar la medida en que un usuario de la prueba puede evaluar a algunos grupos conforme a la prueba y a otros grupos conforme a una prueba diferente.

Estándar 3.17

Cuando se informen públicamente puntajes agregados para subgrupos relevantes —por ejemplo, hombres y mujeres, individuos de diferente nivel socioeconómico, individuos que difieren en cuanto a raza/origen étnico, individuos con diferentes orientaciones sexuales, individuos con características lingüísticas y culturales diversas, individuos con discapacidades, niños pequeños o adultos mayores— los usuarios de la prueba son responsables de proporcionar evidencia de comparabilidad y de incluir declaraciones de advertencia cuando la investigación creíble o la teoría indique que es posible que los puntajes de la prueba no tengan significado comparable entre estos subgrupos.

Comentario: Reportar puntajes para subgrupos relevantes se justifica solo si los puntajes tienen significado comparable entre estos grupos y existe un tamaño de la muestra suficiente por grupo para proteger la identidad individual y justificar la agregación. Este estándar tiene por objeto ser aplicable a contextos en los que los puntajes se presenten implícita o explícitamente como comparables en significado entre subgrupos. Se debe tener la precaución de que los términos utilizados para describir subgrupos reportados se definan claramente, de conformidad con el uso común, y sean comprendidos claramente por quienes interpretan los puntajes de la prueba.

La terminología para describir subgrupos específicos para los que pueden y no pueden extraerse inferencias válidas de puntajes de la prueba debe ser lo más precisa posible, y las categorías deben ser coherentes con los usos previstos de los resultados. Por ejemplo, los términos *latino* o *hispano* pueden ser ambiguos si no se definen específicamente, en el sentido de que pueden denotar individuos de origen cubano, mexicano, portorriqueño, sudamericano o centroamericano o de otra cultura hispana, independientemente de la raza/origen étnico, y pueden combinar a quienes son inmigrantes recientes con quienes son nativos nacidos en EE. UU., quienes pueden no ser competentes en inglés, y quienes son de un nivel socioeconómico diverso. De manera similar, el término "individuos con discapacidades" abarca una amplia variedad de afecciones y características de antecedentes específicas. Incluso las referencias a categorías específicas de individuos con discapacidades, como problemas auditivos, deben estar

acompañadas de una explicación del significado del término y una indicación de la variabilidad de individuos dentro del grupo.

Estándar 3.18

En la evaluación de individuos para fines de diagnóstico y/o colocación en un programa especial, los usuarios de la prueba no deben usar puntajes de la prueba como los únicos indicadores para caracterizar el funcionamiento, la competencia, las actitudes y/o las predisposiciones de un individuo. En cambio, deben utilizarse múltiples fuentes de información, deben considerarse explicaciones alternativas para el desempeño en la prueba, y el juicio profesional de alguien familiarizado con la prueba debe aplicarse a la decisión.

Comentario: Muchos manuales de prueba señalan variables que deberían considerarse en la interpretación de los puntajes de la prueba, como antecedentes clínicamente relevantes, medicamentos, registro escolar, estado vocacional y motivación del examinando. Las influencias asociadas con variables tales como edad, cultura, discapacidad, género y características lingüísticas o raciales/étnicas también pueden ser relevantes.

La oportunidad de aprendizaje es otra variable que puede ser necesario tener en cuenta en los contextos educativos y/o clínicos. Por ejemplo, si inmigrantes recientes que se evalúan en un inventario de personalidad o una medida de capacidad tienen escasa exposición previa a la escuela, es posible que no hayan tenido la oportunidad de aprender conceptos que la prueba supone son conocimientos comunes o experiencias comunes, incluso si la prueba es administrada en la lengua nativa. No tener en cuenta la oportunidad previa de aprendizaje puede conducir a diagnósticos equivocados, colocaciones y/o servicios inapropiados y consecuencias negativas imprevistas.

Las inferencias sobre la competencia general en idioma de los examinandos deben basarse en pruebas que midan una serie de características del idioma, no una sola habilidad lingüística. Una variedad más completa de capacidades comunicativas (p. ej., conocimiento de palabras, sintaxis, así como variación cultural) por lo general deberán evaluarse. Los usuarios de la prueba son responsables de interpretar puntajes individuales a la luz de explicaciones alternativas y/o variables individuales relevantes observadas en el manual de la prueba.

Estándar 3.19

En contextos en los que la misma autoridad es responsable tanto de la provisión del plan de estudios como de las decisiones de alto riesgo basadas en la evaluación del dominio del plan de estudios por parte de los individuos examinados, estos últimos no deberían sufrir consecuencias negativas permanentes si la evidencia indica que no han tenido la oportunidad de aprender el contenido de la prueba.

Comentario: En contextos educativos, la oportunidad de los estudiantes de aprender el contenido y las habilidades evaluadas por una prueba de rendimiento puede afectar seriamente su desempeño en la prueba y la validez de las interpretaciones de los puntajes de la prueba para el uso previsto para las decisiones individuales de alto riesgo. Si no hay una correspondencia apropiada entre el contenido del plan de estudios y la instrucción y el de los constructos evaluados para algunos estudiantes, no se puede esperar que esos estudiantes salgan bien en la prueba y pueden ser desfavorecidos injustamente por decisiones individuales de alto riesgo, como la denegación de la graduación de la escuela secundaria, que se toman sobre la base de los resultados de la prueba. Cuando una autoridad, como un estado o distrito, es responsable de indicar y/o impartir el plan de estudios y la instrucción, no debe penalizar a los individuos por el desempeño en la prueba en cuanto al contenido que la autoridad no proporcionó.

Obsérvese que este estándar no es aplicable en situaciones en las que diferentes autoridades son responsables del plan de estudios, la evaluación y/o la interpretación y el uso de resultados. Por ejemplo, la oportunidad de aprendizaje puede escapar al conocimiento o control de los usuarios de la prueba, y es posible que no influya en la

validez de las interpretaciones de la prueba como las predicciones de desempeño futuro.

Estándar 3.20

Cuando un constructo puede medirse de diferentes maneras que son iguales en su grado de representación del constructo y validez (incluyendo la ausencia de varianza irrelevante de constructo), los usuarios de la prueba deben considerar, entre otros factores, evidencia de diferencias de los subgrupos en los puntajes medios o en porcentajes de individuos examinados cuyos puntajes excedan los puntajes de corte, en la decisión de qué puntajes de prueba y/o de corte usar.

Comentario: La evidencia de desempeño diferencial de los subgrupos es un factor importante que influye en la elección entre una prueba u otra. Sin embargo, otros factores, como costo, tiempo de evaluación, seguridad de la prueba y cuestiones logísticas (p. ej., la necesidad de cribar cantidades muy grandes de individuos examinados en muy poco tiempo), también deben ser parte de los juicios profesionales sobre la selección y uso de la prueba. Si los puntajes de dos pruebas conducen a interpretaciones igualmente válidas e imponen costos y otras cargas similares, las consideraciones legales pueden requerir seleccionar la prueba que minimice las diferencias de subgrupos. Debe establecerse la articulación clara de cada interpretación prevista de los puntajes de la prueba para un uso especificado, y debe proporcionarse evidencia de validación apropiada que respalde cada interpretación prevista.

PARTE II

II Operaciones

4. DISEÑO Y DESARROLLO DE PRUEBAS

ANTECEDENTES

El *desarrollo de la prueba* es el proceso de producir una medida de algún aspecto del conocimiento, las habilidades, capacidades, intereses, actitudes u otras características de un individuo mediante el desarrollo de preguntas o tareas y la combinación de estas para formar una prueba, según un plan especificado. Los pasos y consideraciones para este proceso se articulan en el plan de diseño de la prueba. El diseño de la prueba comienza con la consideración de interpretaciones esperadas para usos previstos de los puntajes que generará la prueba. El contenido y formato de la prueba luego se especifican para proporcionar evidencia para respaldar las interpretaciones para los usos previstos. El diseño de la prueba también incluye especificación de la administración de la prueba y procedimientos de calificación, y cómo deben reportarse los puntajes. Las preguntas o tareas (en adelante denominados ítems) se desarrollan siguiendo las especificaciones de la prueba y se seleccionan utilizando criterios apropiados a los usos previstos de la prueba. Los procedimientos para calificar ítems individuales y la prueba en conjunto también se desarrollan, revisan y corrigen según sea necesario. El diseño de la prueba es comúnmente iterativo, con ajustes y revisiones que se realizan en respuesta a datos de ensayos y uso operativo.

Los procedimientos de diseño y desarrollo deben respaldar la validez de las interpretaciones de los puntajes de la prueba para sus usos previstos. Por ejemplo, las evaluaciones educativas actuales suelen utilizarse para indicar la competencia de los estudiantes con respecto a estándares para el conocimiento y la habilidad que un estudiante debería mostrar; por lo tanto, la relación entre el contenido de la prueba y los estándares de contenido establecidos es clave. En este caso, las especificaciones de contenido deben describir claramente el contenido y/o las categorías cognitivas que se cubrirán para que pueda reunirse evidencia de la alineación de las preguntas de la prueba con estas categorías. Cuando se prevean interpretaciones normativas, los procedimientos de desarrollo deben incluir una definición precisa de la población de referencia y planes para reunir datos normativos apropiados. Muchas pruebas, como las pruebas de empleo o de selección universitaria, dependen de evidencia de validación predictiva. Las especificaciones para dichas pruebas deben incluir descripciones de los resultados que la prueba se ha diseñado para predecir y planes para reunir evidencia de la efectividad de los puntajes de la prueba en la predicción de estos resultados.

Las cuestiones que inciden en la validez, confiabilidad e imparcialidad se entrelazan dentro de las etapas de desarrollo de la prueba. Cada uno de estos temas se aborda integralmente en otros capítulos de los Estándares: validez en el capítulo 1, confiabilidad en el capítulo 2 e imparcialidad en el capítulo 3. En el capítulo 6 se brinda material adicional sobre la administración y calificación de las pruebas, y sobre la presentación de reportes e interpretación de puntajes y resultados. El capítulo 5 analiza escalas de puntajes y el capítulo 7 cubre requisitos de documentación.

Además, los desarrolladores de la prueba deben respetar los derechos de los participantes en el proceso de desarrollo, incluyendo los participantes de la prueba previa. En particular, los desarrolladores de la prueba deben tomar medidas para garantizar la notificación y el consentimiento adecuados de los participantes y para proteger la información personal de identificación de los participantes de conformidad con los requisitos legales y profesionales aplicables. Los derechos de los examinandos se tratan en el capítulo 8.

Este capítulo describe cuatro fases del proceso de desarrollo de la prueba que abarcan desde la declaración original de la(s) finalidad(es) hasta el producto final: (a) desarrollo y evaluación de las especificaciones de la prueba; (b) desarrollo,

ensayo y evaluación de los ítems; (c) reunión y evaluación de nuevos formularios de la prueba; y (d) desarrollo de procedimientos y materiales para administración y calificación. Lo que sigue es una descripción de los procedimientos de desarrollo típicos de la prueba, aunque puede haber motivos sólidos por los que algunos pasos cubiertos en la descripción se sigan en algunos contextos y no en otros.

Especificaciones de la prueba

Consideraciones generales

En casi todos los casos, el desarrollo de la prueba está guiado por un conjunto de especificaciones de la prueba. La naturaleza de estas especificaciones y el modo en que se crean pueden variar ampliamente como una función de la naturaleza de la prueba y sus usos previstos. El término *especificaciones de la prueba* a veces se limita a la descripción del contenido y formato de la prueba. En los *Estándares*, las especificaciones de la prueba se definen en líneas más generales para incluir también documentación de la finalidad y los usos previstos de la prueba, así como decisiones detalladas sobre contenido, formato, extensión de la prueba, características psicométricas de los ítems y de la prueba, modo de ejecución, administración, calificación, y reporte de puntajes.

La responsabilidad del desarrollo de especificaciones de la prueba también varía ampliamente entre los programas de evaluación. En la mayoría de las pruebas comerciales, las especificaciones de la prueba son creadas por el desarrollador de la prueba. En otros contextos, como las pruebas utilizadas en rendición de cuentas en materia educativa, muchos aspectos de las especificaciones de la prueba se establecen a través del proceso de política pública. Como se analizó en la introducción, el término genérico desarrollador de la prueba se utiliza en este capítulo con preferencia respecto de otros términos, como editor de la prueba, para cubrir tanto a los responsables del desarrollo como a los responsables de la implementación de las especificaciones de la prueba en una amplia variedad de procesos de desarrollo de la prueba.

Declaración de finalidad y usos previstos

El proceso de desarrollar pruebas educativas y psicológicas debe comenzar con una declaración de la(s) finalidad(es) de la prueba, los usuarios y usos previstos, el constructo o dominio de contenido sometido a medición, y la población prevista de individuos examinados. Las pruebas del mismo constructo o dominio pueden diferir de maneras importantes porque factores tales como finalidad, usos previstos y población de individuos examinados pueden variar. Además, las pruebas previstas para diversas poblaciones de individuos examinados deben ser desarrolladas para minimizar los factores irrelevantes del constructo que puedan deprimir o inflar injustamente el desempeño de algunos individuos examinados. En muchos casos, es posible que deban especificarse adecuaciones y/o versiones alternativas de las pruebas para eliminar obstáculos irrelevantes al desempeño para subgrupos en particular en la población prevista de individuos examinados.

La especificación de los usos previstos incluirá una indicación de que las interpretaciones de los puntajes de la prueba son principalmente *conformes a normas* o *conformes a criterios*. Cuando los puntajes son conformes a normas, las interpretaciones de puntajes relativas son de principal interés. Un puntaje para un individuo o para un grupo definible se clasifica dentro de una distribución de puntajes o se compara con el desempeño promedio de examinandos en una población de referencia (p. ej., basada en edad, grado, categoría de diagnóstico o clasificación del trabajo). Cuando las interpretaciones son conformes a criterios, las interpretaciones de puntajes absolutas son de principal interés. El significado de dichos puntajes no depende de la información de clasificación. En cambio, el puntaje de la prueba transmite directamente un nivel de competencia en algún dominio de criterios definido. Tanto las interpretaciones relativas como absolutas suelen utilizarse con una prueba dada, pero el desarrollador de la prueba determina qué enfoque es el más relevante para los usos específicos de la prueba.

Especificaciones de contenido

El primer paso en el desarrollo de especificaciones de la prueba es extender la declaración de finalidad(es), y el constructo o dominio de contenido que se considera, en un marco para la prueba que describa el grado del dominio, o el alcance del constructo sometido a medición. Las *especificaciones de contenido* a veces denominadas *marcos de contenido*, delinean los aspectos (p. ej., contenido, habilidades, procesos y características de diagnóstico) del constructo o dominio sometido a medición. Las especificaciones deben abordar preguntas sobre qué debe incluirse, como "¿Las matemáticas de octavo grado incluyen álgebra?", "¿La capacidad verbal incluye comprensión de textos y vocabulario?", "¿La autoestima incluye tanto sentimientos como actos?". La delineación de las especificaciones de contenido puede orientarse por la teoría o por un análisis del dominio de contenido (p. ej., un análisis de los requisitos del puesto en el caso de muchas pruebas de acreditación y empleo). Las especificaciones de contenido sirven como una guía para la evaluación de pruebas posteriores. El capítulo sobre validez proporciona un análisis más profundo de las relaciones entre el constructo o dominio de contenido, el marco de la prueba, y la(s) finalidad(es) de la prueba.

Especificaciones de formato

Una vez que se han tomado decisiones sobre qué debe medir la prueba y qué significado tienen por objeto transmitir sus puntajes, el próximo paso es crear especificaciones de formato. Las especificaciones de formato delinean el formato de ítems (es decir, tareas o preguntas); el formato de respuesta o condiciones para responder; y el tipo de procedimientos de calificación. Si bien las decisiones de formato a menudo están impulsadas por consideraciones de conveniencia, como la facilidad de respuesta o el costo de calificación, las consideraciones de validez no deben pasarse por alto. Por ejemplo, si las preguntas de la prueba requieren que los examinandos posean una habilidad lingüística significativa para interpretarlas pero la prueba no tiene la intención de constituir una medida de habilidad lingüística, la complejidad de las preguntas puede conducir a varianza irrelevante de constructo en los puntajes de la prueba. Esto sería injusto para los examinandos con habilidades lingüísticas limitadas, lo que reduce la validez de los puntajes de la prueba como una medida del contenido previsto. Las especificaciones de formato deben incluir una justificación respecto de cómo el formato elegido respalda la validez, confiabilidad e imparcialidad de los usos previstos de los puntajes resultantes.

La naturaleza de los formatos de los ítems y respuestas que pueden especificarse depende de las finalidades de la prueba, el dominio definido de la prueba y la plataforma de evaluación. Los formatos de respuesta seleccionada, como ítems de verdadero-falso o de opciones múltiples, son adecuados para muchas finalidades de evaluación. Las pruebas basadas en computadora permiten diferentes maneras de indicar respuestas, como arrastrar y soltar. Otras finalidades pueden cumplirse de manera más efectiva mediante un formato de respuesta corta. Los ítems de respuesta corta requieren una respuesta de no más de algunas palabras. Los formatos de respuesta extendida requieren que el examinando escriba una respuesta más extensa de una o más oraciones o párrafos. Las evaluaciones de desempeño a menudo buscan emular el contexto o las condiciones en las que efectivamente se aplican el conocimiento o las habilidades previstas. Un tipo de evaluación de desempeño, por ejemplo, es la muestra estandarizada de empleo o trabajo en la que una tarea se presenta al examinando en un formato estandarizado en condiciones estandarizadas. Las muestras de empleo o trabajo podrían incluir la evaluación de la capacidad de un profesional médico para hacer un diagnóstico exacto y recomendar tratamiento para una afección definida, la capacidad de un gerente para articular metas para una organización, o la competencia de un estudiante en la realización de un experimento de laboratorio de ciencias.

Accesibilidad de los formatos de ítems. Como se describe en el capítulo 3, diseñar pruebas para que sean accesibles y válidas para todos los individuos

examinados previstos, en la máxima medida posible, es fundamental. Los formatos que pueden no ser conocidos para algunos grupos de examinandos o que presentan exigencias inapropiadas deben evitarse. Los principios del *diseño universal* describen el uso de formatos de prueba que permiten tomar pruebas sin adaptación a la variedad más amplia posible de individuos, pero no necesariamente eliminan la necesidad de adaptaciones. Las especificaciones de formato deben incluir la consideración de formatos alternativos que también podrían ser necesarios para eliminar obstáculos irrelevantes al desempeño, como letra grande o formato braille para individuos examinados que tienen problemas de la vista o, cuando corresponda al constructo sometido a medición, diccionarios bilingües para examinandos que son más competentes en un idioma que no es el idioma de la prueba. La cantidad y tipos de adaptaciones a especificarse dependen tanto de la naturaleza del constructo que se evalúa como de la población de destino de examinandos.

Formatos de ítems complejos. Algunos programas de evaluación emplean formatos de ítems más complejos. Los ejemplos incluyen evaluaciones de desempeño, simulaciones y porfolios. Las especificaciones para formatos de ítems más complejos deben describir el dominio del que se toman muestras de ítems o tareas, componentes del dominio que se evaluará mediante las tareas o ítems, y características críticas de los ítems que deberían replicarse en la creación de ítems para formularios alternativos. Consideraciones especiales para formatos de ítems complejos se describen en el siguiente análisis de evaluaciones de desempeño, simulaciones y porfolios.

Evaluaciones de desempeño. Las evaluaciones de desempeño requieren que los individuos examinados demuestren la capacidad de desempeñar tareas que a menudo son complejas en su naturaleza y por lo general requieren que los examinandos demuestren sus capacidades o habilidades en contextos que se asemejan mucho a situaciones de la vida real. Una distinción entre evaluaciones de desempeño y otras formas de pruebas es el tipo de respuesta que se requiere de los examinandos. Las evaluaciones de desempeño requieren que los examinandos lleven a cabo un proceso tal como tocar un instrumento musical o afinar el motor de un auto o crear un producto como un ensayo escrito. Una evaluación de un psicólogo clínico en capacitación puede requerir que el examinando entreviste a un cliente, elija pruebas apropiadas, llegue a un diagnóstico y planifique la terapia.

Debido a que las evaluaciones de desempeño habitualmente consisten en una pequeña cantidad de tareas, establecer el grado en que los resultados pueden generalizarse a un dominio más amplio descripto en las especificaciones de la prueba es especialmente importante. Las especificaciones de la prueba deben indicar dimensiones críticas a medir (p. ej., habilidades y conocimiento, procesos cognitivos, contexto para realizar las tareas) de modo que las tareas seleccionadas para la evaluación representen sistemáticamente las dimensiones críticas, lo que conduce a una cobertura integral del dominio, así como cobertura coherente entre los formularios de prueba. La especificación del dominio a cubrir es también importante para aclarar fuentes posiblemente irrelevantes de variación en el desempeño. Además, tanto la evidencia teórica como la empírica son importantes para documentan la medida en que las evaluaciones de desempeño —tareas como así también criterios de calificación— reflejan los procesos o habilidades que son especificados por la definición del dominio. Cuando las tareas se diseñan para suscitar procesos cognitivos complejos, los análisis detallados de las tareas y criterios de calificación y análisis tanto teóricos como empíricos de los desempeños de los examinandos en las tareas proporcionan la evidencia de validación necesaria.

Simulaciones. Las evaluaciones de simulación son similares a las evaluaciones de desempeño en cuanto a que requieren que el individuo examinado se involucre en un conjunto complejo de comportamientos durante un período especificado. Las simulaciones a veces reemplazan a las evaluaciones de desempeño, cuando el desempeño real de la tarea podría ser costoso o peligroso. Las especificaciones para tareas de simulación deben describir el dominio de actividades a ser cubierto

por las tareas, dimensiones críticas de desempeño a reflejarse en cada tarea, y consideraciones de formato específicas como la cantidad o duración de las tareas y aspectos esenciales de cómo interactúa el usuario con las tareas. Las especificaciones deben ser suficientes para permitir que los expertos juzguen la comparabilidad de diferentes conjuntos de tareas de simulación incluidas en formularios alternativos.

Porfolios. Los porfolios son recopilaciones sistemáticas de productos educativos o de trabajo, por lo general reunidos a lo largo del tiempo. El diseño de una evaluación de porfolio, al igual que el de otros procedimientos de evaluación, debe surgir de la finalidad de la evaluación. Las finalidades típicas incluyen juzgar la mejora en el desempeño laboral o educativo y la evaluación de la elegibilidad para un empleo, promoción o graduación. Las especificaciones del porfolio indican la naturaleza del trabajo que ha de incluirse en el porfolio. El porfolio puede incluir entradas tales como productos representativos, el mejor trabajo del examinando, o indicadores de progreso. Por ejemplo, en un contexto laboral que involucra decisiones de promoción, se puede instruir a los empleados para que incluyan sus mejores productos o trabajo. Alternativamente, si la finalidad es juzgar el crecimiento educativo de los estudiantes, se puede pedir a los estudiantes que proporcionen evidencia de mejora con respecto a competencias o habilidades en particular. También se puede pedir a los estudiantes que proporcionen justificaciones para sus elecciones o una nota de presentación que refleje el trabajo presentado y lo que el estudiante ha aprendido de ello. Otros métodos pueden requerir el uso de videos, exhibiciones o demostraciones.

Las especificaciones para el porfolio indican quién es responsable de seleccionar sus contenidos. Por ejemplo, las especificaciones deben indicar si el examinando, el examinador o ambas partes que trabajan juntas deben involucrarse en la selección de los contenidos del porfolio. Las responsabilidades particulares de cada parte se delinean en las especificaciones. En contextos laborales, los empleados pueden involucrarse en la selección de su trabajo y productos que demuestren sus competencias para fines de promoción. De manera análoga, en aplicaciones educativas, los estudiantes pueden participar en la selección de parte de su trabajo y los productos a incluir en sus porfolios.

Las especificaciones en cuanto a cómo se califican los porfolios y quién los califica variarán como una función del uso de los puntajes del porfolio. La evaluación centralizada de porfolios es común cuando estos se utilizan en decisiones de alto riesgo. Cuanto más estandarizados sean los contenidos y procedimientos para recopilar y calificar el material, más comparables serán los puntajes de los porfolios resultantes. Independientemente de los métodos usados, todas las evaluaciones de desempeño, simulaciones y porfolios se evalúan según los mismos estándares de calidad técnica que de otras formas de pruebas.

Extensión de la prueba

Los desarrolladores de la prueba con frecuencia siguen *proyectos básicos* de prueba que especifican la cantidad de ítems para cada área de contenido que se incluirá en cada formulario de prueba. Las especificaciones para la extensión de la prueba deben equilibrar los requisitos de tiempo de evaluación con la precisión de los puntajes resultantes; las pruebas más largas generalmente conducen a puntajes más precisos. Los desarrolladores de la prueba con frecuencia siguen proyectos básicos de prueba que proporcionan orientación sobre la cantidad o porcentaje de ítems para cada área de contenido y que también pueden incluir la especificación de la distribución de ítems por requisitos cognitivos o por formato de ítem. Las especificaciones de extensión y del proyecto básico de la prueba suelen actualizarse en función de datos de ensayos sobre requisitos de tiempo, cobertura de contenido y precisión de puntajes. Cuando las pruebas se administran en forma adaptable, la extensión de la prueba (la cantidad de ítems administrados a cada individuo examinado) es determinada por reglas de espera, que pueden basarse en una cantidad fija de preguntas de la prueba o pueden basarse en un nivel deseado de precisión de puntajes.

Especificaciones psicométricas

Las precisiones psicométricas indican propiedades estadísticas deseadas de los ítems (p. ej., dificultad, discriminación y correlaciones entre ítems) así como las propiedades estadísticas deseadas de toda la prueba, incluyendo la naturaleza de la escala de presentación de reportes, dificultad y precisión de la prueba, y la distribución de ítems entre categorías de contenido y cognitivas. Cuando los índices psicométricos de los ítems se estiman utilizando teoría de respuesta al ítem (TRI), también se evalúa el ajuste del modelo a los datos. Esto se logra evaluando el grado en que se satisfacen las suposiciones subyacentes al modelo de respuesta al ítem (p. ej., unidimensionalidad e independencia local).

Especificaciones de calificación

Las especificaciones de la prueba describirán cómo deben calificarse los ítems individuales de la prueba y cómo deben combinarse los puntajes de los ítems para arrojar uno o más puntajes generales de la prueba. Todos los tipos de ítems requieren alguna indicación de cómo calificar las respuestas. Para ítems de respuesta seleccionada, una de las opciones de respuesta se considera la respuesta correcta en algunos programas de evaluación. En otros programas de evaluación, cada opción de respuesta puede arrojar un puntaje de ítem diferente. Para ítems de respuesta corta, una lista de respuestas aceptables puede ser suficiente, aunque a veces se requieren instrucciones de calificación más generales. Los ítems de respuesta extendida requieren reglas más detalladas para calificación, en ocasiones denominadas *rúbricas de puntajes*. Las rúbricas de puntajes especifican los criterios para evaluar el desempeño y pueden variar en el grado de juicio que conllevan, la cantidad de niveles de puntaje empleados y los modos en que se describen los criterios para cada nivel de puntaje. Es práctica común que los desarrolladores de la prueba proporcionen a los evaluadores ejemplos de desempeños en cada uno de los niveles de puntaje para ayudar a aclarar los criterios.

Para ítems de respuesta extendida, incluyendo tareas de desempeño, simulaciones y porfolios, se utilizan dos tipos principales de procedimientos de puntaje: analítico y holístico. Ambos procedimientos requieren criterios de desempeño explícitos que reflejen el marco de la prueba. Sin embargo, los enfoques conducen a algunas diferencias en las especificaciones de calificación. En el procedimiento de puntaje analítico, cada dimensión crítica de los criterios de desempeño se juzga de manera independiente, y se obtienen puntajes separados para cada una de estas dimensiones además de un puntaje general. En el procedimiento de puntaje holístico, implícitamente pueden considerarse los mismos criterios de desempeño, pero solo se proporciona un puntaje general. Debido a que el procedimiento analítico puede proporcionar información sobre una serie de dimensiones críticas, potencialmente proporciona información valiosa para fines de diagnóstico y se presta a evaluar fortalezas y debilidades de los examinandos. Sin embargo, se requerirá validación para interpretaciones diagnósticas para usos particulares de los puntajes separados. Por el contrario, el procedimiento holístico puede ser preferible cuando se desea un juicio general y cuando las habilidades evaluadas son complejas y están altamente interrelacionadas. Independientemente del tipo de procedimiento de calificación, diseñar los ítems y desarrollar rúbricas y procedimientos de puntajes es un proceso integrado.

Cuando los procedimientos de calificación requieren juicio humano, las especificaciones de calificación deben describir cualificaciones esenciales de los evaluadores, cómo deben capacitarse y supervisarse los evaluadores, como deben identificarse y resolverse las discrepancias de calificación, y cómo debe verificarse la ausencia de sesgo en el juicio del evaluador. En algunos casos, se utilizan algoritmos por computadora para calificar repuestas complejas de individuos examinados, como los ensayos. En esos casos, las especificaciones de calificación deben indicar cómo son generados los puntajes por estos algoritmos y cómo han de verificarse y validarse.

Las especificaciones de calificación también incluirán si los puntajes de la prueba son sumas simples de puntajes de ítems, involucran ponderación diferencial de ítems o secciones, o se basan en un modelo de medición más complejo. Si se utiliza un modelo de TRI, las especificaciones

deben indicar el formulario del modelo, cómo han de estimarse los parámetros del modelo y cómo ha de evaluarse el ajuste del modelo.

Especificaciones de la administración de la prueba

Las especificaciones de administración de la prueba describen cómo tiene que administrarse la prueba. Los procedimientos de administración incluyen el modo de ejecución de la prueba (p. ej., papel y lápiz o basada en computadora), límites de tiempo, procedimientos de adecuación, instrucciones y materiales provistos a los examinadores e individuos examinados y procedimientos para supervisar la ejecución de la prueba y garantizar la seguridad de la prueba. Para pruebas administradas por computadora, las especificaciones de administración también incluirán una descripción de cualquier requisito de hardware o software, incluyendo consideraciones de conectividad para pruebas basadas en Internet.

Perfeccionamiento de las especificaciones de la prueba

A menudo existe una sutil interacción entre el proceso de conceptualizar un constructo o dominio de contenido y el desarrollo de una prueba de ese constructo o dominio. Las especificaciones para la prueba proporcionan una descripción de cómo se representará el constructo o dominio y es posible que deban perfeccionarse a medida que avanza el desarrollo. Los procedimientos utilizados para desarrollar ítems y rúbricas de puntajes y para examinar las características de los ítems y la prueba a menudo pueden contribuir a aclarar las especificaciones. La medida en que el constructo se define completamente a priori depende de la aplicación de la evaluación. En muchas aplicaciones de evaluación, las especificaciones de la prueba bien definidas y detalladas orientan el desarrollo de ítems y sus rúbricas de puntajes y procedimientos asociados. En algunas áreas de medición psicológica, el desarrollo de la prueba puede ser menos dependiente de un marco definido a priori y puede depender más de un enfoque basado en datos que da por resultado una definición derivada en forma empírica del constructo sometido a medición. En esos casos, los ítems se seleccionan principalmente sobre la base de su relación empírica con un criterio externo, sus relaciones entre sí, o el grado en que discriminan entre grupos de individuos. Por ejemplo, ítems para una prueba para personal de ventas podrían seleccionarse sobre la base de las correlaciones de puntajes de ítems con medidas de productividad del personal de ventas actual. De manera similar, un inventario para ayudar a identificar diferentes patrones de psicopatología podría desarrollarse utilizando pacientes de diferentes subgrupos de diagnóstico. Cuando el desarrollo de la prueba se basa en un enfoque basado en datos, es probable que algunos ítems se seleccionen sobre la base de ocurrencias al azar en los datos. Los estudios de validación cruzada se realizan habitualmente para determinar la tendencia a seleccionar ítems al azar, lo cual involucra administrar la prueba a una muestra comparable que no estuvo involucrada en el esfuerzo de desarrollo de la prueba original.

En otras aplicaciones de evaluación, no obstante, las especificaciones de la prueba se fijan con antelación y orientan el desarrollo de ítems y procedimientos de calificación. Las relaciones empíricas pueden entonces utilizarse para informar decisiones sobre conservar, rechazar o modificar ítems. Las interpretaciones de puntajes de las pruebas desarrolladas mediante este proceso tienen la ventaja de un fundamento teórico y uno empírico para las dimensiones subyacentes representadas por la prueba.

Consideraciones para pruebas adaptables

En las pruebas adaptables, los ítems o conjuntos de ítems de la prueba se seleccionan a medida que se administra la prueba sobre la base de las respuestas del examinando a ítems anteriores. La especificación de los algoritmos de selección de ítems puede involucrar la consideración de cobertura de contenido como así también el aumento de la precisión de la estimación de puntajes. Cuando varios ítems están relacionados a un solo pasaje o tarea, se necesitan algoritmos más complejos para seleccionar el siguiente pasaje o tarea. En algunos casos, se desarrolla una cantidad mayor de ítems para cada pasaje o tarea y el algoritmo de

selección elige ítems específicos para administrar basados en consideraciones de contenido y precisión. Las especificaciones también deben indicar si se debe administrar una cantidad fija de ítems o si la prueba debe continuar hasta que se cumplan los criterios de precisión o cobertura de contenido.

El uso de pruebas adaptables y de modelos de pruebas basadas en computadora también involucra consideraciones especiales relacionadas con desarrollo de ítems. Cuando un conjunto de ítems operativos se desarrolla para una prueba adaptable computarizada, las especificaciones se refieren tanto al conjunto de ítems como a las reglas o procedimientos por los cuales se selecciona un conjunto de ítems individualizado para cada examinando. Algunas de las características atractivas de las pruebas adaptables computarizadas, como crear a medida el nivel de dificultad de los ítems de acuerdo con la capacidad del examinando, colocan restricciones adicionales sobre el diseño de dichas pruebas. En la mayoría de los casos, se necesitan grandes cantidades de ítems para construir una prueba adaptable computarizada para garantizar que el conjunto de ítems administrado a cada examinando cumpla todos los requisitos de las especificaciones de la prueba. Además, a menudo se desarrollan pruebas en el contexto de sistemas o programas de mayor tamaño. Se pueden crear múltiples conjuntos de ítems, por ejemplo, para usar con diferentes grupos de examinandos o en diferentes fechas de evaluación. Las preocupaciones sobre la seguridad de la prueba se intensifican cuando la disponibilidad limitada de equipos hace que sea imposible evaluar a todos los examinandos al mismo tiempo. Una serie de cuestiones, incluyendo la seguridad de la prueba, la complejidad de los requisitos de cobertura de contenido, niveles de precisión de puntajes requeridos, y si podría permitirse que los examinandos vuelvan a dar la prueba utilizando el mismo conjunto, deben considerarse al especificar el tamaño de los conjuntos de ítems asociados con cada formulario de la prueba adaptable.

El desarrollo de ítems para pruebas adaptables por lo general requiere que se desarrolle una mayor proporción de ítems a niveles altos o bajos de dificultad en relación con la población de la prueba de destino. Los datos de ensayos para ítems desarrollados para usar en pruebas adaptables deben examinarse para detectar posibles efectos de contexto para evaluar cuánto podrían cambiar los parámetros de los ítems cuando los ítems se administran en órdenes diferentes. Además, si los ítems se asocian con un pasaje o estímulo común, el desarrollo debe estar informado por una comprensión de cómo funcionará la selección de ítems. Por ejemplo, el enfoque para desarrollar ítems asociados con un pasaje puede diferir dependiendo de que el algoritmo de selección de ítems seleccione todos los ítems disponibles relacionados con el pasaje o pueda elegir subconjuntos de los ítems disponibles relacionados con el pasaje. Debido a los problemas que surgen cuando los ítems o tareas están anidados dentro de pasajes o estímulos en común, a menudo se consideran variaciones de las pruebas adaptables. Por ejemplo, la *evaluación de múltiples etapas* comienza con una serie de ítems de direccionamiento. Una vez que estos se dan y se califican, la computadora hace una ramificación a grupos de ítems que están explícitamente destinados a niveles de dificultad apropiados, basados en la evaluación del desempeño observado de los individuos examinados en los ítems de direccionamiento. En general, los requisitos especiales de las pruebas adaptables exigen algún cambio en el modo en que se desarrollan y prueban los ítems. Si bien los principios de calidad fundamentales del desarrollo de ítems no son diferentes, debe prestarse mayor atención a las interacciones entre contenido, formato y dificultad de los ítems para lograr conjuntos de ítems que sean más adecuados a este enfoque de evaluación.

Sistemas que respaldan el desarrollo de ítems y pruebas

La mayor dependencia de la tecnología y la necesidad de velocidad y eficiencia en el proceso de desarrollo de pruebas requieren la consideración de los sistemas que respaldan el desarrollo de ítems y pruebas. Dichos sistemas pueden mejorar la buena práctica de desarrollo de ítems y pruebas facilitando la creación y revisión de ítems/tareas, proporcionando un banco de ítems y herramientas automatizadas para asistir con el desarrollo de

formularios de prueba, e integrando información estadística de ítems/tareas con texto y gráficos de ítems/tareas. Estos sistemas pueden desarrollarse para cumplir con estándares y marcos de interoperabilidad y accesibilidad que faciliten a los usuarios de la prueba la transición de sus programas de evaluación de un desarrollador de la prueba a otro. Si bien los aspectos específicos de las bases de datos de ítems y sistemas de respaldo están fuera del alcance de los Estándares, el aumento de disponibilidad de esos sistemas obliga a los responsables de desarrollar esas pruebas a considerar la aplicación de tecnología al diseño y desarrollo de pruebas. Los desarrolladores de pruebas deben evaluar los costos y beneficios de diferentes aplicaciones, considerando cuestiones tales como velocidad de desarrollo, transportabilidad entre plataformas de evaluación, y seguridad.

Desarrollo y revisión de ítems

El desarrollador de la prueba por lo general reúne un conjunto de ítems que consiste en más preguntas o tareas que las necesarias para llenar el formulario o los formularios de la prueba que se elaborarán. Esto permite al desarrollador de la prueba seleccionar un conjunto de ítems para uno o más formularios de la prueba que cumplen las especificaciones de la prueba. La calidad de los ítems suele determinarse a través de procedimientos de revisión de ítems y ensayos de ítems, a menudo denominados *evaluación previa*. Los ítems se revisan en cuanto a calidad de contenido, claridad y aspectos de contenido irrelevantes del constructo que influyen en las respuestas de los examinandos. En la mayoría de los casos, la práctica acertada dicta que los ítems se revisen en cuanto a sensibilidad y potencial de resultar ofensivos que podrían introducir varianza irrelevante de constructo para individuos o grupos de examinandos. Por lo general se intenta evitar palabras y temas que puedan ofender o de otro modo molestar a examinandos, si material menos ofensivo es igualmente útil (véase el cap. 3). Para preguntas de respuestas construidas y tareas de desempeño, el desarrollo incluye rúbricas de puntajes específicas de ítems así como indicaciones o descripciones de tareas. Los revisores deben ser conocedores del contenido de la prueba y de los grupos de individuos examinados cubiertos por esta revisión.

A menudo, nuevos ítems de prueba se administran a un grupo de examinandos que son lo más representativos posible de la población de destino para la prueba, y cuando es posible, que representan adecuadamente a individuos de los subgrupos previstos. Los ensayos de ítems ayudan a determinar algunas de las propiedades psicométricas de los ítems de prueba, como dificultad de un ítem y capacidad para distinguir entre examinandos de diferente situación respecto del constructo que se evalúa. Los programas de evaluación continuos suelen hacer una prueba previa de los ítems insertándolos en pruebas operativas existentes (los ítems de ensayo no contribuyen a los puntajes que reciben los examinandos). Los análisis de las respuestas a estos ítems de ensayo proporcionan datos útiles para evaluar la calidad y pertinencia antes del uso operativo.

Los análisis estadísticos de los datos de los ensayos de ítems incluyen estudios de funcionamiento diferencial de los ítems (véase el cap. 3, "Imparcialidad en las pruebas"). Se dice que existe funcionamiento diferencial de los ítems cuando examinandos de diferentes grupos (p. ej., grupos definidos por género, raza/origen étnico o edad) que tienen capacidad aproximadamente igual respecto del constructo de destino o dominio de contenido difieren en sus respuestas a un ítem. En teoría, la meta máxima de dichos estudios es identificar aspectos irrelevantes del constructo del contenido del ítem, formato de ítems, o criterios de calificación que pueden afectar en forma diferencial los puntajes de la prueba de uno o más grupos de examinandos. Cuando se detecta funcionamiento diferencial de los ítems, los desarrolladores de la prueba intentan identificar explicaciones plausibles de las diferencias, y pueden luego reemplazar o revisar ítems para promover interpretaciones sólidas de puntajes para todos los individuos examinados. Cuando los ítems se abandonan debido a un índice de funcionamiento diferencial de los ítems, el desarrollador de la prueba debe tener cuidado de que ningún

reemplazo o revisión comprometa la cobertura del contenido de la prueba especificado.

Los desarrolladores de la prueba a veces utilizan enfoques que involucran entrevistas estructuradas o protocolos de pensamiento en voz alta con examinandos seleccionados. Dichos enfoques, en ocasiones denominados *laboratorios cognitivos*, se utilizan para identificar obstáculos irrelevantes a responder correctamente que podrían limitar la accesibilidad del contenido de la prueba. Los laboratorios cognitivos también se utilizan para proporcionar evidencia de que los procesos cognitivos que siguen quienes se someten a la evaluación son coherentes con el constructo sometido a medición.

Hay pasos adicionales en la evaluación de rúbricas de puntajes para ítems de respuesta extendida o tareas de desempeño. Los desarrolladores de la prueba deben identificar respuestas que ilustren cada nivel de calificación, para usar en la capacitación y verificación de evaluadores. Los desarrolladores también identifican respuestas en los límites entre niveles de puntajes adyacentes para utilizar en discusiones más detalladas durante la capacitación de evaluadores. Los análisis estadísticos de coherencia y exactitud de puntajes (concordancia con puntajes asignados por expertos) deben incluirse en el análisis de los datos de ensayos.

Reunión y evaluación de formularios de prueba

El próximo paso en el desarrollo de la prueba es reunir ítems en uno o más formularios de prueba o identificar uno o más conjuntos de ítems para una prueba adaptable o de múltiples etapas. El desarrollador de la prueba es responsable de documentar que los ítems seleccionados para la prueba cumplen los requisitos de las especificaciones de la prueba. En particular, el conjunto de ítems seleccionados para un nuevo formulario de prueba o conjunto de ítems para una prueba adaptable debe cumplir tanto las especificaciones de contenido como las psicométricas. Además, habitualmente se realizan revisiones editoriales y de contenido para reemplazar ítems que son demasiado similares a otros ítems o que pueden proporcionar pistas para las respuestas de otros ítems en el mismo formulario de prueba o conjunto de ítems. Cuando se preparan múltiples formularios de una prueba, las especificaciones de la prueba rigen cada uno de los formularios.

En ocasiones se prueban nuevos formularios de prueba o se hacen pruebas de campo de estos antes del uso operativo. La finalidad de una prueba de campo es determinar si los ítems funcionan según lo previsto en el contexto del nuevo formulario de prueba y evaluar las propiedades estadísticas, como la precisión o confiabilidad de puntajes, del nuevo formulario. Cuando se llevan a cabo pruebas de campo, todos los grupos de individuos examinados relevantes deben incluirse de modo que los resultados y conclusiones se generalicen al uso operativo previsto de los nuevos formularios de prueba y respalden más análisis de la imparcialidad de los nuevos formularios.

Desarrollo de procedimientos y materiales para administración y calificación

Muchas personas interesadas (p. ej., profesionales, profesores) pueden estar involucrados en el desarrollo de ítems y rúbricas de puntajes y/o en la evaluación de los desempeños posteriores. Si se utiliza un enfoque participativo, el conocimiento de los participantes sobre el dominio que se evalúa y su capacidad para aplicar las rúbricas de puntajes revisten una importancia fundamental. Igualmente importante para las personas involucradas en el desarrollo de pruebas y la evaluación de desempeños es su conocimiento de la naturaleza de la población que se evalúa. Las características relevantes de la población que se evalúa pueden incluir el rango típico de niveles de habilidad esperados, familiaridad con los modos de respuesta requeridos de ellos, maneras típicas en que se muestran el conocimiento y las habilidades, y el idioma principal utilizado.

El desarrollo de la prueba incluye la creación de una serie de documentos para respaldar la administración de la prueba según lo descripto en las especificaciones de la prueba. Las instrucciones a los usuarios de la prueba se desarrollan y prueban como parte de los procedimientos de pruebas piloto o de campo. Las instrucciones y la capacitación

para administradores de pruebas también deben desarrollarse y probarse. Una consideración clave en el desarrollo de procedimientos y materiales de administración de pruebas es que la administración de la prueba debe ser imparcial para todos los individuos examinados. Esto significa que las instrucciones para dar la prueba deben ser claras y que las condiciones de administración de la prueba deben ser estandarizadas para todos los individuos examinados. También significa que deben considerarse con antelación las adecuaciones de la prueba correspondientes para individuos examinados que las necesiten, como se analiza en el capítulo 3.

Para pruebas administradas por computadora, los procedimientos de administración deben ser coherentes con los requisitos de hardware y software incluidos en las especificaciones de la prueba. Los requisitos de hardware pueden cubrir velocidad y memoria del procesador; teclado, mouse u otros dispositivos de entrada de datos; tamaño del monitor y resolución de pantalla; y conectividad a servidores locales o a Internet. Los requisitos de software cubren sistemas operativos, navegadores u otras herramientas comunes y disposiciones para bloquear acceso a otro software o interferencia de este. Los individuos examinados que dan pruebas administradas por computadora deben ser informados acerca de cómo responder a las preguntas, cómo desplazarse por la prueba, si pueden saltar ítems, si pueden volver a ver ítems respondidos previamente más adelante en el período de evaluación, si pueden suspender la sesión de evaluación para un tiempo más tarde, y otras exigencias que pueden ocurrir durante la evaluación.

También deben implementarse procedimientos de seguridad de la prueba junto con la administración y la calificación de las pruebas. Dichos procedimientos a menudo incluyen el seguimiento y almacenamiento de materiales; cifrado de transmisión electrónica del contenido y los puntajes del examen; acuerdos de confidencialidad para examinandos, evaluadores y administradores; y procedimientos para supervisar a los individuos examinados durante la sesión de evaluación. Además, para programas de evaluación que reutilizan ítems de prueba o formularios de prueba, los procedimientos de seguridad deben incluir evaluación de cambios en las estadísticas de ítems para evaluar la posibilidad de una violación de seguridad. Los desarrolladores o usuarios de la prueba podrían considerar la supervisión de sitios web respecto de la posible divulgación del contenido de la prueba.

Revisiones de las pruebas

Las pruebas y sus documentos de respaldo (p. ej., manuales de la prueba, manuales técnicos, guías de usuario) deben revisarse periódicamente para determinar si se necesitan revisiones. Las revisiones o modificaciones son necesarias cuando nuevos datos de investigación, cambios significativos en el dominio o nuevas condiciones del uso y la interpretación de la prueba sugieren que la prueba ha dejado de ser óptima o completamente apropiada para algunos de sus usos previstos. Por ejemplo, las pruebas se revisan si el contenido o el lenguaje de la prueba se ha desactualizado y, por lo tanto, puede afectar posteriormente la validez de las interpretaciones de los puntajes de la prueba. Sin embargo, las normas desactualizadas pueden no tener las mismas implicaciones para las revisiones que una prueba desactualizada. Por ejemplo, es posible que sea necesario actualizar las normas para una prueba de rendimiento después de un período de aumento o descenso del rendimiento en la población de normalización, o cuando hay cambios en la población que se somete a la prueba, pero el contenido de la prueba propiamente dicho puede continuar siendo tan relevante como lo era cuando se desarrolló la prueba. El momento en que se necesite la revisión variará como una función del contenido y los usos previstos de la prueba. Por ejemplo, las pruebas de dominio de planes de estudios educativos o de capacitación deben revisarse cada vez que se actualice el plan de estudios correspondiente. Las pruebas que evalúan constructos psicológicos deben revisarse cuando la investigación sugiere una conceptualización revisada del constructo.

ESTÁNDARES PARA EL DISEÑO Y DESARROLLO DE PRUEBAS

Los estándares en este capítulo comienzan con un estándar global (numerado 4.0), que se ha diseñado para transmitir la intención central o enfoque principal del capítulo. El estándar global también puede verse como el principio rector del capítulo, y es aplicable a todas las pruebas y usuarios de pruebas. Todos los estándares posteriores se han separado en cuatro unidades temáticas denominadas de la siguiente manera:

1. Estándares para especificaciones de la prueba
2. Estándares para desarrollo y revisión de ítems
3. Estándares para desarrollar procedimientos y materiales de administración y calificación de pruebas
4. Estándares para revisión de pruebas

Estándar 4.0

Las pruebas y programas de evaluación deben diseñarse y desarrollarse de una manera que respalde la validez de las interpretaciones de los puntajes de la prueba para sus usos previstos. Los desarrolladores y editores de pruebas deben documentar las medidas tomadas durante el proceso y desarrollo de la prueba para proporcionar evidencia de imparcialidad, confiabilidad y validez para los usos previstos para individuos en la población prevista de individuos examinados.

Comentario: A continuación, se describen estándares específicos para diseñar y desarrollar pruebas de una manera que respalde los usos previstos. Las especificaciones iniciales para una prueba, que tienen por objeto guiar el proceso de desarrollo, pueden modificarse o ampliarse a medida que avanza el desarrollo y se dispone de nueva información. Tanto la documentación inicial como la final de las especificaciones y los procedimientos de desarrollo de la prueba proporcionan una base sobre la cual los expertos externos y los usuarios de la prueba pueden juzgar el grado en que se han respaldado o es probable que se respalden los usos previstos, lo cual conduce a interpretaciones válidas de los resultados de la prueba para todos los individuos. Las especificaciones de la prueba iniciales pueden modificase a medida que se reúne evidencia durante el desarrollo y la implementación de la prueba.

Unidad 1. Estándares para especificaciones de la prueba

Estándar 4.1

Las especificaciones de la prueba deben describir la(s) finalidad(es) de la prueba, la definición del constructo o el dominio medido, la población prevista de individuos examinados y las interpretaciones para los usos previstos. Las especificaciones deben incluir una justificación que respalde las interpretaciones y usos de los resultados de la prueba para el o los fines previstos.

Comentario: La adecuación y utilidad de las interpretaciones de la prueba dependen del rigor con el que se hayan definido y explicado la(s) finalidad(es) de la prueba y el dominio representado por la prueba. La definición del dominio debería ser lo suficientemente detallada y delimitada para mostrar con claridad qué dimensiones de conocimiento, habilidades, procesos cognitivos, actitudes, valores, emociones o comportamientos se incluyen y qué dimensiones se excluyen. Una descripción clara mejorará los juicios exactos de revisores y otras personas acerca del grado de congruencia entre el dominio definido y los ítems de la prueba. La especificación clara de la población prevista de individuos examinados y sus características puede ayudar a proteger contra características irrelevantes del constructo del contenido y el formato de los ítems. Las especificaciones deben incluir planes para recopilar evidencia de la validez de las interpretaciones previstas de los puntajes de la prueba para sus usos previstos. Los desarrolladores de la prueba también deben identificar posibles limitaciones sobre el uso de la prueba o posibles usos inapropiados.

Estándar 4.2

Además de describir los usos previstos de la prueba, las especificaciones de la prueba deben definir el contenido de la prueba, la extensión propuesta de la prueba, los formatos de los ítems, las propiedades psicométricas deseadas de los ítems de la prueba y la prueba, y el orden de los ítems y secciones. Las especificaciones de la prueba también deben establecer la cantidad de tiempo permitido para la evaluación; instrucciones para los examinandos; procedimientos que se usarán para la administración de la prueba, incluyendo variaciones aceptables; cualquier material que se usará; y procedimientos de calificación y presentación de reportes. Las especificaciones para pruebas basadas en computadora deben incluir una descripción de cualquier requisito de hardware y software.

Comentario: El juicio profesional desempeña un rol importante en el desarrollo de las especificaciones de la prueba. Los procedimientos específicos utilizados para desarrollar las especificaciones dependen de la(s) finalidad(es) de la prueba. Por ejemplo, al desarrollar pruebas para obtención de licencias y certificación, los análisis prácticos o análisis de empleo por lo general proporcionan la base para definir las especificaciones de la prueba; los análisis de empleo solos por lo general cumplen esta función para las pruebas de empleo. Para las pruebas de rendimiento que se toman al final de un curso, las especificaciones de la prueba deben basarse en un esquema del contenido y las metas del curso. Para las pruebas de colocación, los desarrolladores examinarán los conocimientos y las habilidades de nivel de ingreso requeridos para diferentes cursos. En el desarrollo de pruebas psicológicas, las descripciones y los criterios de diagnóstico de déficits del comportamiento, mentales y emocionales y psicopatología informan las especificaciones de la prueba.

Los tipos de ítems, los formatos de respuesta, los procedimientos de calificación, y los procedimientos de administración de la prueba deben seleccionarse sobre la base de la(s) finalidad(es) de la prueba, el dominio sometido a medición, y los examinandos previstos. En la medida posible, el contenido y los procedimientos de administración de la prueba deben elegirse para que las inferencias previstas de los puntajes de la prueba sean igualmente válidas para todos los examinandos. Algunos detalles de las especificaciones de la prueba pueden ser revisados sobre la base de pruebas piloto o de campo iniciales. Por ejemplo, las especificaciones de la extensión de la prueba o combinación de tipos de ítems podrían modificarse en función de datos iniciales para lograr la precisión de medida deseada.

Estándar 4.3

Los desarrolladores de la prueba deben documentar la justificación y la evidencia de respaldo para la administración, calificación y reglas de presentación de reportes utilizadas en pruebas adaptables por computadora, adaptables de múltiples etapas u otras ejecutadas utilizando algoritmos de computación para seleccionar ítems. Esta documentación debe incluir procedimientos utilizados en la selección de ítems o conjuntos de ítems para administración, en la determinación de las condiciones de punto de partida y finalización para la prueba, en la calificación de la prueba y en el control de la exposición de ítems.

Comentario: Si una prueba adaptable computarizada tiene por objeto medir una cantidad de subcategorías de contenido diferentes, los procedimientos de selección de ítems deben asegurar que las subcategorías estén representadas adecuadamente por los ítems presentados al examinando. Las justificaciones comunes para las pruebas adaptables computarizadas son que aumenta la precisión de los puntajes, en particular para individuos examinados de alta y baja calificación, o que se logra precisión comparable mientras se reduce el tiempo de evaluación. Obsérvese que estas pruebas están sujetas a los mismos requisitos para la documentación de la validez de las interpretaciones de los puntajes para su uso previstos que otros tipos de pruebas. Las especificaciones de la prueba deben incluir planes para recopilar evidencia requerida para dicha documentación.

Estándar 4.4

Si los desarrolladores de la prueba preparan diferentes versiones de una prueba con algún cambio en las especificaciones de la prueba, deben documentar el contenido y las especificaciones psicométricas de cada versión. La documentación debe describir el impacto de las diferencias entre versiones sobre la validez de las interpretaciones de los puntajes para los usos previstos y sobre la precisión y comparabilidad de puntajes.

Comentario: Los desarrolladores de la prueba pueden tener diversos motivos para crear diferentes versiones de una prueba, como permitir diferentes cantidades de tiempo para la administración de la prueba reduciendo o aumentando la cantidad de ítems en la prueba original, o permitir la administración a diferentes poblaciones traduciendo las preguntas de la prueba a diferentes idiomas. Los desarrolladores de la prueba deben documentar el grado en que las especificaciones difieren de las de la prueba original, proporcionar una justificación para las diferentes versiones y describir las implicaciones de dichas diferencias para interpretar los puntajes derivados de las diferentes versiones. Los desarrolladores y usuarios de la prueba deben supervisar y documentar cualquier diferencia psicométrica entre versiones de la prueba sobre la base de evidencia recopilada durante el desarrollo y la implementación. La evidencia de diferencias puede involucrar juicios cuando la cantidad de individuos examinados que reciben una versión en particular es pequeña (p. ej., una versión en braille). Obsérvese que estos requisitos son además de los requisitos normales para demostrar la equivalencia de puntajes de diferentes formularios de la misma prueba. Cuando se utilizan diferentes idiomas en diferentes versiones de la prueba, los procedimientos utilizados para desarrollar y verificar las traducciones a cada idioma deben documentarse.

Estándar 4.5

Si el desarrollador de la prueba indica que se permite que varíen las condiciones de administración de un examinando o grupo a otro, debe identificarse la variación aceptable en las condiciones para la administración. Deben documentarse una justificación para permitir las diferentes condiciones y cualquier requisito para permitir las diferentes condiciones.

Comentario: La variación en las condiciones de administración puede reflejar restricciones de administración en diferentes ubicaciones o, más comúnmente, puede estar diseñada como adecuaciones de la evaluación para individuos examinados o grupos de individuos examinados específicos. Un ejemplo de una variación común es el uso de administración por computadora de un formulario de prueba en algunas ubicaciones y administración con papel y lápiz del mismo formulario en otras ubicaciones. Otro ejemplo es la administración en grupos pequeños o individual para examinandos cuyo desempeño en la prueba podría estar limitado por distracciones en contextos de grupos grandes. Las adecuaciones de la prueba, como se analiza en el capítulo 3 (“Imparcialidad en las pruebas”), son cambios hechos en una prueba para aumentar la imparcialidad para individuos que de otro modo serían desfavorecidos por características irrelevantes del constructo de los ítems de la prueba. Los desarrolladores de la prueba deben especificar procedimientos para supervisar variaciones y para reunir evidencia para mostrar que el constructo de destino está o no está alterado por variaciones aceptables. Estos procedimientos deben documentarse sobre la base de datos recopilados durante la implementación.

Estándar 4.6

Cuando corresponda para documentar la validez de las interpretaciones de los puntajes de la prueba para los usos previstos, expertos relevantes externos al programa de evaluación deben revisar las especificaciones de la prueba para evaluar su adecuación para los fines previstos de los puntajes de la prueba e imparcialidad para los examinandos previstos. La finalidad de la revisión, el proceso por el cual se realiza la revisión y los resultados de la revisión deben

documentarse. Las cualificaciones, experiencias relevantes y características demográficas de los jueces expertos también deben documentarse.

Comentario: Pueden considerarse varios factores al decidir si es necesaria la revisión externa de especificaciones de la prueba, incluyendo el alcance del uso previsto, si las interpretaciones de los puntajes pueden tener consecuencias importantes, y la disponibilidad de expertos externos. La revisión de expertos de las especificaciones de la prueba puede servir a muchos fines útiles, como ayudar a garantizar la calidad y representatividad del contenido. El uso de expertos externos al proceso de desarrollo de la prueba respalda la objetividad en los juicios de la calidad de las especificaciones de la prueba. La revisión de las especificaciones antes de comenzar el desarrollo de los ítems puede evitar problemas significativos durante las revisiones posteriores de ítems de la prueba. Los jueces expertos pueden incluir individuos que representen poblaciones definidas de interés para las especificaciones de la prueba. Por ejemplo, si la prueba debe administrarse a diferentes grupos lingüísticos y culturales, la revisión de expertos habitualmente incluye a miembros de estos grupos y expertos en la evaluación de cuestiones específicas a estos grupos.

Unidad 2. Estándares para el desarrollo y la revisión de ítems

Estándar 4.7

Los procedimientos utilizados para desarrollar, revisar y probar ítems y para seleccionar ítems del conjunto de ítems deben documentarse.

Comentario: Las cualificaciones de individuos que desarrollan y revisan ítems y los procesos utilizados para capacitarlos y guiarlos en estas actividades son aspectos importantes de la documentación del desarrollo de la prueba. Por lo general, varios grupos de individuos participan en el proceso de desarrollo de la prueba, incluyendo redactores de ítems y personas que participan en revisiones de los ítems y del contenido de la prueba en cuanto a sensibilidad o para otros fines.

Estándar 4.8

El proceso de revisión de la prueba debe incluir análisis empíricos y/o el uso de jueces expertos para revisar ítems y criterios de calificación. Cuando se utilizan jueces expertos, sus cualificaciones, experiencias relevantes y características demográficas deben documentarse, junto con las instrucciones y la capacitación en el proceso de revisión de ítems que los jueces reciben.

Comentario: Cuando el tamaño de la muestra lo permita, se necesitan análisis empíricos para verificar las propiedades psicométricas de los ítems de la prueba y también para verificar si los ítems de la prueba funcionan en forma similar para grupos diferentes. Se puede pedir que jueces expertos verifiquen la calificación de ítems e identifiquen material que probablemente sea inapropiado, confuso u ofensivo para grupos en la población de examinandos. Por ejemplo, se puede pedir a los jueces que identifiquen si la falta de exposición a contextos de problemas en problemas de planteo de matemáticas puede constituir una preocupación para algunos grupos de estudiantes. Varios grupos de examinandos pueden ser definidos por características tales como edad, origen étnico, cultura, género, discapacidad o región demográfica. Cuando sea viable, la evidencia tanto empírica como basada en juicios de la medida en que los ítems de la prueba funcionan de manera similar para diferentes grupos debe utilizarse en el cribado de los ítems. (Véase el cap. 3 para consultar ejemplos de tipos apropiados de evidencia).

En ocasiones se realizan estudios de la alineación de los formularios de la prueba con las especificaciones de contenido para respaldar interpretaciones de que los puntajes de la prueba indican dominio del contenido de la prueba de destino. Expertos independientes de los desarrolladores de la prueba juzgan el grado en que el contenido de los ítems se corresponde con las categorías de contenido en las especificaciones de la prueba y si los formularios de prueba proporcionan cobertura equilibrada del contenido de destino.

Estándar 4.9

Cuando se realizan ensayos de ítems o formularios de prueba, deben documentarse los procedimientos utilizados para seleccionar la(s) muestra(s) de examinandos, así como las características resultantes de la(s) muestra(s). Las muestras deben ser tan representativas como sea posible de las poblaciones para las que está prevista la prueba.

Comentario: Deben documentarse cuando corresponda las condiciones que pueden afectar diferencialmente el desempeño en los ítems de la prueba según las muestras de los ensayos en comparación con las poblaciones previstas. Por ejemplo, los examinandos pueden estar menos motivados cuando saben que sus puntajes no tendrán un impacto en ellos. Cuando sea posible, deben examinarse y documentarse las características de los ítems y las pruebas para subgrupos relevantes en la población prevista de individuos examinados.

En la medida en que sea viable, los ensayos de ítems y formularios de prueba deben incluir grupos de individuos examinados relevantes. Cuando el tamaño de la muestra lo permita, los desarrolladores de la prueba deben determinar si los puntajes de los ítems tienen diferentes relaciones con el constructo sometido a medición para diferentes grupos (funcionamiento diferencial de los ítems). Cuando se diseñan adecuaciones de la prueba para grupos de individuos examinados específicos, también debe recopilarse información sobre el desempeño en el ítem en condiciones adaptadas. Para grupos relativamente pequeños, la información cualitativa puede ser útil. Por ejemplo, las entrevistas con examinandos podrían utilizarse para evaluar la efectividad de las adecuaciones en la eliminación de la varianza irrelevante.

Estándar 4.10

Cuando un desarrollador de pruebas evalúa las propiedades psicométricas de los ítems, el modelo utilizado para ese fin (p. ej., teoría clásica de los tests, teoría de respuesta al ítem u otro modelo) debe documentarse. La muestra utilizada para estimar las propiedades de los ítems debe describirse y debe ser de un tamaño y diversidad adecuados para el procedimiento. El proceso por el cual se criban los ítems y los datos utilizados para cribado, como dificultad del ítem, discriminación de ítems, o funcionamiento diferencial de los ítems (DIF) para grupos importantes de individuos examinados también deben documentarse. Cuando se utilizan métodos basados en modelos (p. ej., TRI) para estimar los parámetros de los ítems en el desarrollo de pruebas, el modelo de respuesta al ítem, los procedimientos de estimación y la evidencia de ajuste del modelo deben documentarse.

Comentario: Si bien el tamaño general de la muestra es relevante, también debe haber una cantidad adecuada de casos en regiones críticas para la determinación de las propiedades psicométricas de los ítems. Si la prueba debe lograr la mayor precisión en una parte en particular de la escala de puntajes y esta consideración afecta la selección de ítems, la manera en que las estadísticas de ítems se utilizan para la selección de ítems debe documentarse cuidadosamente. Cuando se utiliza TRI como base para el desarrollo de la prueba, es importante documentar la adecuación del ajuste del modelo a los datos. Esto se logra proporcionando información sobre la medida en que se satisfacen las suposiciones de TRI (p. ej., unidimensionalidad, independencia del ítem local o, para ciertos modelos, igualdad de parámetros de pendiente).

Se deben describir las estadísticas utilizadas para indicar ítems que funcionan de manera diferente para diferentes grupos, incluyendo especificación de los grupos a analizar, los criterios para la indicación, y los procedimientos para revisar y tomar decisiones definitivas sobre los ítems indicados. Los tamaños de la muestra para grupos de interés deben ser adecuados para detectar DIF significativo.

Los desarrolladores de la prueba deben considerar cómo cualquier diferencia entre las condiciones de administración de la prueba de campo y el formulario final podría afectar el desempeño del ítem. Las condiciones que pueden afectar las estadísticas

de los ítems incluyen motivación de los examinandos, posición de los ítems, límites de tiempo, extensión de la prueba, modo de evaluación (p. ej., papel y lápiz frente a administración por computadora) y uso de calculadoras u otras herramientas.

Estándar 4.11

Los desarrolladores de la prueba deben realizar estudios de validación cruzada cuando los ítems o pruebas se seleccionan principalmente sobre la base de relaciones empíricas más que sobre la base de consideraciones de contenido o teóricas. Debe documentarse el grado a en que los diferentes estudios muestran resultados coherentes.

Comentario: Cuando se utilizan enfoques basados en datos para el desarrollo de la prueba, los ítems se seleccionan principalmente sobre la base de sus relaciones empíricas con un criterio externo, sus relaciones entre sí, o su poder para discriminar entre grupos de individuos. En estas circunstancias, es probable que algunos ítems se seleccionen sobre la base de ocurrencias al azar en los datos usados. Administrar la prueba a una muestra comparable de examinandos o el uso de una muestra de validación separada proporciona verificación independiente de las relaciones utilizadas en la selección de ítems.

Las técnicas de optimización estadística como la regresión escalonada se utilizan a veces para desarrollar compuestos de pruebas o para seleccionar pruebas para otro uso en una batería de pruebas. Al igual que con la selección empírica de ítems, puede ocurrir capitalización del azar. La validación cruzada de una muestra independiente o el uso de una fórmula que prediga la reducción de correlaciones en una muestra independiente pueden proporcionar un índice menos sesgado del poder predictivo de las pruebas o compuesto.

Estándar 4.12

Los desarrolladores de la prueba deben documentar el grado en que el dominio de contenido de una prueba representa el dominio definido en las especificaciones de la prueba.

Comentario: Los desarrolladores de la prueba deben proporcionar evidencia del grado en que los ítems de la prueba y los criterios de calificación arrojan resultados que representan el dominio definido. Esto ofrece una base para ayudar a determinar si el desempeño en la prueba puede generalizarse al dominio que se evalúa. Esto es especialmente importante para las pruebas que contienen una pequeña cantidad de ítems, como las evaluaciones de desempeño. Dicha evidencia puede ser proporcionada por jueces expertos. En algunas situaciones, se lleva a cabo un estudio independiente de la alineación de las preguntas de la prueba a las especificaciones de contenido para validar el procesamiento interno del desarrollador para garantizar la cobertura de contenido apropiada.

Estándar 4.13

Cuando evidencia creíble indica que la varianza irrelevante podría afectar los puntajes de la prueba, en la medida en que sea viable, el desarrollador de la prueba debe investigar las fuentes de varianza irrelevante. Cuando sea posible, dichas fuentes de varianza irrelevante deben ser eliminadas o reducidas por el desarrollador de la prueba.

Comentario: Se pueden utilizar diversos métodos para verificar la influencia de factores irrelevantes, incluyendo análisis de correlaciones con medidas de otros constructos relevantes e irrelevantes y, en algunos casos, análisis cognitivos más profundos (p. ej., uso de sondeos de seguimiento para identificar motivos relevantes e irrelevantes de respuestas correctas e incorrectas) de la situación del individuo examinado respecto del constructo de destino. Un entendimiento más profundo de las fuentes de varianza irrelevantes también puede conducir al perfeccionamiento de la descripción del constructo sometido a examen.

Estándar 4.14

Para una prueba que tiene un límite de tiempo, la investigación del desarrollo de la prueba debe examinar el grado en que los puntajes incluyen

un componente de velocidad y debe evaluar la adecuación de ese componente, dado el dominio que la prueba está diseñada para medir.

Comentario: Como mínimo, los desarrolladores de la prueba deben examinar la proporción de individuos examinados que completan toda la prueba, así como la proporción que no responde (omite) preguntas individuales de la prueba. Cuando la velocidad es una parte significativa del constructo de destino, la distribución de la cantidad de ítems respondidos debe analizarse para verificar la variabilidad apropiada en la cantidad de ítems en los que se hizo un intento así como la cantidad de respuestas correctas. Cuando la velocidad no es una parte significativa del constructo de destino, deben determinarse los límites de tiempo de modo que los individuos examinados tengan tiempo adecuado para demostrar el conocimiento y la habilidad de destino.

Unidad 3. Estándares para desarrollar procedimientos y materiales de administración y calificación de pruebas

Estándar 4.15

Las instrucciones para la administración de la prueba deben presentarse con suficiente claridad para que sea posible que otros repliquen las condiciones de administración en las que se obtuvieron los datos sobre confiabilidad, validez y (cuando corresponda) normas. Las variaciones admisibles en los procedimientos de administración deben describirse claramente. El proceso para revisar solicitudes de variaciones adicionales en la evaluación también debe documentarse.

Comentario: Debido a que todas las personas que administran pruebas, incluyendo aquellas en escuelas, la industria y clínicas, necesitan seguir procedimientos de administración de pruebas con atención, es esencial que los administradores de la prueba reciban instrucciones detalladas sobre directrices y procedimientos de administración de la prueba. Es posible que se necesiten adecuaciones de la prueba para permitir la medición exacta de constructos previstos para grupos específicos de examinandos, como individuos con discapacidades e individuos cuya lengua nativa no sea el inglés. (Véase el cap. 3, "Imparcialidad en las pruebas").

Estándar 4.16

Las instrucciones presentadas a los examinandos deben contener suficiente detalle para que los examinandos puedan responder a una tarea de la manera prevista por el desarrollador de la prueba. Cuando corresponda, deben proporcionarse los materiales de la muestra, preguntas prácticas o de la muestra, criterios para calificación y un ítem representativo identificado con cada formato de ítem o área importante en la clasificación o dominio de la prueba a los examinandos antes de la administración de la prueba, o deben incluirse en el material de evaluación como parte de las instrucciones de administración estándar.

Comentario: Por ejemplo, en un inventario de personalidad la intención puede ser que los examinandos den la primera respuesta que se les ocurra. Una expectativa de este tipo debe dejarse en claro en las instrucciones del inventario. En otro ejemplo, en las instrucciones para inventarios de intereses u ocupacionales, es posible que sea importante especificar si los examinandos deben marcar las actividades que preferirían en condiciones ideales o si deben considerar tanto su oportunidad como su capacidad en forma realista.

Las instrucciones y cualquier material práctico deben estar disponibles en formatos a los que todos los examinandos puedan acceder. Por ejemplo, si se proporciona una versión en braille de la prueba, las instrucciones y cualquier material práctico deben también proporcionarse en una forma a la que puedan acceder los estudiantes que realizan la versión en braille.

El alcance y la naturaleza de los materiales prácticos y las instrucciones dependen de los niveles esperados de conocimiento entre los examinandos. Por ejemplo, al usar un formato de prueba novedoso, es posible que sea importante

proporcionar al examinando una oportunidad práctica como parte de la administración de la prueba. En algunas situaciones de prueba, es posible que sea importante que las instrucciones aborden asuntos tales como límites de tiempo y los efectos que hacer conjeturas tiene en los puntajes de la prueba. Si se permite la ampliación o elaboración de las instrucciones de la prueba, las condiciones en las que esto puede hacerse deben indicarse claramente en el formulario de reglas generales y dando ejemplos representativos. Si no se permitirá ampliación o elaboración, esto debe indicarse explícitamente. Los desarrolladores de la prueba deben incluir orientación para tratar con preguntas típicas de los examinandos. Se debe instruir a los administradores de la prueba sobre cómo ocuparse de las preguntas que pueden surgir durante el período de evaluación.

Estándar 4.17

Si una prueba o parte de una prueba está prevista para uso de investigación únicamente y no se distribuye para uso operativo, deben mostrarse de manera prominente declaraciones a ese efecto en todos los materiales de administración e interpretación de la prueba relevantes que se proporcionan al usuario de la prueba.

Comentario: Este estándar se refiere a pruebas que están previstas para uso en investigación únicamente. No se refiere a las funciones de desarrollo estándar de pruebas que ocurren antes del uso operativo de una prueba (p. ej., ensayos de ítems o formularios). Es posible que existan requisitos legales para informar a los participantes sobre cómo el desarrollador de la prueba utilizará los datos generados de la prueba, incluyendo información personal de identificación, cómo se protegerá esa información y con quién podría compartirse.

Estándar 4.18

Los procedimientos para calificación y, si corresponde, los criterios de calificación, deben ser presentados por el desarrollador de la prueba con suficiente detalle y claridad para maximizar la exactitud de la calificación. Las instrucciones para usar escalas de calificación o para derivar puntajes obtenidos por codificación, escalamiento o clasificando respuestas construidas deben ser claras. Esto es especialmente crítico para ítems de respuesta extendida como las tareas de desempeño, porfolios y ensayos.

Comentario: Al calificar respuestas más complejas, los desarrolladores de la prueba deben proporcionar rúbricas detalladas y capacitación en su uso. Proporcionar múltiples ejemplos de respuestas en cada nivel de puntajes para usarse en la capacitación de evaluadores y supervisar la coherencia de calificaciones es también práctica común, aunque estos suelen agregarse a las especificaciones de calificación durante el desarrollo y los ensayos de ítems. Para supervisar la efectividad de la calificación, deben especificarse criterios de coherencia para la cualificación de evaluadores, según corresponda, junto con procedimientos, tales como calificación doble de algunas o todas las respuestas. Según corresponda, los desarrolladores de la prueba deben especificar los criterios de selección para evaluadores y procedimientos para capacitación, cualificación y supervisión de evaluadores. Si se utilizan diferentes grupos de evaluadores con diferentes administraciones, deben especificarse e implementarse procedimientos para verificar la comparabilidad de puntajes generados por los diferentes grupos.

Estándar 4.19

Cuando deban usarse algoritmos automatizados para calificar respuestas complejas de los individuos examinados, deben documentarse las características de las respuestas en cada nivel de puntaje junto con los fundamentos teóricos y empíricos para el uso de los algoritmos.

Comentario: Los algoritmos de puntaje automático deben estar respaldados por una articulación de los fundamentos teóricos y metodológicos para su uso que sean suficientemente detallados para establecer una justificación para vincular los puntajes de la prueba resultantes con el constructo de interés subyacente. Además, el algoritmo de

puntaje automático debe tener un respaldo de investigación empírica, como tasas de concordancia con evaluadores humanos, antes del uso operativo, así como evidencia de que los algoritmos de puntaje no introducen sesgo sistemático contra algunos subgrupos.

Debido a lo que los algoritmos de puntaje automático a menudo se consideran patentados, sus desarrolladores rara vez están dispuestos a revelar las reglas de calificación y ponderación en documentación pública. Además, en algunos casos, la divulgación completa de detalles del algoritmo de puntaje podría dar por resultado estrategias de orientación que aumentarían los puntajes sin ningún cambio real en el o los constructos que se evalúan. En esos casos, los desarrolladores deben describir las características generales de los algoritmos de puntaje. También pueden hacer que los algoritmos sean revisados por expertos independientes, en condiciones de confidencialidad, y reunir juicios independientes de la medida en que los puntajes resultantes implementarán exactamente las rúbricas de puntajes previstas y estarán libres de sesgo para las subpoblaciones previstas de individuos examinados.

Estándar 4.20

El proceso para seleccionar, capacitar, cualificar y supervisar a evaluadores debe ser especificado por el desarrollador de la prueba. Los materiales de capacitación, como las rúbricas de puntajes y ejemplos de respuestas de examinandos que ejemplifican los niveles en la escala de puntajes de rúbrica, y los procedimientos para la capacitación de evaluadores deben dar por resultado un grado de exactitud y concordancia entre evaluadores que permita que los puntajes se interpreten según lo previsto originalmente por el desarrollador de la prueba. Las especificaciones también deben describir procesos para evaluar la coherencia de evaluadores y la posible desviación con el tiempo en la calificación de los evaluadores.

Comentario: En la medida posible, los procesos y materiales de calificación deben prever problemas que podrían surgir durante la calificación. Los materiales de capacitación deben abordar cualquier idea equivocada común sobre las rúbricas utilizadas para describir niveles de puntajes. Cuando se califica texto escrito, es común incluir un conjunto de respuestas precalificadas para usar en la capacitación y para juzgar la exactitud de la calificación. La base para determinar la coherencia de calificación (p. ej., porcentaje de concordancia exacta, porcentaje dentro de un punto de puntaje, o algún otro índice de concordancia) debe indicarse. La información sobre la coherencia de calificación es fundamental para estimar la precisión de los puntajes resultantes.

Estándar 4.21

Cuando los usuarios de la prueba son responsables de calificar y la calificación requiere el juicio del evaluador, el usuario de la prueba es responsable de proporcionar capacitación e instrucción adecuadas a los evaluadores y de examinar la concordancia y exactitud de los evaluadores. El desarrollador de la prueba debe documentar el nivel esperado de concordancia y exactitud del evaluador y debe proporcionar tanta orientación técnica como sea posible para ayudar a los usuarios de la prueba a cumplir con este estándar.

Comentario: Una práctica común de los desarrolladores de pruebas es proporcionar materiales de capacitación (p. ej., rúbricas de puntajes, ejemplos de respuestas de examinandos en cada nivel de puntaje) y procedimientos cuando la calificación es realizada por usuarios de la prueba y requiere el juicio del evaluador. La capacitación proporcionada para respaldar la calificación local debe incluir estándares para verificar la exactitud de los evaluadores durante la capacitación y la calificación operativa. La capacitación también debe cubrir cualquier consideración especial para grupos de examinandos que podrían interactuar de manera diferente con la tarea que se calificará.

Estándar 4.22

Los desarrolladores de la prueba deben especificar los procedimientos utilizados para interpretar

puntajes de la prueba y, cuando corresponda, las muestras normativas o de estandarización o el criterio utilizado.

Comentario: Las especificaciones de la prueba pueden indicar que los puntajes previstos deben interpretarse como indicación de un nivel absoluto del constructo sometido a medición o como indicación de la situación respecto del constructo en relación con otros individuos examinados, o ambas. En las interpretaciones de puntaje absolutas, se supone que el puntaje o promedio refleja directamente un nivel de competencia o destreza en algún dominio de criterios definido. En las interpretaciones de puntaje relativas el estado de un individuo (o grupo) se determina comparando el puntaje (o puntaje medio) con el desempeño de otros en una o más poblaciones definidas. Las pruebas diseñadas para facilitar un tipo de interpretación pueden funcionar de manera menos efectiva para el otro tipo de interpretación. Dado el diseño de la prueba adecuado y los datos de respaldo adecuados, no obstante, los puntajes que surgen de programas de evaluación conformes a normas pueden proporcionar interpretaciones de puntajes absolutas razonables, y los puntajes que surgen de programas conformes a criterios pueden proporcionar interpretaciones de puntajes relativas razonables.

Estándar 4.23

Cuando el puntaje de una prueba se deriva de la ponderación diferencial de ítems o subpuntajes, el desarrollador de la prueba debe documentar la justificación y el proceso utilizados para desarrollar, revisar y asignar ponderaciones de ítems. Cuando se obtienen ponderaciones de ítems sobre la base de datos empíricos, la muestra utilizada para obtener las ponderaciones de los ítems debe ser representativa de la población para la que está prevista la prueba y suficientemente grande para proporcionar estimaciones exactas de ponderaciones óptimas. Cuando se obtienen ponderaciones de ítems sobre la base de juicio de expertos, las cualificaciones de los jueces deben documentarse.

Comentario: Los cambios en la población de examinandos, junto con otros cambios, por ejemplo, en instrucciones, capacitación o requisitos de empleo, pueden afectar las ponderaciones de ítems derivadas originales, lo cual necesita estudios posteriores. En muchos casos, las áreas de contenido se ponderan especificando una cantidad diferente de ítems de áreas diferentes. La justificación para ponderar las diferentes áreas de contenido debe también documentarse y revisarse en forma periódica.

Unidad 4. Estándares para revisión de pruebas

Estándar 4.24

Las especificaciones de la prueba deben modificarse o revisarse cuando nuevos datos de investigación, cambios significativos en el dominio representado o condiciones recientemente recomendadas del uso de la prueba pueden reducir la validez de las interpretaciones de los puntajes de la prueba. Si bien no es necesario que una prueba que mantiene su utilidad sea retirada o revisada simplemente debido al paso del tiempo, los desarrolladores de la prueba y los editores de la prueba son responsables de supervisar condiciones cambiantes y de modificar, revisar o retirar la prueba según lo indicado.

Comentario: Los desarrolladores de la prueba deben considerar una serie de factores que pueden justificar la revisión de una prueba, incluyendo contenido y lenguaje de la prueba desactualizados, nueva evidencia de relaciones entre los constructos medidos y predichos, o cambios en los marcos de prueba para reflejar cambios en el plan de estudios, la instrucción o los requisitos de empleo. Si se utiliza una versión más antigua de una prueba cuando se ha publicado o puesto a disposición una versión más nueva, los usuarios de la prueba son responsables de proporcionar evidencia de que la versión más antigua es tan apropiada como la nueva para ese uso en particular de la prueba.

Estándar 4.25

Cuando las pruebas se revisan, se debe informar a los usuarios de los cambios en las especificaciones, de cualquier ajuste hecho a la escala de puntajes y del grado de comparabilidad de puntajes de las pruebas originales y revisadas. Las pruebas deben indicarse como "revisadas" solo cuando las especificaciones de la prueba hayan sido actualizadas de maneras significativas.

Comentario: Es responsabilidad del desarrollador de la prueba determinar si las revisiones a una prueba influirían en las interpretaciones de los puntajes de la prueba. Si las interpretaciones de los puntajes de la prueba serían afectadas por las revisiones, es apropiado indicar la prueba como "revisada". Cuando las pruebas se revisan, deben documentarse la naturaleza de las revisiones y sus implicaciones para las interpretaciones de los puntajes de la prueba. Ejemplos de cambios que requieren consideración incluyen agregar nuevas áreas de contenido, mejorar las descripciones de contenido, redistribuir el énfasis entre diferentes áreas de contenido, e incluso solo cambiar especificaciones del formato de los ítems. Obsérvese que crear un nuevo formulario de prueba usando las mismas especificaciones no se considera una revisión dentro del contexto de este estándar.

5. PUNTAJES, ESCALAS, NORMAS, VINCULACIÓN DE PUNTAJES Y PUNTAJES DE CORTE

ANTECEDENTES

Los puntajes de la prueba se reportan en escalas diseñadas para ayudar con la interpretación de los puntajes. Por lo general, la calificación comienza con respuestas a ítems de la prueba por separado. Estos puntajes de los ítems se combinan, a veces mediante la suma, para obtener un puntaje bruto cuando se usa la teoría clásica de los tests o para producir un puntaje de TRI cuando se utilizan la *teoría de respuesta al ítem* (TRI) u otras técnicas basadas en modelos. Los puntajes brutos y los puntajes de TRI a menudo son difíciles de interpretar en ausencia de mayor información. La interpretación puede facilitarse mediante la conversión de los puntajes brutos o puntajes de TRI a puntajes de escala. Ejemplos incluyen varios puntajes de escala utilizados en las pruebas de admisiones universitarias y los usados para reportar resultados para pruebas de inteligencia o inventarios de interés vocacional y de personalidad. El proceso de desarrollar una escala de puntajes se denomina *escalamiento de una prueba*. Los puntajes de escala pueden contribuir a la interpretación indicando cómo es un puntaje dado en comparación con los de otros examinandos, mejorando la comparabilidad de puntajes obtenidos a través de diferentes formularios de una prueba y ayudando a evitar confusión con otros puntajes.

Otra manera de ayudar en la interpretación de puntajes es establecer puntajes de corte que distingan diferentes rangos de puntajes. En algunos casos, un solo puntaje de corte define el límite entre aprobar y reprobar. En otros casos, una serie de puntajes de corte define distintos niveles de competencia. Los puntajes de escala, niveles de competencia y puntajes de corte pueden ser centrales para el uso y la interpretación de puntajes de la prueba. Por ese motivo, la posibilidad de defenderlos es una consideración importante en la validación de puntajes de la prueba para los fines previstos.

Las decisiones sobre cuántos puntos de puntaje de escala usar suele basarse en cuestiones de confiabilidad de los puntajes de la prueba. Si se utilizan muy pocos puntos de puntaje de escala, entonces la confiabilidad de los puntajes de escala se reduce a medida que se descarta información. Si se utilizan demasiados puntos de puntajes de escala, los usuarios de la prueba podrían intentar interpretar diferencias de puntajes de escala que son pequeñas en relación con la cantidad de error de medida en los puntajes.

Además de facilitar interpretaciones de puntajes en un solo formulario de prueba, los puntajes de escala a menudo se crean para mejorar la comparabilidad entre *formularios alternativos*[2] de la misma prueba, usando *métodos de equiparación*. Vinculación de puntajes es un término general para métodos utilizados para desarrollar escalas con propiedades de escala similares. La vinculación de puntajes incluye la equiparación y otros métodos para transformar puntajes para mejorar su comparabilidad en pruebas diseñadas para medir diferentes constructos (p. ej., subpruebas relacionadas en una batería). Los métodos de vinculación también se usan para relacionar puntajes de escala en diferentes medidas de constructos

[2]El término *formulario alternativo* se utiliza en este capítulo para indicar formularios de prueba que se han elaborado según las mismas especificaciones de contenido y estadísticas y desarrollado para medir el mismo constructo. Este término no debe confundirse con el término evaluación alternativa como se utiliza en el capítulo 3, para indicar una prueba que se ha modificado o cambiado para aumentar el acceso al constructo para subgrupos de la población. La evaluación alternativa puede o no medir el mismo constructo que la evaluación no alterada.

similares (p. ej., pruebas de un constructo en particular de diferentes desarrolladores de pruebas) y para relacionar puntajes de escala en pruebas que miden constructos similares dados en modos diferentes de administración (p. ej., administraciones por computadora y con papel y lápiz). *Los métodos de escalamiento vertical* se utilizan a veces para colocar puntajes de diferentes niveles de una prueba de rendimiento en una sola escala con el fin de facilitar inferencias sobre crecimiento o desarrollo. El grado de comparabilidad de puntajes que se deriva de la aplicación de un procedimiento de vinculación varía a lo largo de un continuum. La equiparación tiene por objeto permitir que puntajes en formularios alternativos de una prueba se utilicen de manera intercambiable, mientras que la comparabilidad de puntajes asociada con otros tipos de vinculación puede ser más restringida.

Interpretaciones de puntajes

Los puntajes brutos o puntajes de escala de un individuo a menudo se comparan con la distribución de puntajes para uno o más grupos de comparación para derivar inferencias útiles sobre el desempeño relativo de la persona. Se dice que las interpretaciones de los puntajes de la prueba basadas en esas comparaciones son conformes a normas. Las normas de rango de percentil, por ejemplo, indican la situación de un individuo o grupo dentro de una población definida de individuos o grupos. Un ejemplo podrían ser los puntajes de percentil utilizados en las pruebas de reclutamiento militar, que comparan el puntaje de cada postulante con puntajes para la población de jóvenes estadounidenses de 18 a 23 años. Los percentiles, promedios u otras estadísticas para dichos grupos de referencia se llaman normas. Al mostrar cómo es el puntaje de la prueba de un individuo examinado determinado en comparación con los de otros, las normas ayudan en la clasificación o descripción de individuos examinados.

Otras interpretaciones de puntajes de la prueba no hacen referencia directa al desempeño de otros individuos examinados. Estas interpretaciones pueden adoptar diversas formas; la mayoría se denominan en forma conjunta como interpretaciones *conformes a criterios*. Los puntajes de escala que respaldan esas interpretaciones pueden indicar la proporción probable de respuestas correctas que se obtendrían en algún dominio más grande de ítems similares, o la probabilidad de que un individuo examinado responda tipos particulares de ítems correctamente. Otras interpretaciones conformes a criterios pueden indicar la probabilidad de que haya presente alguna psicopatología. Además, otras interpretaciones conformes a criterios pueden indicar la probabilidad de que el nivel de conocimiento o habilidad evaluado de un individuo examinado sea adecuado para desempeñarse con éxito en algún otro contexto. Los puntajes de escala para respaldar esas interpretaciones de puntaje conformes a criterios suelen desarrollarse sobre la base de análisis estadísticos de las relaciones de los puntajes de la prueba con otras variables.

Algunos puntajes de escala se desarrollan principalmente para respaldar interpretaciones conformes a normas; otros respaldan interpretaciones conformes a criterios. En la práctica, no obstante, no siempre hay una distinción marcada. Tanto las escalas conformes a criterios como las conformes a normas pueden desarrollarse y utilizarse con los mismos puntajes de la prueba si se usan métodos apropiados para validar cada tipo de interpretación. Sin embargo, una escala de puntajes conforme a normas originalmente desarrollada, por ejemplo, para indicar desempeño en relación con alguna población de referencia específica podría, con el tiempo, también llegar a respaldar interpretaciones conformes a criterios. Esto podría ocurrir puesto que la investigación y la experiencia aportan mayor comprensión de las capacidades implícitas por los diferentes niveles de puntajes de escala. Al contrario, los resultados de una evaluación educativa podrían reportarse en una escala compuesta por varios niveles de competencia ordenados, definidos por descripciones de las clases de tareas que pueden realizar los estudiantes en cada nivel. Esa sería una escala conforme a criterios, pero una vez que se reporta la distribución de puntajes en niveles, supongamos, para todos los estudiantes de octavo grado en un estado

determinado, los puntajes de cada estudiante también transmitirán información sobre su situación en relación con la población evaluada.

Las interpretaciones basadas en puntajes de corte pueden del mismo modo ser conformes a criterios o conformes a normas. Si descripciones cualitativamente diferentes se asocian a rangos de puntajes sucesivos, se admite una interpretación conforme a criterios. Por ejemplo, las descripciones de niveles de competencia en algunas rúbricas de puntajes de tareas de evaluaciones pueden mejorar la interpretación de puntajes resumiendo las capacidades que deben demostrarse para merecer un puntaje dado. En otros casos, las interpretaciones conformes a criterios pueden basarse en relaciones determinadas empíricamente entre los puntajes de la prueba y otras variables. Pero cuando las pruebas se utilizan para selección, es posible que sea apropiado clasificar a los individuos examinados de acuerdo con su desempeño en la prueba y establecer un puntaje de corte para seleccionar una cantidad o proporción preespecificada de individuos examinados de un extremo de la distribución, siempre que el uso de la selección esté suficientemente respaldado por evidencia de confiabilidad y validez relevante para respaldar la clasificación. En esos casos, la interpretación de los puntajes de corte es conforme a normas; las etiquetas "rechazar" o "reprobar" frente a "aceptar" o "aprobar" son determinadas principalmente por la situación del individuo examinado en relación con otros evaluados en el proceso de selección actual.

Las interpretaciones conformes a criterios basadas en puntajes de corte a veces son criticadas con el argumento de que pocas veces existe una distinción marcada entre aquellos apenas por encima y aquellos apenas por debajo de un puntaje de corte. Una prueba neuropsicológica puede ser útil en el diagnóstico de algún deterioro en particular, por ejemplo, pero es probable que la probabilidad de que el deterioro esté presente aumente en forma continua como una función del puntaje de la prueba en lugar de cambiar notoriamente en un puntaje en particular. Los puntajes de corte pueden ayudar a formular reglas para arribar a decisiones sobre la base del desempeño en la prueba. Debe reconocerse, no obstante, que la probabilidad de clasificación errónea por lo general será relativamente alta para personas con puntajes cercanos a los puntajes de corte.

Normas

La validez de interpretaciones conformes a normas depende en parte de la adecuación del grupo de referencia con el cual se comparan los puntajes de la prueba. Las normas basadas en pacientes hospitalizados, por ejemplo, podrían ser inapropiadas para algunas interpretaciones de puntajes de pacientes no hospitalizados. Por lo tanto, es importante que las poblaciones de referencia se definan cuidadosamente y se describan con claridad. La validez de esas interpretaciones también depende de la exactitud con la que las normas resumen el desempeño de la población de referencia. La población puede ser suficientemente pequeña para que básicamente toda la población pueda evaluarse (p. ej., todos los examinandos en un nivel de grado dado en un distrito dado evaluados en la misma ocasión). A menudo, no obstante, solo se evalúa una muestra de individuos examinados de la población de referencia. Es por lo tanto importante que las normas se basen en una muestra representativa, técnicamente sólida de examinandos de tamaño suficiente. Es poco probable que los pacientes en algunos hospitales en una región geográfica pequeña sean representativos de todos los pacientes en Estados Unidos, por ejemplo. Además, la utilidad de las normas basadas en una muestra determinada puede reducirse con el tiempo. Por lo tanto, para pruebas que han estado en uso durante varios años, por lo general se requiere una revisión periódica para asegurar la utilidad continua de sus normas. Es posible que se requiera renormalización para mantener la validez de interpretaciones de puntajes de la prueba conformes a normas.

Más de una población de referencia puede ser apropiada para la misma prueba. Por ejemplo, el desempeño en una prueba de rendimiento podría interpretarse por referencia a normas locales sobre la base de muestreo de un distrito escolar en particular para uso en la toma de decisiones sobre

instrucción locales, o a normas para un estado o tipo de comunidad para usar en la interpretación de resultados de evaluación a nivel estatal, o a normas nacionales para usarse al hacer comparaciones con grupos nacionales. Para otras pruebas, las normas podrían basarse en clasificaciones ocupacionales o educativas. Las estadísticas descriptivas para todos los individuos examinados que resultan ser evaluados durante un período de tiempo determinado (a veces denominadas *normas de usuario* o *normas de programa*) pueden ser útiles para algunos fines, como describir tendencias conforme avanza el tiempo. Pero debe haber un motivo sólido para considerar a ese grupo de examinandos como una base apropiada para dichas inferencias. Cuando existe una justificación adecuada para usar a dicho grupo, las estadísticas descriptivas deben caracterizarse claramente como basadas en una muestra de personas habitualmente evaluadas como parte de un programa continuo.

Vinculación de puntajes

Vinculación de puntajes es un término general que se refiere a relacionar puntajes de diferentes pruebas o formularios de prueba. Cuando diferentes formularios de una prueba se construyen según las mismas especificaciones de contenido y estadísticas, y se administran en las mismas condiciones, se denominan formularios alternativos o a veces formularios *paralelos* o *equivalentes*. El proceso de colocar puntajes brutos de dichos formularios alternativos en una escala común se denomina *equiparación*. La equiparación involucra pequeños ajustes estadísticos para representar diferencias menores en la dificultad de los formularios alternativos. Después de la equiparación, los formularios alternativos de la misma prueba arrojan puntajes de escala que pueden usarse en forma intercambiable aunque se basen en diferentes conjuntos de ítems. En muchos programas de evaluación que administran pruebas múltiples veces, pueden plantearse preocupaciones sobre la seguridad de la prueba si el mismo formulario se usa en forma reiterada. En otros programas de evaluación, los mismos examinandos pueden ser medidos en forma reiterada, tal vez para medir cambios en los niveles de disfunción psicológica, actitudes o rendimiento educativo. En estos casos, reutilizar los mismos ítems de la prueba puede dar lugar a estimaciones de cambio sesgadas. La equiparación de puntajes permite el uso de formularios alternativos, con lo cual se evitan estas preocupaciones.

Si bien los formularios alternativos se elaboran según las mismas especificaciones de contenido y estadísticas, ocurrirán diferencias en la dificultad de la prueba, lo que generará la necesidad de equiparación. Un enfoque hacia la equiparación implica administrar los formularios a equiparar a la misma muestra de individuos examinados o a muestras equivalentes. Otro enfoque involucra administrar un conjunto común de ítems, denominados ítems de anclaje, a las muestras que toman cada formulario. Cada enfoque tiene fortalezas exclusivas, pero también involucra suposiciones que podrían influir en los resultados de la equiparación, y por lo tanto estas suposiciones deben verificarse. Elegir entre enfoques de equiparación puede incluir las siguientes consideraciones:

- Administrar formularios a la misma muestra permite una estimación de la correlación entre los puntajes de los dos formularios, así como proporcionar datos necesarios para ajustar por diferencias en la dificultad. Sin embargo, podría haber efectos de orden relacionados con la práctica o fatiga que pueden afectar la distribución de puntajes para el formulario administrado en segundo lugar.
- Administrar formularios alternativos a muestras equivalentes, por lo general mediante asignación aleatoria, evita cualquier efecto de orden pero no proporciona una estimación directa de la correlación entre los puntajes; otros métodos son necesarios para demostrar que los dos formularios miden el mismo constructo.
- Incorporar un conjunto de ítems de anclaje en cada uno de los formularios que se equiparan proporciona una base para ajustar por diferencias en las muestras de individuos examinados que completan cada formulario. Los

ítems de anclaje deben cubrir el mismo contenido y rango de dificultad que cada uno de los formularios completos que se equiparan de modo que las diferencias en los ítems de anclaje reflejarán de manera exacta diferencias en los formularios completos. Además, la posición de los ítems de anclaje y otros factores de contexto deben ser los mismos en ambos formularios. Es importante verificar que los ítems de anclaje funcionen de manera similar en los formularios que se equiparan. Los ítems de anclaje a menudo se retiran del anclaje si su dificultad relativa es sustancialmente diferente en los formularios que se equiparan.

- A veces se utiliza una prueba de anclaje externa en la que los ítems de anclaje se administran en una sección separada y no contribuyen al puntaje total de la prueba. Este enfoque elimina algunos factores de contexto dado que la presentación de los ítems de anclaje es idéntica para cada muestra de individuos examinados. Nuevamente, no obstante, la prueba de anclaje debe reflejar el contenido y la dificultad de los formularios operativos que se equiparan. Los diseños de pruebas de anclaje tanto incorporadas como externas involucran fuertes suposiciones estadísticas respecto de la equivalencia del anclaje y los formularios que se equiparan. Estas suposiciones son particularmente críticas cuando las muestras de individuos examinados que completan los diferentes formularios varían considerablemente en el constructo que se mide.

Cuando se afirma que los puntajes en los formularios de prueba están equiparados, es importante documentar cómo los formularios se elaboran según las mismas especificaciones de contenido y estadísticas y demostrar que los puntajes en los formularios alternativos son medidas del mismo constructo y tienen confiabilidad similar. La equiparación debe proporcionar conversiones de puntaje exactas para cualquier conjunto de personas tomado de la población de individuos examinados para la que se diseña la prueba; por lo tanto, la estabilidad de las conversiones entre subgrupos relevantes debe documentarse. Cuando sea posible, las definiciones de poblaciones importantes de individuos examinados deben incluir grupos para los que la imparcialidad puede ser una cuestión particular, como individuos examinados con discapacidades o de características lingüísticas y culturales diversas. Cuando los tamaños de la muestra lo permitan, es importante examinar la estabilidad de las conversiones de equiparación entre estas poblaciones.

El mayor uso de pruebas ejecutadas por computadora plantea consideraciones especiales para la equiparación y la vinculación porque se hacen posibles modelos más flexibles para ejecutar pruebas. Estos incluyen pruebas adaptables así como enfoques en los que se seleccionan ítems exclusivos o múltiples conjuntos intactos de ítems de un conjunto más grande de ítems disponibles. Hace mucho tiempo que se reconoce que poco se aprende de las respuestas de los individuos examinados a ítems que son demasiado fáciles o demasiado difíciles para ellos. En consecuencia, algunos procedimientos de evaluación utilizan solo un subconjunto de los ítems disponibles con cada individuo examinado. Una prueba adaptable consiste en un conjunto de ítems junto con reglas para seleccionar un subconjunto de los ítems que se administrarán a cada individuo examinado y un procedimiento para colocar los puntajes de diferentes individuos examinados en una escala común. La selección de ítems sucesivos se basa en parte en las respuestas de los individuos examinados a ítems anteriores. Pueden diseñarse reglas de selección de ítems y de conjuntos de ítems de modo que cada individuo examinado reciba un conjunto representativo de ítems de dificultad apropiada. Con algunas pruebas adaptables, puede resultar que dos individuos examinados casi nunca, o nunca, reciban el mismo conjunto de ítems. Además, es posible dar a dos individuos examinados que hacen la misma prueba adaptable conjuntos de ítems que difieren marcadamente en cuanto a dificultad. No obstante, los puntajes de la prueba adaptable pueden reportare en una escala común y funcionar de manera muy similar a puntajes de un solo formulario alternativo de una prueba que no es adaptable.

A menudo, la adaptación de la prueba se hace ítem por ítem. En otras situaciones, como en evaluaciones de múltiples etapas, el proceso de examen puede dividirse desde elegir entre conjuntos de ítems que son en líneas generales representativos del contenido y la dificultad hasta elegir entre conjuntos de ítems que son destinados explícitamente para un nivel mayor o menor del constructo sometido a medición, sobre la base de una evaluación provisional del desempeño del individuo examinado.

En muchas situaciones, los conjuntos de ítems para pruebas adaptables se actualizan reemplazando algunos de los ítems en el conjunto con nuevos ítems. En otros casos, se reemplazan los conjuntos de ítems enteros. En cualquier caso, se utilizan procedimientos estadísticos para vincular estimaciones de parámetros de ítems para los nuevos ítems con la escala de TRI existente de modo que los puntajes de conjuntos alternativos puedan usarse en forma intercambiable, de manera muy similar en que los puntajes en formularios alternativos se utilizan cuando los puntajes en los formularios alternativos se equiparan. Para respaldar la comparabilidad de puntajes en pruebas adaptables entre conjuntos, es necesario construir los conjuntos según las mismas especificaciones explícitas de contenido y estadísticas y administrarlos en las mismas condiciones. Generalmente, un diseño de ítems comunes se utiliza en la vinculación de estimaciones de parámetros para los nuevos ítems a la escala de TRI utilizada para pruebas adaptables. En esos casos, deben hacerse verificaciones de estabilidad sobre las características estadísticas de los ítems comunes, y la cantidad de ítems comunes debe ser suficiente para arrojar resultados estables. Debe verificarse la adecuación de las suposiciones necesarias para vincular puntajes entre conjuntos.

Existen muchos otros ejemplos de vinculación que pueden no resultar en puntajes intercambiables, incluyendo los siguientes:

- Para la evaluación del crecimiento de los individuos examinados con el tiempo, es posible que sea aconsejable desarrollar escalas verticales que abarquen un amplio rango de niveles de desarrollo o educativos. El desarrollo de escalas verticales suele requerir vinculación de pruebas que se construyen deliberadamente para diferir en dificultad.
- La revisión de la prueba a menudo genera la necesidad de vincular puntajes obtenidos utilizando especificaciones de la prueba más nuevas y más viejas.
- Estudios comparativos internacionales pueden requerir vinculación de puntajes en las pruebas dadas en diferentes idiomas.
- Los puntajes pueden vincularse en pruebas que miden diferentes constructos, tal vez comparando una aptitud con una forma de comportamiento, o vinculando medidas de rendimiento en varias áreas de contenido o entre diferentes editores de la prueba.
- En ocasiones se hacen vinculaciones para comparar el desempeño de grupos (p. ej., distritos escolares, estados) en diferentes medidas de constructos similares, como cuando se vinculan puntajes en una prueba de rendimiento estatal con puntajes en una evaluación internacional.
- Los resultados de los estudios de vinculación a veces se alinean o presentan en una tabla de concordancia para ayudar a los usuarios a estimar el desempeño en una prueba a partir del desempeño en otra.
- En situaciones en las que se utilizan tipos de ítems complejos, la vinculación de puntajes a veces se realiza a través de juicios sobre la comparabilidad del contenido del ítem de una prueba a otra. Por ejemplo, indicaciones de redacción elaboradas para ser similares, en las que las respuestas se califiquen utilizando una rúbrica común, podrían suponerse equivalente en términos de dificultad. Cuando sea posible, estas vinculaciones deben verificarse empíricamente.
- En algunas situaciones, se utilizan métodos basados en juicios para vincular puntajes entre pruebas. En estas situaciones, los procesos de juicio y su confiabilidad deben estar bien documentados y la justificación para su uso debe ser clara.

Los procesos utilizados para facilitar comparaciones pueden describirse con términos tales como *vinculación, calibración, concordancia, escalamiento vertical, proyección* o *moderación*. Estos procesos pueden ser técnicamente sólidos y pueden satisfacer completamente las metas de compatibilidad deseadas para una finalidad o para un subgrupo relevante de individuos examinados, pero no puede suponerse que sean estables con el tiempo o invariantes entre múltiples subgrupos de la población de individuos examinados, y tampoco hay ninguna garantía de que los puntajes obtenidos utilizando diferentes pruebas sean igualmente precisos. Por lo tanto, su uso para otros fines o con otras poblaciones que no sean la población originalmente prevista puede requerir respaldo adicional. Por ejemplo, una conversión de puntajes que fue exacta para un grupo de hablantes nativos podría sistemáticamente sobrepredecir o infrapredecir los puntajes de un grupo de hablantes no nativos.

Puntajes de corte

Un paso crítico en el desarrollo y uso de algunas pruebas es establecer uno o más puntajes de corte dividiendo el rango de puntajes para separar la distribución de puntajes en categorías. Estas categorías pueden utilizarse solo para fines descriptivos o pueden usarse para distinguir entre individuos examinados para los que se consideran aconsejables diferentes programas o para los que se justifican diferentes predicciones. Un empleador puede determinar un puntaje de corte para seleccionar posibles empleados o para promover a los empleados actuales; pueden establecerse niveles de competencia de "básico", "competente" y "avanzado" utilizando métodos de fijación de estándares para fijar puntajes de corte en una prueba estatal de rendimiento matemático en cuarto grado; es posible que los educadores quieran usar puntajes de la prueba para identificar a estudiantes que están preparados para continuar con la universidad y tomar cursos que dan créditos; o en la obtención de una licencia profesional, un estado puede especificar un puntaje de aprobación mínimo en una prueba para la obtención de la licencia.

Estos ejemplos difieren en aspectos importantes, pero todos involucran delinear categorías de individuos examinados sobre la base de puntajes de la prueba. Estos puntajes de corte proporcionan la base para usar e interpretar resultados de la prueba. Por lo tanto, en algunas situaciones, la validez de las interpretaciones de los puntajes de la prueba puede depender de los puntajes de corte. No puede haber un solo método para determinar puntajes de corte para todas las pruebas o para todos los fines, ni un único conjunto de procedimientos para establecer su posibilidad de defenderlos. Además, aunque los puntajes de corte son útiles para informar la selección, colocación, y otras clasificaciones, debe reconocerse que dichas decisiones categóricas rara vez se toman sobre la base del desempeño en la prueba únicamente. Las situaciones a continuación solo ejemplos.

El primer ejemplo, de un empleador que entrevista a todos lo que obtienen puntajes por encima de un nivel determinado en una prueba de empleo, es el más directo. Suponiendo que se haya proporcionado evidencia de validación para los puntajes de la prueba de empleo para su uso previsto, por lo general se esperaría que el desempeño laboral promedio aumente en forma constante, aunque lenta, con cada incremento en el puntaje de la prueba, al menos para algún rango de puntajes cercanos al puntaje de corte. En ese caso, la designación del valor particular para el puntaje de corte puede determinarse principalmente por la cantidad de personas a ser entrevistadas o que continuarán siendo cribadas.

En el segundo ejemplo, un departamento de educación estatal establece estándares de contenido para lo que los estudiantes de cuarto grado deben aprender en matemáticas e implementa una prueba para evaluar el rendimiento de los estudiantes en estos estándares. Utilizando un proceso de fijación de estándares basado en juicios, estructurado, comités de expertos en la materia desarrollan o elaboran descriptores *de nivel de desempeño* (a veces denominados *descriptores de nivel de rendimiento*) que indican qué deberían saber y poder hacer en matemáticas de cuarto grado los estudiantes en los niveles de rendimiento "básico", "competente" y "avanzado".

Además, comités examinan ítems de la prueba y desempeño estudiantil para recomendar puntajes de corte que se usarán para asignar a estudiantes a cada nivel de rendimiento sobre la base de su desempeño en la prueba. La decisión final sobre los puntajes de corte es una decisión de políticas que por lo general toma un organismo de políticas como un consejo de educación para el estado.

En el tercer ejemplo, educadores desean utilizar puntajes de la prueba para identificar a estudiantes que están preparados para continuar con la universidad y tomar cursos que otorgan créditos. Los puntajes de corte podrían identificarse inicialmente sobre la base de juicios sobre requisitos para tomar cursos que otorgan créditos en una serie de universidades. Alternativamente, podrían reunirse juicios sobre estudiantes individuales y luego utilizarse para buscar un nivel de puntaje que diferencie de manera más efectiva a quienes se considera preparados de los que se considera no están preparados. En esos casos, los jueces deben estar familiarizados tanto con los requisitos del curso universitario como con los propios estudiantes. Cuando sea posible, podría hacerse un seguimiento de los juicios iniciales con datos longitudinales que indiquen si anteriores individuos examinados tomaron o no cursos de apoyo.

En el último ejemplo, el de un examen para la obtención de una licencia profesional, el puntaje de corte representa un juicio informado de que quienes obtienen puntajes por debajo de él están en riesgo de cometer graves errores porque carecen del conocimiento o las habilidades evaluadas. Ninguna prueba es perfecta, por supuesto, e independientemente de los puntajes de corte elegidos, es probable que algunos individuos examinados con habilidades insuficientes aprueben, y algunos con habilidades suficientes reprueben. Las probabilidades relativas de esos errores de falso positivo y falso negativo variarán dependiendo del puntaje de corte elegido. Una probabilidad dada de exponer al público a posible daño emitiendo una licencia a un individuo incompetente (falso positivo) debe ponderarse frente a alguna probabilidad correspondiente de denegar una licencia, y así inhabilitar, a un individuo examinado cualificado (falso negativo). Cambiar el puntaje de corte para reducir cualquiera de las dos probabilidades aumentará la otra, aunque ambas clases de errores pueden minimizarse mediante un diseño de la prueba sólido que prevea el rol del puntaje de corte en el uso y la interpretación de la prueba. Determinar puntajes de corte en esas situaciones no puede ser un asunto meramente técnico, aunque estudios empíricos y modelos estadísticos pueden ser de gran valor para informar el proceso.

Los puntajes de corte incorporan juicios de valor, así como consideraciones técnicas y empíricas. Cuando los resultados del proceso de fijación de estándares tienen consecuencias altamente significativas, los involucrados en el proceso de fijación de estándares deben preocuparse de que el proceso por el cual se determinan los puntajes de corte se documente claramente y que sea defendible. Cuando la fijación de estándares involucra a jueces o expertos en la materia, sus cualificaciones y el proceso por el cual fueron seleccionados son parte de esa documentación. Se debe tener cuidado de garantizar que estas personas comprendan lo que deben hacer y que sus juicios sean tan razonados y objetivos como sea posible. El proceso debe ser tal que participantes bien cualificados puedan aplicar su conocimiento y experiencia para arribar a juicios significativos y relevantes que reflejen exactamente sus entendimientos e intenciones. Debe emplearse un grupo de participantes suficientemente grande y representativo para proporcionar una seguridad razonable de que las calificaciones de expertos entre jueces sean suficientemente confiables y que los resultados de los juicios no varíen en gran medida si el proceso se replicara.

ESTÁNDARES PARA PUNTAJES, ESCALAS, NORMAS, VINCULACIÓN DE PUNTAJES Y PUNTAJES DE CORTE

Los estándares en este capítulo comienzan con un estándar global (numerado 5.0), que se ha diseñado para transmitir la intención central o enfoque principal del capítulo. El estándar global también puede verse como el principio rector del capítulo, y es aplicable a todas las pruebas y usuarios de pruebas. Todos los estándares posteriores se han separado en cuatro unidades temáticas denominadas de la siguiente manera:

1. Interpretaciones de puntajes
2. Normas
3. Vinculación de puntajes
4. Puntajes de corte

Estándar 5.0

Los puntajes de la prueba deben derivarse de una manera que respalde las interpretaciones de los puntajes de la prueba para los usos propuestos de las pruebas. Los desarrolladores y usuarios de la prueba deben documentar evidencia de imparcialidad, confiabilidad y validez de los puntajes de la prueba para su uso propuesto.

Comentario: A continuación, se describen diversos usos e interpretaciones de los puntajes de la prueba y escalas de puntajes. Estos incluyen estándares para interpretaciones conformes a normas y conformes a criterios, interpretaciones de puntajes de corte, posibilidad de intercambiar puntajes en formularios alternativos tras la equiparación, y comparabilidad de puntajes tras el uso de otros procedimientos para vinculación de puntajes. La documentación que respalda esas interpretaciones proporciona una base para que expertos externos y usuarios de la prueba juzguen en qué grado es probable que las interpretaciones sean respaldadas y pueden conducir a interpretaciones válidas de puntajes para todos los individuos en la población prevista de individuos examinados.

Unidad 1. Interpretaciones de puntajes

Estándar 5.1

Se deben proporcionar a los usuarios de la prueba explicaciones claras de las características, el significado y la interpretación prevista de los puntajes de escala, así como de sus limitaciones.

Comentario: Los ejemplos de interpretaciones apropiadas e inapropiadas pueden ser útiles, en especial para tipos de escalas o interpretaciones que no son conocidas para la mayoría de los usuarios. Este estándar corresponde a escalas de puntajes previstas para interpretaciones conformes a criterios y conformes a normas. Todos los puntajes (puntajes brutos y puntajes de escala) pueden estar sujetos a interpretación errónea. Si la naturaleza o los usos previstos de una escala son novedosos, es especialmente importante que sus usos, interpretaciones y limitaciones se describan claramente.

Estándar 5.2

Los procedimientos para construir escalas utilizadas para reportar puntajes y la justificación para estos procedimientos deben describirse claramente.

Comentario: Cuando las escalas, normas u otros sistemas interpretativos sean proporcionados por el desarrollador de la prueba, la documentación técnica debe describir su justificación y permitir que los usuarios juzguen la calidad y precisión de los puntajes de escala resultantes. Por ejemplo, el desarrollador de la prueba debe describir cualquier información normativa, de contenido o de precisión de puntajes que esté incorporada en la escala y proporcionar una justificación para

la cantidad de puntos de puntaje que se utilizan. Este estándar corresponde a escalas de puntajes previstas para interpretaciones conformes a criterios y conformes a normas.

Estándar 5.3

Si existe un motivo sólido para creer que son probables las interpretaciones erróneas específicas de una escala de puntajes, se debe advertir explícitamente a los usuarios de la prueba.

Comentario: Los editores y usuarios de la prueba pueden reducir las interpretaciones erróneas de puntajes de escala si describen explícitamente tanto los usos apropiados como los posibles usos indebidos. Por ejemplo, un punto de escala de puntajes originalmente definido como la media de alguna población de referencia debe dejar de interpretarse como representación del desempeño promedio si la escala se mantiene constante con el tiempo y la población de individuos examinados cambia. De manera similar, se necesita precaución si los significados de los puntajes pueden variar para algunos examinandos, como el significado de los puntajes de rendimiento para estudiantes que no han tenido la oportunidad adecuada de aprender el material cubierto por la prueba.

Estándar 5.4

Cuando está previsto que los puntajes brutos sean directamente interpretables, sus significados, interpretaciones previstas y limitaciones deben describirse y justificarse de la misma manera en que se hace para puntajes de escala.

Comentario: En algunos casos, los ítems en una prueba son una muestra representativa de un dominio bien definido de ítems con respecto tanto al contenido como a la dificultad de los ítems. La proporción respondida correctamente en la prueba puede entonces interpretarse como una estimación de la proporción de ítems en el dominio que podría responderse correctamente. En otros casos, diferentes interpretaciones pueden atribuirse a puntajes por encima o por debajo de un puntaje de corte en particular. Se debe ofrecer apoyo a cualquier interpretación de este tipo recomendada por el desarrollador de la prueba.

Estándar 5.5

Cuando los puntajes brutos o puntajes de escala se diseñan para interpretación conforme a criterios, incluyendo la clasificación de individuos examinados en categorías separadas, la justificación para las interpretaciones de puntajes recomendadas debe explicarse claramente.

Comentario: Las interpretaciones conformes a criterios son descripciones o inferencias basadas en puntajes que no adoptan la forma de comparaciones del desempeño en la prueba de un individuo examinado con el desempeño en la prueba de otros individuos examinados. Los ejemplos incluyen declaraciones de que probablemente haya alguna psicopatología presente, de que un potencial empleado posee habilidades específicas requeridas en un puesto dado, o de que un niño con un puntaje superior a determinado punto de puntaje puede aplicar con éxito un conjunto determinado de habilidades. Esas interpretaciones pueden referirse a los niveles absolutos de puntajes de la prueba o a patrones de puntajes para un solo individuo examinado. Cada vez que el desarrollador de la prueba recomienda dichas interpretaciones, deben presentarse claramente la justificación y el fundamento empírico. Deben hacerse esfuerzos serios cada vez que sea posible para obtener evidencia independiente respecto de la solidez de tales interpretaciones de puntajes.

Estándar 5.6

Los programas de evaluación que intentan mantener una escala común conforme avanza el tiempo deben realizar verificaciones periódicas de la estabilidad de la escala en la que se reportan los puntajes.

Comentario: La frecuencia de dichas verificaciones depende de varias características del programa de evaluación. En algunos programas de evaluación, los ítems se introducen y retiran de conjuntos de ítems en forma continua. En otros casos, los

ítems en formularios de prueba sucesivos pueden superponerse muy poco, o nada en absoluto. En cualquier caso, si se utiliza una escala fija para la presentación de reportes, es importante asegurar que el significado de los puntajes de escala no cambie con el transcurso del tiempo. Cuando las escalas se basan en la aplicación posterior de estimaciones de parámetros de ítems precalibrados utilizando teoría de respuesta al ítem, deben realizarse como rutina análisis de la estabilidad de los parámetros de ítems.

Estándar 5.7

Cuando se cambian pruebas o procedimientos de evaluación estandarizados para subgrupos relevantes de examinandos, el individuo o grupo que hace el cambio debe proporcionar evidencia de la comparabilidad de puntajes en las versiones cambiadas con puntajes obtenidos en las versiones originales de las pruebas. Si falta evidencia, se debe proporcionar documentación que advierta a los usuarios que los puntajes de la prueba o del procedimiento de evaluación cambiado pueden no ser comparables con los de la versión original.

Comentario: A veces se hace necesario cambiar versiones originales de una prueba o procedimiento de evaluación cuando la prueba se da a subgrupos relevantes de la población de evaluación, por ejemplo, individuos con discapacidades o individuos con características lingüísticas y culturales diversas. Una prueba puede traducirse a braille de modo que sea accesible a individuos que son ciegos, o el procedimiento de evaluación puede cambiarse para incluir tiempo extra para determinados grupos de individuos examinados. Estos cambios pueden o no tener un efecto en los constructos subyacentes medidos por la prueba y, en consecuencia, en las conversiones de puntajes utilizadas con la prueba. Si los puntajes en la prueba cambiada se compararán con puntajes en la prueba original, el desarrollador de la prueba debe proporcionar evidencia empírica de la comparabilidad de puntajes en la prueba cambiada y original cada vez que los tamaños de la muestra sean suficientemente grandes para proporcionar este tipo de evidencia.

Unidad 2. Normas

Estándar 5.8

Las normas, si se utilizan, deben referirse a poblaciones descriptas claramente. Estas poblaciones deben incluir individuos o grupos con los que los usuarios de la prueba desearán comúnmente comparar a sus propios individuos examinados.

Comentario: Es responsabilidad de los desarrolladores de la prueba describir normas claramente y responsabilidad de los usuarios de la prueba utilizar las normas de manera apropiada. Los usuarios deben conocer la aplicabilidad de una prueba a diferentes grupos. Normas diferenciadas o información resumida sobre diferencias entre grupos de género, raciales/étnicos, de idioma, discapacidad, grado o edad, por ejemplo, pueden ser útiles en algunos casos. Los usos aceptables de dichas normas diferenciadas e información relacionada pueden estar limitados por ley. Los usuarios también deben ser alertados sobre situaciones en las que las normas sean menos apropiadas para algunos grupos o individuos que para otros. En un inventario de interés ocupacional, por ejemplo, las normas para personas que realmente se dedican a una ocupación pueden ser inapropiadas para interpretar los puntajes de personas que no dedican a ella.

Estándar 5.9

Los reportes de estudios de normalización deben incluir especificación precisa de la población que se muestreó, los procedimientos de muestreo y las tasas de participación, cualquier ponderación de la muestra, las fechas de evaluación, y estadísticas descriptivas. La documentación técnica debe indicar la precisión de las normas propiamente dichas.

Comentario: La información proporcionada debe ser suficiente para permitir a los usuarios juzgar la adecuación de las normas para interpretar los puntajes de individuos examinados locales. La información debe presentarse de modo que cumpla

con los requisitos legales y estándares profesionales aplicables relacionados con la privacidad y la seguridad de los datos.

Estándar 5.10

Cuando las normas se utilizan para caracterizar a grupos de individuos examinados, las estadísticas utilizadas para resumir el desempeño de cada grupo y las normas a las que se refieren dichas estadísticas deben definirse claramente y deben respaldar el uso o interpretación previstos.

Comentario: No es posible determinar el rango de percentil del puntaje de prueba promedio de una escuela si todo lo que se conoce es el rango de percentil de cada uno de los estudiantes de esa escuela. Es posible que en ocasiones sea útil desarrollar normas especiales para medias de grupos, pero cuando los tamaños de los grupos difieren sustancialmente o cuando algunos grupos son más heterogéneos que otros, la construcción e interpretación de las normas de grupo es problemática. Un procedimiento común y aceptable es reportar el rango de percentil del miembro mediano del grupo, por ejemplo, el rango de percentil mediano de los alumnos evaluados en una escuela determinada.

Estándar 5.11

Si el editor de una prueba proporciona normas para usar en la interpretación de puntajes de la prueba, siempre que la prueba se mantenga en formato impreso, es responsabilidad del editor de la prueba renormalizar la prueba con suficiente frecuencia para permitir la continuidad de las interpretaciones de puntajes exactas y apropiadas.

Comentario: Los editores de la prueba deben asegurarse de que haya normas actualizadas inmediatamente disponibles o de proporcionar evidencia de que las normas anteriores aún son apropiadas. Sin embargo, continúa siendo responsabilidad del usuario de la prueba evitar el uso inapropiado de normas que estén desactualizadas y esforzarse por garantizar interpretaciones de puntajes exactas y apropiadas.

Unidad 3. Vinculación de puntajes

Estándar 5.12

Deben proporcionarse una justificación clara y evidencia de respaldo para cualquier afirmación de que los puntajes de escala obtenidos en formularios alternativos de una prueba pueden utilizarse en forma intercambiable.

Comentario: Para que los puntajes en formularios alternativos se utilicen en forma intercambiable, los formularios alternativos deben elaborarse según especificaciones detalladas de contenido y estadísticas en común. Deben reunirse datos adecuados y debe aplicarse metodología estadística apropiada para realizar equiparación de puntajes en formularios alternativos de la prueba. La calidad de la equiparación debe evaluarse para determinar si los puntajes de escala resultantes en los formularios alternativos pueden usarse en forma intercambiable.

Estándar 5.13

Cuando las afirmaciones de equivalencia de puntajes de un formulario a otro se basan en procedimientos de equiparación, debe proporcionarse información técnica detallada sobre el método por el cual se establecieron las funciones de equiparación y sobre la exactitud de las funciones de equiparación.

Comentario: Se debe proporcionar evidencia para demostrar que los puntajes equiparados en formularios alternativos miden esencialmente el mismo constructo con niveles muy similares de confiabilidad y errores estándares de medida condicionales y que los resultados son apropiados para subgrupos relevantes. La información técnica debe incluir el diseño del estudio de equiparación, los métodos estadísticos utilizados, el tamaño y las características relevantes de las muestras de individuos examinados utilizadas en los estudios de equiparación, y las características de cualquier prueba de anclaje o ítem de anclaje. En las pruebas para las que se realiza equiparación antes del

uso operativo (es decir, preequiparación), debe proporcionarse documentación del proceso de calibración de los ítems y la adecuación de las funciones de equiparación debe evaluarse tras la administración operativa. Cuando formularios equivalentes de pruebas basadas en computadora se construyan dinámicamente, deben documentarse los algoritmos utilizados y las características técnicas de los formularios alternativos deben evaluarse en función de simulación y/o análisis de datos de administración. Se deben estimar y reportar siempre que sea posible los errores estándares de las funciones de equiparación. Si los tamaños de la muestra lo permiten, es posible que resulte informativo evaluar si las funciones de equiparación desarrolladas para subgrupos relevantes de individuos examinados son similares. Es posible que sea informativo utilizar dos o más formularios de anclaje y realizar la equiparación utilizando cada uno de los anclajes. Para ser más útil, el error de equiparación debe presentarse en unidades de la escala de puntajes reportada. Para los programas de evaluación con puntajes de corte, el error de equiparación cercano al puntaje de corte es de primordial importancia.

Estándar 5.14

En estudios de equiparación que se basan en la equivalencia estadística de grupos de individuos examinados que reciben diferentes formularios, los métodos para establecer dicha equivalencia deben describirse en detalle.

Comentario: Determinados diseños de equiparación dependen de la equivalencia aleatoria de grupos que reciben diferentes formularios. A menudo, una manera de asegurar dicha equivalencia es mezclar sistemáticamente diferentes formularios de prueba y luego distribuirlos en forma aleatoria de modo que cantidades de individuos examinados aproximadamente iguales reciban cada formulario. Debido a que los diseños de administración que tienen por objeto arrojar grupos equivalentes no siempre se siguen en la práctica, la equivalencia de grupos debe evaluarse estadísticamente.

Estándar 5.15

En estudios de equiparación que emplean un diseño de prueba de anclaje, deben presentarse las características de la prueba de anclaje y su similitud con los formularios que se equiparan, incluyendo tanto especificaciones de contenido como relaciones determinadas en forma empírica entre los puntajes de la prueba. Si los ítems de anclaje se utilizan en el estudio de equiparación, deben presentarse la representatividad y las características psicométricas de los ítems de anclaje.

Comentario: Los puntajes en las pruebas o formularios de prueba pueden equipararse mediante ítems en común incorporados dentro de cada uno de ellos, o una prueba en común administrada junto con cada uno de ellos. Estos ítems o pruebas en común se *denominan ítems de vinculación*, ítems en común, ítems de anclaje o *pruebas de anclaje*. Los procedimientos estadísticos aplicados a ítems de anclaje hacen suposiciones que sustituyen la equivalencia alcanzada con un diseño de grupos equivalentes. Los desempeños en estos ítems son la única evidencia empírica utilizada para ajustar por diferencias en capacidad entre grupos antes de hacer ajustes por dificultad de la prueba. Con tales enfoques, la calidad de la equiparación resultante depende mucho de la cantidad de ítems de anclaje utilizados y de cuán bien los ítems de anclaje reflejen proporcionalmente el contenido y las características estadísticas de la prueba. El contenido de los ítems de anclaje debe ser exactamente el mismo en cada formulario de prueba que se equiparará. Los ítems de anclaje deben estar en posiciones similares para ayudar a reducir el error en la equiparación debido a efectos de contexto de los ítems. Además, deben hacerse verificaciones para asegurar que, después de controlar las diferencias de grupos de individuos examinados, los ítems de anclaje tengan características estadísticas similares en cada formulario de prueba.

Estándar 5.16

Cuando los puntajes de la prueba se basan en procedimientos psicométricos basados en modelos,

como los utilizados en pruebas adaptables computarizadas o de múltiples etapas, se debe proporcionar documentación para indicar que los puntajes tienen significado comparable en conjuntos alternativos de ítems de prueba.

Comentario: Cuando se utilizan procedimientos psicométricos basados en modelos, se debe proporcionar documentación técnica que respalde la comparabilidad de puntajes en conjuntos de ítems alternativos. Dicha documentación debe incluir las suposiciones y procedimientos que se utilizaron para establecer la comparabilidad, incluyendo descripciones claras de algoritmos basados en modelos, software utilizado, procedimientos de control de calidad que se siguieron, y análisis técnicos realizados que justifiquen el uso de modelos psicométricos para los puntajes de prueba en particular que tienen por objeto ser comparables.

Estándar 5.17

Cuando se vinculan puntajes en pruebas que no pueden equipararse, debe proporcionarse evidencia directa de la comparabilidad de puntajes, y la población de individuos examinados para la que se aplica la comparabilidad de puntajes debe especificarse claramente. La justificación específica y la evidencia requerida dependerán en parte de los usos previstos para los cuales se afirma la comparabilidad de puntajes.

Comentario: Se debe proporcionar respaldo para cualquier aseveración respecto de que puntajes vinculados obtenidos con uso de pruebas elaboradas según diferentes especificaciones de contenido y estadísticas, pruebas que utilizan diferentes materiales de prueba o pruebas que se administran en diferentes condiciones de administración de la prueba son comparables para la finalidad prevista. Para estas vinculaciones, debe especificarse claramente la población de individuos examinados para la que se establece la comparabilidad de puntajes. Este estándar se aplica, por ejemplo, a pruebas que difieren en extensión, pruebas administradas en diferentes formatos (p. ej., pruebas con papel y lápiz y basadas en computadora), formularios de prueba diseñados para administración individual frente a grupal, pruebas que se escalan verticalmente, pruebas adaptables computarizadas, pruebas que son sustancialmente revisadas, pruebas dadas en diferentes idiomas, pruebas administradas con varias adecuaciones, pruebas que miden diferentes constructos y pruebas de diferentes editores.

Estándar 5.18

Cuando se utilizan procedimientos de vinculación para relacionar puntajes en pruebas o formularios de prueba que no son muy paralelos, la construcción, la interpretación prevista y las limitaciones de esas vinculaciones deben describirse claramente.

Comentario: Se han realizado varias vinculaciones relacionando puntajes en pruebas desarrolladas en diferentes niveles de dificultad, relacionando formularios anteriores con formularios revisados de pruebas publicadas, creando concordancias entre diferentes pruebas de constructos similares o diferentes o para otros fines. Esas vinculaciones suelen ser útiles, pero también pueden estar sujetas a interpretación errónea. Las limitaciones de dichas vinculaciones deben describirse claramente. Se debe proporcionar información técnica detallada sobre la metodología de vinculación y la calidad de la vinculación. Se debe incluir información técnica sobre la vinculación, según corresponda, la confiabilidad de los conjuntos de puntajes que se vinculan, la correlación entre los puntajes de la prueba, una evaluación de la similitud del contenido, las condiciones de medición para cada prueba, el diseño de recopilación de datos, los métodos estadísticos utilizados, los errores estándares de la función de vinculación, evaluaciones de estabilidad de muestreo, y evaluaciones de comparabilidad de puntajes.

Estándar 5.19

Cuando las pruebas se crean tomando un subconjunto de los ítems en una prueba existente o reorganizando ítems, se debe proporcionar

evidencia de que no hay distorsiones de puntajes de escala, puntajes de corte o normas para las diferentes versiones o para vinculaciones de puntajes entre ellas.

Comentario: Algunas pruebas y baterías de pruebas se publican tanto en versión completa como en formato de sondeo o versión corta. En otros casos, pueden crearse múltiples versiones de un solo formulario de prueba reorganizando sus ítems. No debe suponerse que los datos de desempeño derivados de la administración de ítems como parte de la versión inicial pueden usarse para calcular puntajes de escala, calcular puntajes vinculados, construir tablas de conversión, aproximar normas o aproximar puntajes de corte para pruebas intactas alternativas. Se requiere precaución en casos en los que son probables efectos de contexto, incluyendo pruebas aceleradas, pruebas largas en las que la fatiga puede ser un factor, pruebas adaptables, y pruebas desarrolladas a partir de conjuntos de ítems calibrados. Las opciones para reunir evidencia relacionada con efectos de contexto podrían incluir exámenes de ajuste de datos de modelo, recalibraciones operativas de estimaciones de parámetros de ítems inicialmente derivadas utilizando datos de pruebas previas, y comparaciones de desempeño sobre formularios de pruebas originales y revisados según lo administrado a grupos equivalentes en forma aleatoria.

Estándar 5.20

Si las especificaciones de la prueba se cambian de una versión de una prueba a una versión posterior, dichos cambios deben identificarse, y se debe dar una indicación de que los puntajes convertidos para las dos versiones pueden no ser estrictamente equivalentes, incluso cuando se hayan usado procedimientos estadísticos para vincular puntajes de las versiones diferentes. Cuando ocurren cambios importantes en las especificaciones de la prueba, los puntajes deben reportarse en una nueva escala, o debe proporcionarse una declaración clara para alertar a los usuarios de que los puntajes no son directamente comparables con los de versiones anteriores de la prueba.

Comentario: A veces ocurren cambios importantes en las especificaciones de pruebas que se utilizan por períodos de tiempo sustanciales. A menudo, esos cambios aprovechan las mejoras en los tipos de ítems o cambios en el contenido que se haya demostrado mejoran la validez y por lo tanto son muy recomendables. Es importante reconocer, sin embargo, que dichos cambios darán por resultado puntajes que no pueden hacerse estrictamente intercambiables con puntajes en un formulario anterior de la prueba, incluso cuando se utilizan procedimientos de vinculación estadística. Para evaluar la comparabilidad de puntajes, es aconsejable evaluar la relación entre puntajes en las versiones anteriores y nuevas.

Unidad 4. Puntajes de corte

Estándar 5.21

Cuando las interpretaciones de puntajes propuestas involucran uno o más puntajes de corte, deben documentarse claramente la justificación y los procedimientos utilizados para establecer puntajes de corte.

Comentario: Los puntajes de corte pueden establecerse para seleccionar una cantidad especificada de individuos examinados (p. ej., identificar una cantidad fija de solicitantes de empleo para mayor cribado), en cuyo caso es posible que se necesite un poco más de documentación respecto de la pregunta específica de cómo se establecen los puntajes de corte, aunque se debe prestar atención a la justificación para usar la prueba en la selección y la precisión de comparaciones entre individuos examinados. En otros casos, no obstante, los puntajes de corte pueden usarse para clasificar individuos examinados en distintas categorías (p. ej., categorías de diagnóstico, niveles de competencia, o aprobar y reprobar) para las que no hay cuotas preestablecidas. En estos casos, el método de fijación de estándares debe documentarse con mayor detalle. Idealmente, el rol de los puntajes de corte en el uso y la interpretación de pruebas se tiene en cuenta durante el diseño de la prueba. La precisión adecuada en regiones

de escalas de puntajes donde se establecen puntajes de corte es un prerrequisito para la clasificación confiable de individuos examinados en categorías. Si la fijación de estándares emplea datos sobre distribuciones de puntajes para grupos de criterios o sobre la relación de los puntajes de la prueba con una o más variables de criterios, esos datos deben resumirse en la documentación técnica. Si se sigue un proceso de fijación de estándares basado en juicios, el método empleado debe describirse claramente, y debe presentarse la naturaleza precisa y la confiabilidad de los juicios requeridos, sean juicios de personas, de desempeños en ítems o en la prueba, o de desempeños en otros criterios predichos por los puntajes de la prueba. La documentación también debe incluir la selección y cualificaciones de participantes de paneles de fijación de estándares, la capacitación proporcionada, cualquier comentario a los participantes respecto de las implicaciones de sus juicios provisionales, y cualquier oportunidad para que los participantes deliberen entre ellos. Cuando corresponda, debe reportarse la variabilidad entre participantes. Cuando sea viable, se debe proporcionar una estimación de la cantidad de variación en los puntajes de corte que podría esperarse si el procedimiento de fijación de estándares se replicara con un panel de fijación de estándares comparable.

Estándar 5.22

Cuando los puntajes de corte que definen aprobado/reprobado o niveles de competencia se basen en juicios directos sobre la adecuación de los desempeños en el ítem o la prueba, el proceso basado en juicios debe diseñarse de modo que los participantes que proporcionan los juicios puedan aplicar su conocimiento y experiencia de una manera razonable.

Comentario: Los puntajes de corte a veces se basan en juicios sobre la adecuación de los desempeños en los ítems o la prueba (p. ej., respuestas de ensayos a una indicación de redacción) o expectativas de competencia (p. ej., el puntaje de escala que caracterizaría a un individuo examinado que está al límite). Los procedimientos utilizados para obtener dichos juicios deben dar por resultado estándares de competencia razonables y defendibles que reflejen con exactitud los valores e intenciones de los participantes en la fijación de estándares. Llegar a esos juicios puede ser más directo cuando se pide a los participantes que consideren clases de desempeño con las que están familiarizados y para las se han formado conceptos claros de adecuación y calidad. Cuando las repuestas suscitadas por una prueba no muestrean ni simulan de cerca el uso de conocimientos o habilidades evaluados en el dominio de criterios real, es probable que los participantes no aborden la tarea con ese entendimiento claro de adecuación y calidad. Se debe tener especial cuidado de asegurar que los participantes tengan un fundamento sólido para elaborar los juicios solicitados. El conocimiento exhaustivo de las descripciones de los diferentes niveles de competencia, la práctica en el juzgamiento de la dificultad de las tareas con comentarios sobre exactitud, la experiencia de efectivamente tomar un formulario de la prueba, comentarios sobre las tasas de aprobación que conllevan los estándares de competencia provisionales, y otras formas de información pueden ser beneficiosos para ayudar a los participantes a llegar a decisiones sólidas y ejemplares.

Estándar 5.23

Cuando sea viable y apropiado, los puntajes de corte que definan categorías con interpretaciones sustantivas distintas deben informarse mediante datos empíricos sólidos respecto de la relación del desempeño en la prueba con los criterios relevantes.

Comentario: En contextos laborales donde se haya establecido que los puntajes de prueba se relacionan con el desempeño laboral, es posible que la relación precisa de la prueba y el criterio tenga escasa incidencia en la elección de un puntaje de corte, si la elección se basa en la necesidad de una cantidad predeterminada de candidatos. Sin embargo, en contextos en que se aplican interpretaciones distintas a diferentes categorías de puntajes, la relación empírica de la prueba con el

criterio supone mayor importancia. Por ejemplo, si un puntaje de corte debe fijarse en una prueba de matemáticas de la escuela secundaria que indica la preparación para instrucción en matemáticas de nivel universitario, es posible que sea aconsejable reunir datos empíricos que establezcan una relación entre los puntajes de la prueba y las calificaciones obtenidas en cursos universitarios relevantes. Los puntajes de corte utilizados en la interpretación de pruebas de diagnóstico pueden establecerse sobre la base de distribuciones de puntajes determinadas en forma empírica para grupos de criterios. Con muchas pruebas de rendimiento o competencia, como las utilizadas en acreditación, grupos de criterios adecuados (p. ej., profesionales exitosos frente a no exitosos) a menudo no están disponibles. No obstante, cuando corresponda y sea viable, el desarrollador de la prueba debe investigar y reportar la relación entre los puntajes de la prueba y el desempeño en contextos prácticos relevantes. Se requiere juicio profesional para determinar un enfoque apropiado de fijación de estándares (o combinación de enfoques) en cualquier situación dada. En general, se esperaría encontrar una marcada diferencia en niveles de la variable de criterios entre aquellos apenas por debajo y aquellos apenas por encima del puntaje de corte, pero debe proporcionarse evidencia, cuando sea viable, de la relación entre el desempeño en la prueba y en el criterio en un intervalo de puntajes que incluya o aborde el puntaje de corte.

6. ADMINISTRACIÓN, CALIFICACIÓN, PRESENTACIÓN DE REPORTES E INTERPRETACIÓN DE PRUEBAS

ANTECEDENTES

La utilidad e interpretabilidad de los puntajes requieren que la prueba se administre y califique de acuerdo con las instrucciones del desarrollador de la prueba. Cuando las instrucciones, las condiciones de la prueba y la calificación siguen los mismos procedimientos detallados para todos los examinandos, se considera que la prueba ha sido estandarizada. Sin esta estandarización, se reduciría la precisión y comparabilidad de las interpretaciones de los puntajes. En pruebas diseñadas para evaluar los conocimientos, habilidades, capacidades u otras características personales, la estandarización permite garantizar que todos los examinandos tengan las mismas oportunidades de demostrar sus competencias. Mantener la seguridad de la prueba también ayuda a garantizar que nadie tenga una ventaja indebida. La importancia de la adherencia a la estandarización apropiada de los procedimientos de administración aumenta con los riesgos de la prueba.

Sin embargo, a veces se producen situaciones en las que pueden ser recomendables o legalmente obligatorias variaciones respecto de los procedimientos estandarizados. Por ejemplo, es posible que individuos con discapacidades y personas de diferentes contextos lingüísticos, edades o familiaridad con las pruebas necesiten modos no estándar de administración, o una orientación más completa para el proceso de la prueba, de manera que todos los examinandos puedan tener una oportunidad expedita para demostrar su situación respecto del constructo que se mide. Es posible que diferentes modos de presentación de la prueba, de sus instrucciones o de las respuestas, resulten idóneos para determinados individuos, por ejemplo, personas con algún tipo de discapacidad o personas con destrezas limitadas en el idioma de la prueba, a fin de proporcionar un acceso adecuado y reducir la varianza irrelevante de constructo (véase el capítulo 3, "Imparcialidad en las pruebas"). En situaciones de pruebas clínicas o neuropsicológicas, puede ser necesaria flexibilidad en la administración, en función de la capacidad del individuo para entender y responder a los ítems de la prueba o a las tareas, y/o al constructo cuya medición se requiere. Algunas situaciones y/o el constructo (p. ej., las pruebas de deterioro de la memoria de un examinando con demencia que se encuentra hospitalizado) pueden requerir que la evaluación se abrevie o altere. Los programas de pruebas a gran escala suelen establecer procedimientos específicos para la consideración y autorización de adecuaciones y otras variaciones respecto de los procedimientos estandarizados. Por lo general, estas adecuaciones están relativamente estandarizadas; en ocasiones, se pueden indicar algunas alternativas distintas a las adecuaciones previstas y especificadas por el desarrollador de la prueba. Se debe tener especial cuidado para evitar el tratamiento sesgado y la discriminación. Aunque las variaciones se pueden realizar con la intención de mantener la comparabilidad de los puntajes, con frecuencia no es posible determinar el grado en que esto es posible. Se podría poner en riesgo la comparabilidad de los puntajes y, por consiguiente, la prueba no mediría el mismo constructo para todos los examinandos.

Las pruebas y las evaluaciones difieren en el grado de estandarización. En muchos casos, diferentes examinandos no reciben el mismo formulario de prueba, pero reciben formularios equivalentes que producen puntajes comparables, o formularios alternativos donde los puntajes se adaptan para hacerlos comparables. Algunas evaluaciones permiten a los examinandos elegir las tareas que deben realizar o las partes de sus trabajos que van a ser evaluadas. En estas situaciones se puede mantener la estandarización especificando

las condiciones de la elección y el criterio para la evaluación de los productos. Cuando una evaluación permite un determinado tipo de colaboración entre los examinandos o entre el examinando y el administrador de la prueba, se deben especificar los límites de esa colaboración. En algunas evaluaciones cabe esperar que los administradores de la prueba adapten las instrucciones para asegurarse de que todos los examinandos entienden lo que se espera de ellos. En todos estos casos, el objetivo sigue siendo el mismo: proporcionar una medición precisa, imparcial y comparable para todos. El grado de estandarización viene dictado por ese objetivo y por el uso previsto de los puntajes de la prueba.

Las instrucciones estandarizadas ayudan a garantizar que todos los examinandos tengan una compresión común de la mecánica de la evaluación. Por lo general, las instrucciones informan a los examinandos sobre cómo presentar sus respuestas, qué clase de ayuda pueden razonablemente obtener si no comprenden la pregunta o tarea, cómo pueden corregir las respuestas accidentales y la naturaleza de las restricciones temporales si las hubiera. En ocasiones, se proporciona orientación general sobre la omisión de respuestas de ítems. Muchas pruebas, incluyendo las pruebas administradas por computadora, requieren equipos o software especiales. En tales casos, suelen presentarse ejercicios de práctica e instrucción, de manera que los examinandos entiendan el modo de funcionamiento del equipo o software. El principio de estandarización incluye la orientación de los examinandos en los materiales y adecuaciones con los que podrían no estar familiarizados. Algunos equipos se pueden facilitar en la ubicación de la prueba, por ejemplo, herramientas comerciales o sistemas de software. A menudo resulta apropiado que los examinandos tengan la oportunidad de practicar con el equipo, a menos que el constructo en evaluación sea la capacidad de usar el equipo.

En ocasiones, las pruebas se administran a través de medios tecnológicos, ingresando las respuestas mediante el teclado, ratón, entrada de voz u otros dispositivos. Cada vez más examinandos están acostumbrados al uso de computadoras. Es posible que aquellos no familiarizados con el uso de computadoras necesiten capacitación para reducir la varianza irrelevante de constructo. Incluso aquellos examinandos familiarizados con computadoras podrían requerir una breve explicación y práctica para gestionar detalles específicos de la prueba, por ejemplo, la interfaz de la prueba. Se producen problemas especiales en la gestión del entorno de la prueba para reducir la varianza irrelevante de constructo, por ejemplo, evitar reflejos de luz en los monitores que interfieran con la legibilidad de la pantalla, o mantener un entorno tranquilo cuando los examinandos empiezan o terminan la prueba en momentos diferentes con respecto a sus vecinos. Quienes administren pruebas basadas en computadora deben recibir capacitación para resolver los problemas de hardware, software o administración de la prueba. Las pruebas administradas por computadora en aplicaciones basadas en Web pueden requerir apoyos adicionales para mantener entornos estandarizados.

Los procedimientos de calificación estandarizados ayudan a garantizar una calificación y presentación de reportes coherentes, fundamentales en cualquier circunstancia. Cuando la calificación se realiza por máquina, se debe establecer y supervisar la precisión de la máquina, incluyendo cualquier programa o algoritmo de calificación. Cuando la calificación de respuestas complejas la realizan evaluadores humanos o motores automáticos de calificación, se requiere una cuidadosa capacitación. Normalmente, la capacitación requiere que evaluadores humanos expertos proporcionen una muestra de respuestas que abarque el rango de posibles clasificaciones o puntajes. Dentro de los rangos de puntajes, los instructores también deben proporcionar muestras que ejemplifiquen la variedad de respuestas que se traducirán en clasificaciones o niveles de puntaje. La supervisión periódica ayuda a garantizar que todos los desempeños de las pruebas se califiquen de acuerdo con los mismos criterios estandarizados y que los evaluadores no aplique los criterios de manera diferente a medida que avanzan por las respuestas entregadas.

En sí mismos, los puntajes no se interpretan fácilmente sin información adicional como, por

ejemplo, normas o estándares, indicaciones de error de medida y descripciones del contenido de la prueba. Así como una temperatura de 10 grados Celsius en enero es cálida en Minnesota y fría en Florida, un puntaje de prueba de 50 no resulta relevante sin contexto. Se debe facilitar material interpretativo que sea fácilmente comprensible para quienes reciben el reporte. A menudo, el usuario de la prueba proporciona una interpretación de los resultados para el examinando, sugiriendo las limitaciones de los resultados y la relación con otros datos de cualquier puntaje reportado. Los puntajes de algunas pruebas no están diseñados para ser presentados a los examinandos; solo se prevé la comunicación de interpretaciones amplias o de clasificaciones dicotómicas, del tipo "aprobado/reprobado".

En ocasiones, las interpretaciones de los resultados de una prueba se preparan mediante sistemas computarizados. Generalmente, tales interpretaciones se basan en una combinación de datos empíricos, juicio experto y experiencia, y requieren validación. En algunas aplicaciones profesionales de pruebas individualizadas, las interpretaciones preparadas por computadora se comunican mediante un profesional, quien puede modificar la interpretación inicial para adaptarla a circunstancias especiales. Se debe tener especial cuidado para que las interpretaciones de la prueba suministradas por métodos no algorítmicos guarden la coherencia apropiada. Los reportes generados automáticamente no son un sustituto del juicio clínico de un evaluador profesional que haya trabajado directamente con el examinando, o de la integración de información adicional, incluyendo, entre otros, otros resultados de la prueba, entrevistas, registros existentes y observaciones conductuales.

En algunas evaluaciones a gran escala, el objetivo principal de la evaluación no es el examinando individual sino una unidad mayor, por ejemplo, un distrito escolar o una planta industrial. A menudo, se proporciona diferentes conjuntos de ítems a diferentes examinandos, siguiendo un plan de muestreo de matriz cuidadosamente equilibrado, con el fin de ampliar la gama de información que se puede obtener en un periodo de tiempo razonable. Los resultados adquieren significado cuando se lleva a cabo la agregación de muchos individuos que responden a diferentes muestras de ítems. Es posible que tales evaluaciones no aporten suficiente información que respalde puntuaciones mínimamente válidas o confiables para los individuos, ya que cada individuo puede realizar solo una parte de la prueba, mientras que en la agregación, los resultados de la evaluación podrían ser válidos y de una confiabilidad aceptable para interpretaciones sobre el desempeño de la unidad de mayor tamaño.

En el capítulo 4, "Diseño y desarrollo de pruebas", se tratan algunos temas adicionales sobre administración y calificación.

Los usuarios de la prueba, y aquellos que reciben los materiales y puntajes de las pruebas e información complementaria (como pueden ser los datos de identificación personal de los examinandos), son responsables de mantener correctamente la seguridad y confidencialidad de esta información.

ESTÁNDARES PARA LA ADMINISTRACIÓN, CALIFICACIÓN, PRESENTACIÓN DE REPORTES E INTERPRETACIÓN DE PRUEBAS

Los estándares de este capítulo empiezan con un estándar general (con el número 6.0), diseñado para comunicar el propósito central o el enfoque principal del capítulo. El estándar general también se puede ver como el principio rector del capítulo y se aplica a todas las pruebas y a todos los usuarios de la prueba. Todos los estándares posteriores se han dividido en tres unidades temáticas, etiquetadas de la siguiente manera:

1. Administración de la prueba
2. Calificación de la prueba
3. Presentación de informes e interpretación

Estándar 6.0

Para respaldar las interpretaciones útiles de los resultados de calificación, los instrumentos de evaluación deben haber establecido los procedimientos para la administración, calificación, presentación de informes e interpretación de las pruebas. Los responsables de administrar, calificar, presentar informes e interpretar deben tener la capacitación y el apoyo suficientes para seguir los procedimientos establecidos. Se debe supervisar la adherencia a los procedimientos establecidos y cualquier error material deberá documentarse y, si es posible, corregirse.

Comentario: A fin de respaldar la validez de las interpretaciones de los puntajes, la administración debe seguir todos y cada uno de los procedimientos establecidos, y será necesario supervisar la conformidad con tales procedimientos.

Unidad 1. Administración de la prueba

Estándar 6.1

Los administradores deben seguir cuidadosamente los procedimientos estandarizados de administración y calificación especificados por el desarrollador de la prueba, así como las instrucciones del usuario de la prueba.

Comentario: Los responsables de los programas de pruebas deben proporcionar la capacitación, documentación y supervisión apropiadas, de manera que los individuos que administren o califiquen las pruebas sean competentes en los procedimientos apropiados de administración o calificación de las pruebas y entiendan la importancia de adherirse a las instrucciones facilitadas por el desarrollador. Los programas de pruebas a gran escala deben especificar los procedimientos estandarizados admitidos para la determinación de las adecuaciones u otras variaciones aceptables en la administración. La capacitación deberá habilitar a los administradores para realizar los ajustes apropiados si se requiere una adecuación o modificación que no esté incluida en los procedimientos estandarizados.

Se deben observar estrictamente las especificaciones relacionadas con las instrucciones para los examinandos, los límites de tiempo, la forma de presentación o respuesta de ítems, y los materiales o equipos de la prueba. En general, se deben seguir los mismos procedimientos que se utilizaron para la obtención de los datos para el escalamiento y normalización de los puntajes de la prueba. Algunos programas no se escalan ni establecen normas, por ejemplo, las evaluaciones de porfolio y la mayoría de las evaluaciones académicas para estudiantes con discapacidades cognitivas severas. No obstante, habitualmente estos programas tienen procedimientos estandarizados específicos para la administración y calificación cuando establecen estándares de rendimiento. Un examinando con una discapacidad puede necesitar variaciones que proporcionen acceso sin cambiar el constructo que se mide. Otras circunstancias especiales pueden requerir flexibilidad en la administración, por ejemplo, apoyo lingüístico para facilitar el acceso bajo determinadas condiciones, o evaluaciones clínicas o neuropsicológicas, además de procedimientos relacionados con las adecuaciones. Los juicios sobre la idoneidad

de los ajustes deben estar matizados por la consideración de que las desviaciones respecto de los procedimientos estándar pueden poner en peligro la validez o complicar la comparabilidad de las interpretaciones de los puntajes. Estos juicios se deben llevar a cabo por profesionales cualificados y ser coherentes con las directrices proporcionadas por el usuario o desarrollador de la prueba.

Las políticas relacionadas con las contrapruebas deben ser establecidas por el usuario o desarrollador de la prueba. El usuario o administrador de la prueba debe seguir la política establecida. El usuario de la prueba debe comunicar claramente estas políticas de contrapruebas, como parte de las condiciones para la administración estandarizada de una prueba. Las contrapruebas tienen como finalidad reducir las probabilidades de que se clasifique erróneamente a una persona por no cumplir un determinado estándar. Por ejemplo, algunos programas de pruebas especifican que una persona debe repetir la prueba; otros ofrecen varias oportunidades de hacer una prueba, por ejemplo, después de aprobar una prueba necesaria para la graduación secundaria o para la obtención de autorizaciones.

Los desarrolladores de la prueba deben especificar las condiciones estandarizadas de administración que respalden los usos previstos de las interpretaciones de puntajes. Los usuarios de la prueba deben tener presentes las implicaciones de condiciones de administración con menor control. Los usuarios de la prueba tienen la responsabilidad de ofrecer apoyo técnico o de otro tipo para garantizar que las administraciones cumplan estas condiciones en el mayor grado posible. Sin embargo, la tecnología e Internet han hecho posible administrar pruebas en muchos contextos, incluyendo contextos donde las condiciones de administración no se controlan o supervisan de manera estricta. Quienes permiten deficiencias de estandarización son responsables de proporcionar la evidencia de que esas deficiencias no han afectado al desempeño del examinando o a la calidad o comparabilidad de los puntajes obtenidos. La documentación completa incluirá los informes sobre el grado de incumplimiento de las condiciones estandarizadas de administración.

Características como los límites de tiempo, la elección de tipos de ítems y formatos de respuesta, las interfaces complejas y las instrucciones que potencialmente introducen varianza irrelevante de constructo, se deben analizar en términos de propósito de la prueba y los constructos sometidos a medición. Si es factible, se deben llevar a cabo investigaciones empíricas y de usabilidad apropiadas para documentar (y de ser posible, minimizar) el impacto de las fuentes o condiciones que contribuyen a la variabilidad irrelevante de constructo.

Estándar 6.2

Cuando se han establecido procedimientos formales para la solicitud y obtención de adecuaciones, se debe informar a los examinandos sobre estos procedimientos con antelación a la prueba.

Comentario: Cuando los programas de pruebas han establecido procedimientos y criterios para identificar y facilitar adecuaciones para los examinandos, los procedimientos y criterios se deben seguir y documentar cuidadosamente. De forma óptima, estos procedimientos incluyen la consideración de los casos donde, además de las adecuaciones previstas y especificadas por el desarrollador de la prueba, puede resultar apropiada una alternativa. Los examinandos deben recibir información sobre cualquier adecuación que puedan tener a disposición, y sobre los procesos y requisitos (si existe alguno) para obtener las adecuaciones necesarias. De forma similar, en contextos educativos, el personal de la escuela y los padres o tutores legales deben recibir información de los requisitos (si existe alguno) para obtener las adecuaciones necesarias para los estudiantes que realizan la prueba.

Estándar 6.3

Los cambios o alteraciones en los procedimientos estandarizados de administración o calificación de pruebas se deben documentar y presentar al usuario de la prueba.

Comentario: La información sobre la naturaleza de los cambios en los procedimientos estandarizados de administración o calificación se debe

mantener en archivos de datos seguros, de manera que puedan ser tenidos en cuenta por los estudios de investigación o las revisiones de casos que se basen en los registros de la prueba. Esto incluye no solo las adecuaciones o modificaciones para examinandos específicos, sino también las alteraciones en el entorno de la prueba que puedan afectar a todos los examinandos de la sesión. Es posible que un investigador quiera usar únicamente los registros basados en la administración estandarizada. En otros casos, los estudios de investigación podrían ser dependientes de esta información para la formación de grupos de examinandos. Los usuarios o promotores de la prueba deben establecer políticas que especifiquen quién protege los archivos de datos, quién puede tener acceso a esos archivos y, si es necesario, cómo mantener la confidencialidad de los entrevistados, por ejemplo, mediante la supresión de cualquier dato de identificación. Si se proporciona o no la información sobre desviaciones respecto de los procedimientos estándar a usuarios de los datos de la prueba depende de consideraciones como, por ejemplo, si los usuarios son funcionarios de admisiones o usuarios de reportes psicológicos individualizados en centros clínicos. Si se llevan a cabo tales reportes, puede resultar apropiado incluir documentación clara sobre cualquier desviación respecto de los procedimientos estándar de administración, las deliberaciones sobre los efectos de estas variaciones administrativas en los resultados, y tal vez determinadas precauciones. Por ejemplo, es posible que los usuarios de la prueba necesitan disponer de información sobre la comparabilidad de los puntajes cuando se establecen modificaciones (véase el capítulo 3, "Imparcialidad en las pruebas" y el capítulo 9, "Derechos y responsabilidades de los usuarios de la prueba"). Si una desviación o cambio en un procedimiento estandarizado de administración de pruebas se considera lo suficientemente importante como para afectar negativamente a la validez de la interpretación de los puntajes, se deben tomar las medidas correspondientes (por ejemplo, la anulación de los puntajes) o facilitar oportunidades para una nueva administración bajo las circunstancias apropiadas. Los entornos de pruebas que no se supervisan (p. ej., en condiciones temporales o en Internet) deben cumplir estas condiciones estandarizadas de administración; de otro modo, los reportes sobre calificaciones deben indicar que no se garantizaron las condiciones estandarizadas.

Estándar 6.4

El entorno de la prueba debe disponer de un grado razonable de comodidad, con mínimas distracciones para evitar la varianza irrelevante de constructo.

Comentario: Los desarrolladores de la prueba deben proporcionar información relacionada con las condiciones y el entorno previstos de administración. El ruido, las interrupciones en el área de la prueba, temperaturas extremas, iluminación insuficiente, un espacio de trabajo inadecuado, materiales ilegibles y computadoras averiadas son algunas de las condiciones que se deben evitar en situaciones de evaluación, a menos que la medida del constructo requiera de tales condiciones. La ubicación donde se realiza la prueba debe ser fácilmente accesible. Las administraciones basadas en tecnología deben evitar distracciones como, por ejemplo, fallos de equipos o de la conexión a Internet, o grandes variaciones en el tiempo que se dedica a presentar los ítems o en el modo de respuesta de la prueba. Las sesiones de las pruebas se deben supervisar donde resulte apropiado para solventar rápidamente las necesidades de los examinandos y mantener correctamente los procedimientos administrativos. En general, las condiciones de las pruebas deben ser equivalentes a las que prevalecían cuando se obtuvieron las normas u otros datos interpretativos.

Estándar 6.5

Se debe facilitar a los examinandos las instrucciones y la práctica apropiadas, y cualquier otro apoyo necesario para reducir la varianza irrelevante de constructo.

Comentario: Las instrucciones a los examinandos deben indicar con claridad cómo formular las respuestas, excepto cuando esto obstaculice

la medida del constructo previsto (p. ej., cuando se evalúa la actitud espontánea de un individuo a la situación de la prueba). También se deben proporcionar instrucciones sobre el uso de cualquier equipo o software con el que los examinandos no estuvieran familiarizados, a menos que la adaptación a tales herramientas sea parte de la evaluación. Es posible que algunos examinandos no estén familiarizados con las funciones o interfaces de las pruebas administradas por computadora y que necesiten cierto aprendizaje sobre el inicio de sesión, la navegación o el acceso a herramientas. Cuando se utilizan equipos, se deben proporcionar oportunidades de práctica, a menos que la evaluación sea el propio uso del equipo. Es posible que algunos examinandos necesiten practicar las respuestas con los medios específicos que demanda la prueba, por ejemplo, rellenar recuadros de respuestas múltiples o interactuar con una simulación multimedia. Donde sea posible, se debe supervisar la práctica de las respuestas para confirmar que el examinando responde de forma aceptable. Si un examinando no puede usar el equipo o formular las respuestas, puede ser conveniente considerar modos alternativos de prueba. Además, se debe informar con claridad a los examinandos sobre la forma en que su velocidad de trabajo puede afectar a los puntajes y sobre cómo se tratarán algunas respuestas en el puntaje (por ejemplo, no responder, hacer suposiciones o responder de forma incorrecta), a menos que tales instrucciones perjudiquen al constructo en evaluación.

Estándar 6.6

Se deben tomar las medidas razonables para garantizar la integridad de los puntajes de las pruebas, eliminando las oportunidades para que los examinandos logren puntajes mediante medios engañosos o fraudulentos.

Comentario: En los programas de pruebas donde se considera que los resultados pueden tener importantes consecuencias, se debe mantener la integridad de los puntajes a través de medidas activas que eviten, detecten y corrijan los puntajes obtenidos por medios engañosos o fraudulentos. Tales medidas pueden incluir, cuando sea factible y apropiado, la estipulación de requisitos de identificación, el diseño de gráficos de asientos, la asignación de asientos a los examinandos, la necesidad de espacio apropiado entre asientos y la supervisión continua del proceso de la prueba. Los desarrolladores deben diseñar los materiales y procedimientos de la prueba para minimizar la posibilidad de trampas. Un cambio local en la fecha u hora de la prueba puede ofrecer una oportunidad de engaño. Se debe capacitar a los instructores sobre las precauciones apropiadas para evitar y detectar oportunidades de engaño, por ejemplo, las oportunidades que ofrece la tecnología para que un examinando se comunique con un cómplice fuera del área de prueba, o sobre tecnología que permite a un examinando copiar información de la prueba para su divulgación posterior. Los administradores deben seguir las políticas establecidas para tratar con cualquier caso de irregularidades en las pruebas. En general, se deben tomar medidas para minimizar la posibilidad de vulnerabilidades en la seguridad de las pruebas y para detectar cualquier punto vulnerable. En las evaluaciones de productos de trabajo (p. ej., porfolios) se deben tomar medidas para garantizar que el producto represente el propio trabajo del examinando y que la cantidad y la clase de asistencia proporcionada es coherente con la finalidad de la evaluación. Puede ser útil documentación complementaria, como la fecha en que se realizó el trabajo. Los programas de pruebas pueden usar tecnologías durante la calificación para detectar posibles irregularidades (p. ej., análisis computarizado de patrones de borraduras, patrones de respuestas similares para varios examinandos, plagio de fuentes online o cambios inusuales en los parámetros de ítems). Los usuarios de tales tecnologías son responsables de su precisión y aplicación apropiada. Es posible que los desarrolladores y usuarios de las pruebas tengan que supervisar la divulgación de los ítems de la prueba en Internet o desde otras fuentes. Los programas de pruebas con consecuencias de alto riesgo deben tener políticas y procedimientos definidos para detectar y procesar potenciales

irregularidades (incluyendo un proceso mediante el cual una persona acusada de irregularidades pueda optar por o presentar una apelación) y para anular la validez de los puntajes y proporcionar oportunidades de repetición de pruebas.

Estándar 6.7

Los usuarios de la prueba tienen la responsabilidad de proteger la seguridad de los materiales de la prueba en todo momento.

Comentario: Quienes tiene los materiales de la prueba bajo su control deben, con la consideración debida a los requisitos éticos y legales, tomar todas las medidas necesarias para asegurarse de que únicamente las personas con necesidades y cualificaciones legítimas para el acceso a estos materiales puedan obtener dicho acceso antes de la administración de la prueba, y también después, si se prevé reutilizar alguna parte de la prueba en un momento posterior. Las preocupaciones relativas al acceso inapropiado a los materiales de la prueba incluyen la divulgación inadecuada del contenido, la adulteración de las respuestas y resultados de la prueba, y la protección de los derechos de privacidad de los examinandos. Los usuarios de la prueba deben compaginar la seguridad de la prueba con los derechos de todos los examinandos y usuarios de la prueba. Cuando documentos sensibles de la prueba se encuentren en procesos de litigio en los tribunales o formen parte de problemas administrativos, es importante identificar desde el principio los temas de privacidad y seguridad y las protecciones necesarias. Las partes se deben asegurar que la divulgación o exposición de tales documentos (incluyendo secciones específicas de esos documentos que pueden requerir redacción) a terceros, expertos y a los propios tribunales/organismos es coherente con condiciones (reflejadas a menudo en órdenes de protección) que no den como resultado la divulgación inapropiada y no pongan en riesgo la divulgación justificada más allá del contexto específico en el cual se ha producido el problema. Bajo ciertas circunstancias, cuando los documentos sensibles de la prueba se encuentren cuestionados, puede resultar apropiado contratar a un tercero independiente, a través de un procedimiento seguro estrechamente supervisado, para llevar a cabo una revisión de los materiales relevantes en lugar de poner las pruebas, manuales o las respuestas de los examinandos en el registro público. Quienes tengan información confidencial relacionada con las pruebas, como la información de registro, la programación y los pagos, tienen una responsabilidad similar de proteger esa información. Quienes dispongan de los materiales bajo su control deben usar y divulgar esa información únicamente de acuerdo con las leyes de privacidad vigentes.

Unidad 2. Calificación de la prueba

Estándar 6.8

Los responsables de la calificación de las pruebas deben establecer protocolos de calificación. La calificación de pruebas que involucra juicio humano debe incluir rúbricas, procedimientos y criterios de calificación. Cuando la calificación de respuestas complejas se lleva a cabo por computadora, se debe documentar la precisión de los algoritmos y procesos.

Comentario: Se debe establecer un protocolo de calificación, el cual puede ser tan sencillo como una clave de respuestas para preguntas de opciones múltiples. Para respuestas construidas, se puede suministrar a los evaluadores (humanos o programas computarizados) respuestas alternativas aceptables, listados de rúbricas de calificación, así como criterios generales. Una práctica común de los desarrolladores de pruebas es proporcionar materiales de capacitación sobre calificación y ejemplos de respuestas de examinandos de cada nivel de puntaje. Los materiales de calificación se deben revisar periódicamente cuando se usen las pruebas o ítems a lo largo de un periodo de tiempo.

Estándar 6.9

Los responsables de la calificación de pruebas deben establecer y documentar los procesos y

criterios de control de calidad. Se debe proporcionar una capacitación adecuada. La calidad de la calificación se debe supervisar y documentar. Cualquier fuente sistemática de errores de calificación se debe documentar y corregir.

Comentario: Se deben establecer criterios para una calidad de calificación aceptable. Se deben establecer procedimientos para calibrar a los evaluadores (humanos o máquinas) antes de la calificación operativa, y para supervisar su coherencia en la calificación según los estándares establecidos durante la calificación operativa. Cuando la calificación se divide entre varios calificadores, los procedimientos para supervisar la precisión y confiabilidad de los evaluadores pueden ser útiles como procedimiento de control de calidad. Con frecuencia, la coherencia en la aplicación de los criterios de calificación se comprueba mediante la recalificación independiente de respuestas aleatoriamente seleccionadas. Las comprobaciones periódicas de las propiedades estadísticas (p. ej. las medias, las desviaciones estándar, el porcentaje de concordancia con puntajes cuya precisión se ha determinado anteriormente) de los puntajes asignados por evaluadores individuales durante una sesión de calificación pueden proporcionar información a los evaluadores y ayudarles a mantener los estándares de calificación. Además, el análisis podría controlar los posibles efectos sobre la precisión de la calificación de variables como el evaluador, la tarea, el tiempo o el día de calificación, el instructor de calificación, el emparejamiento de evaluadores, etc., para informar las acciones correctivas o preventivas apropiadas. Cuando se usan los mismos ítems en varias administraciones, los programas deben tener establecidos procedimientos para supervisar la coherencia de la calificación entre las administraciones (p. ej., comparabilidad interanual). Una manera de revisar la coherencia a lo largo del tiempo es recalificar algunas respuestas de administraciones anteriores. Una calificación imprecisa o incoherente puede requerir reentrenamiento, recalificación, la revocación de algunos evaluadores o el reexamen de las rúbricas o programas de calificación. Se deben corregir los errores de calificación sistemáticos y esto puede conllevar la recalificación de respuestas previamente puntuadas, así como la corrección de la fuente del error. Se deben examinar los errores administrativos o mecánicos. Los errores de calificación se deben minimizar y, cuando se encuentren, se deben tomar medidas rápidamente para minimizar su recurrencia.

Habitualmente, los responsables de la calificación documentarán los procedimientos seguidos para la calificación, los procedimientos seguidos para el control de calidad de esa calificación, los resultados del control de calidad y cualquier circunstancia inusual. En función del usuario de la prueba, se puede facilitar esa documentación periódicamente o por peticiones razonables. Las aplicaciones de calificación computarizadas de texto, voz u otras respuestas construidas deben proporcionar documentación similar de la precisión y confiabilidad, incluyendo comparaciones con la calificación humana.

Cuando la calificación se hace localmente y requiere del juicio de un evaluador, el usuario de la prueba es responsable de facilitar capacitación e instrucción adecuadas a los evaluadores y de examinar la concordancia y precisión de los evaluadores. Cuando se posible, se debe documentar el nivel previsto de concordancia y precisión de un evaluador.

Unidad 3. Presentación de informes e interpretación

Estándar 6.10

Cuando se divulgue la información de puntajes de la prueba, los responsables de los programas de calificación deben ofrecer interpretaciones apropiadas a la audiencia. Las interpretaciones deben describir, en lenguaje sencillo, el ámbito de la prueba, lo que representan los puntajes, la precisión/confiabilidad de los puntajes y su uso previsto.

Comentario: Los usuarios de la prueba deben consultar el material interpretativo preparado por el desarrollador de la prueba y deben revisar o complementar el material según sea necesario para

presentar los resultados individuales de manera precisa y clara a la audiencia objetivo, que puede incluir clientes, representantes legales, medios de comunicación, fuentes de referencia, examinandos, padres o profesores. Los reportes y comentarios deben estar diseñados para respaldar las interpretaciones válidas y el uso, y para minimizar las consecuencias negativas potenciales. La precisión del puntaje podría representarse mediante márgenes de error o rangos probables de puntajes que muestren el error estándar de medida. Los reportes deben incluir las deliberaciones sobre las variaciones administrativas u observaciones de conducta en contextos clínicos que puedan afectar a los resultados e interpretaciones. Los usuarios de la prueba deben evitar las interpretaciones erróneas o el uso indebido de la información de calificación. Aunque los usuarios de la prueba son los principales responsables de evitar las interpretaciones erróneas o el uso indebido, los materiales interpretativos preparados por el desarrollador o editor de la prueba pueden resolver los usos indebidos o errores de interpretación comunes. Para conseguir esto, los desarrolladores de reportes y materiales interpretativos pueden llevar a cabo investigaciones para verificar que los reportes y materiales pueden interpretarse como se pretende (p. ej., grupos temáticos con usuarios finales representativos de los reportes). El desarrollador debe informar a los usuarios de la prueba sobre los cambios en la prueba a lo largo del tiempo que puedan afectar a la interpretación del puntaje, por ejemplo, los cambios en las normas, los modelos de contenido de la prueba o los significados de los puntajes de escala.

Estándar 6.11

Cuando se reportan interpretaciones de protocolos de respuestas de pruebas o de desempeño en pruebas generados automáticamente, las fuentes, justificaciones y bases empíricas de estas interpretaciones deben estar disponibles y se deben describir sus limitaciones.

Comentario: En ocasiones, las interpretaciones de resultados de pruebas se generan automáticamente, ya sea por programas computarizados que funcionan con calificaciones computarizadas o mediante materiales preparados manualmente. Es posible que las interpretaciones generadas automáticamente no tomen en consideración el contexto de circunstancias de los individuos. Las interpretaciones generadas automáticamente se deben usar con cuidado en contextos de diagnóstico, ya que es posible que no tomen en cuenta otra información relevante sobre el examinando individual que proporcione contexto a los resultados, por ejemplo, la edad, el género, el nivel educativo, el empleo anterior, la situación psicológica, el estado de salud, los antecedentes psicológicos y la sintomatología. De forma similar, los desarrolladores y usuarios de las pruebas con interpretaciones generadas automáticamente del desempeño académico y de prescripciones complementarias de seguimiento instruccional deben reportar las bases y limitaciones de las interpretaciones. Las interpretaciones de las pruebas no deben implicar la existencia de evidencia empírica de una relación entre los resultados de pruebas específicas, intervenciones prescritas y conclusiones deseadas, a menos que la evidencia empírica esté disponible para poblaciones similares a las representativas del examinando.

Estándar 6.12

Cuando se obtiene información de nivel de grupo mediante la agregación de los resultados de pruebas parciales realizadas por individuos, se debe reportar la evidencia de validación y confiabilidad/precisión del nivel de agregación en el cual se presentan los resultados. No se deben reportar los puntajes por individuos sin la evidencia apropiada que respalde las interpretaciones para los usos previstos.

Comentario: Con frecuencia, las interpretaciones a gran escala logran eficiencia mediante un "muestreo de matriz" del contenido del dominio, para el cual se hacen diferentes preguntas a diferentes examinandos. De este modo, la evaluación requiere menos tiempo para cada examinando, en tanto que la agregación de resultados individuales confiere

cobertura de dominio que puede resultar adecuada para interpretaciones relevantes en un nivel de grupo o programa, por ejemplo, para escuelas o niveles de grado en una localidad o en áreas temáticas específicas. No obstante, debido a que se solo administra una prueba incompleta al individuo, los puntajes individuales tendrían un significado limitado, si lo tienen.

Estándar 6.13

Cuando se encuentra un error material en los puntajes de las pruebas u otra información importante publicada por una organización de evaluación u otra institución, se debe distribuir esta información y un reporte de calificación corregida tan pronto como sea posible a todos los destinatarios conocidos quienes, de lo contrario, podrían usar los puntajes erróneos como base para la toma de decisiones. El reporte corregido se debe etiquetar como tal. Se deben documentar las acciones realizadas para corregir los reportes. Los motivos del reporte de calificación corregida deben presentarse claramente a los destinatarios del reporte.

Comentario: Un error material es un error que puede cambiar la interpretación del puntaje de la prueba y suponer una diferencia importante. Un ejemplo es un puntaje de prueba erróneo (p. ej., calculado de manera incorrecta u obtenido de forma fraudulenta) que afectaría a una decisión importante sobre el examinando, por ejemplo, la decisión de otorgar una acreditación o la concesión de un título de secundaria. Se excluirían los errores tipográficos. La pertinencia temporal es crucial en las decisiones que se toman poco después de recibir los puntajes de las pruebas. Cuando los resultados de las pruebas se han usado para informar decisiones de alto riesgo, es posible que los usuarios de la prueba tengan que llevar a cabo acciones correctivas para rectificar las circunstancias afectadas por los puntajes erróneos, además de publicar los reportes corregidos. En determinados trabajos u otros contextos, es posible que no sea factible o no se puedan llevar a cabo acciones correctivas y de presentación de reportes. Los usuarios de las pruebas deben desarrollar una política de gestión de errores materiales en los puntajes de pruebas y deben documentar las acciones realizadas en el caso de errores materiales reales o supuestos.

Estándar 6.14

Las organizaciones que mantienen información de puntajes con identificación personal deben desarrollar un conjunto claro de directrices sobre la duración de la conservación de los registros de los individuos y sobre la disponibilidad y uso a lo largo del tiempo de tales datos para investigación u otros fines. La política debe estar documentada y disponible para el examinando. Los usuarios de la prueba deben mantener una seguridad de datos apropiada, que debe incluir protecciones administrativas, técnicas y físicas.

Comentario: En algunos casos, los puntajes de las pruebas quedan obsoletos a lo largo del tiempo y ya no reflejan el estado actual del examinando. En general, los puntajes desactualizados no se deben usar ni estar disponibles, excepto para fines de investigación. En otros casos, los puntajes obtenidos en años anteriores pueden ser útiles, como en las evaluaciones longitudinales o en el seguimiento del deterioro de una función o cognición. El factor clave es el uso válido de la información. Las organizaciones e individuos que mantienen información de puntajes con identificación personal deben tener en cuenta y cumplir los requisitos legales y profesionales. Es posible que se solicite a las organizaciones e individuos que mantienen puntajes de pruebas sobre individuos que proporcionen los datos a investigadores u otros usuarios terceros. Cuando la divulgación de los datos se considera apropiada y no esté prohibida por estatutos o normativas, el usuario de la prueba debe proteger la confidencialidad de los examinandos a través de políticas adecuadas, por ejemplo, suprimiendo cualquier dato de identificación o mediante acuerdos de no revelación y confidencialidad de los datos. Las organizaciones e individuos que mantienen o usan información confidencial sobre los examinandos o sus puntajes deben tener e implementar una política apropiada

para mantener la seguridad e integridad de los datos, incluyendo la protección de modificaciones accidentales o deliberadas, así como la prevención ante pérdidas o destrucción no autorizada. En algunos casos, es posible que las organizaciones deban obtener el consentimiento de los examinandos para usar o revelar los registros. Se deben establecer protocolos apropiados y una seguridad adecuada cuando los datos confidenciales de una prueba forman parte de un registro de mayor tamaño (p. ej., registros médicos electrónicos) o cuando se combinan en un almacén de datos. Si los registros se van a comunicar para evaluaciones clínicas o forenses, se debe tener cuidado en comunicarlos a las personas debidamente autorizadas, con las autorizaciones de publicación firmadas por el examinando o la autoridad legal pertinente.

Estándar 6.15

Cuando se retienen datos individuales de las pruebas, se debe conservar de alguna forma tanto el protocolo de la prueba como cualquier reporte escrito.

Comentario: Es posible que el protocolo sea necesario para responder una potencial recusación de un examinando o para facilitar la interpretación en un momento posterior. Normalmente, el protocolo debería adjuntar los materiales y puntajes de la prueba. La retención de más registros detallados de respuestas dependería de las circunstancias y debe estar incluida en una política de retención. El mantenimiento de los registros debe estar sujeto a los requisitos legales y profesionales. La política de publicación de cualquier información de la prueba para fines diferentes a la investigación se trata en el capítulo 9, "Derechos y responsabilidades de los usuarios de la prueba".

Estándar 6.16

La transmisión de puntajes de pruebas con identificación personal a individuos o instituciones autorizadas se debe hacer de manera que se proteja la naturaleza confidencial de los puntajes y la información complementaria pertinente.

Comentario: Siempre hace falta poner mucha atención cuando se comunican los puntajes de examinandos identificados, independientemente del medio de comunicación. Puede ser necesaria una atención similar para proteger la confidencialidad de la información complementaria, por ejemplo, información de identificación personal sobre el estado de discapacidad de estudiantes o puntajes de pruebas clínicas que comparten los médicos. Se deben tomar las precauciones apropiadas con respecto a la información confidencial en las comunicaciones presenciales, así como por teléfono, fax y otras formas de comunicación escrita. De forma similar, la transmisión de datos de las pruebas a través de medios electrónicos y la transmisión y almacenamiento en redes computarizadas (incluyendo la transmisión y almacenamiento inalámbricos o el procesamiento en Internet) requieren precauciones para mantener la confidencialidad y seguridad apropiadas. También se debe mantener la integridad de los datos impidiendo la modificación inapropiada de los resultados durante tales transmisiones. Los usuarios de las pruebas son responsables de conocer y adherirse a las obligaciones legales vigentes en materia de gestión, transmisión, uso y prácticas de retención de datos, incluyendo la recopilación, manipulación, almacenamiento y disposición. Los usuarios de las pruebas deben establecer y seguir las políticas de seguridad apropiadas relacionadas con los datos confidenciales de las pruebas y otra información de evaluación. La publicación de datos brutos, pruebas o protocolos clínicos a terceros deben seguir las leyes, normativas y directrices proporcionadas por las organizaciones profesionales y deben tener en cuenta el impacto de la disponibilidad de las pruebas en dominios públicos (p. ej., en procesos judiciales) y las posibilidades de infracción de los derechos de propiedad intelectual.

7. DOCUMENTACIÓN DE RESPALDO DE LAS PRUEBAS

ANTECEDENTES

Este capítulo incluye los estándares generales para la preparación y publicación de la documentación de las pruebas por parte de desarrolladores, editores y otros proveedores de pruebas. Otros capítulos contienen los estándares específicos que se usan en la preparación de los materiales a incluir en la documentación de una prueba. Además, es posible que los usuarios de la prueba tengan sus propios requisitos de documentación. Los derechos y responsabilidades de los usuarios de la prueba se tratan en el capítulo 9.

Los documentos de respaldo de las pruebas son el medio principal que los desarrolladores, editores y otros proveedores de pruebas utilizan para comunicarse con los usuarios de la prueba. Estos documentos se evalúan basándose en su integridad, precisión, actualidad y claridad, y deben estar disponibles para los individuos cualificados cuando proceda. Habitualmente, la documentación de una prueba específica la naturaleza de la prueba; los usos para los que se ha desarrollado; los procesos incluidos en el desarrollo de la prueba; información técnica relacionada con el puntaje, la interpretación y con las evidencias de validez, imparcialidad y confiabilidad/precisión; información sobre el escalamiento, la normalización y el establecimiento de estándares si resulta apropiado para el instrumento; y las directrices para la administración, presentación de reportes e interpretación de la prueba. El objetivo de la documentación es proporcionar a los usuarios de la prueba la información necesaria para evaluar la naturaleza y calidad de la prueba, los puntajes resultantes y las interpretaciones basadas en los puntajes de la prueba. La información puede reportarse en documentos como los manuales de la prueba, manuales técnicos, guías de usuario, reportes de investigación, conjuntos de muestras, kits de examen, instrucciones para los administradores y evaluadores de la prueba, o materiales de vista previa para los examinandos.

Independientemente de quién desarrolle la prueba (p. ej. el editor de la prueba, un consejo de certificación o licenciamiento, un empleador o una institución educativa) o del número de usuarios, el proceso de desarrollo debe incluir una documentación exhaustiva, oportuna y útil. Aunque es importante una documentación correcta de la evidencia que respalda la interpretación de los puntajes para los usos propuestos de una prueba, las deficiencias en documentar formalmente con antelación tales evidencias no se traducen automáticamente en la carencia de validez del uso o la interpretación correspondientes de la prueba. Por ejemplo, considere una prueba de selección de empleo no publicada, desarrollada por un psicólogo únicamente para uso interno dentro de una organización donde existe una necesidad inmediata de cubrir las vacantes. Es posible que la prueba se ponga en uso operativo después de que se haya recolectado la evidencia de validación necesaria, pero antes de completar la documentación formal de la evidencia. De forma similar, es posible que una prueba usada para la certificación deba revisarse con frecuencia, en cuyo caso se deben generar periódicamente reportes técnicos que describan el desarrollo de la prueba, así como la información relativa a los ítems, el examen y el desempeño de los candidatos, pero no necesariamente antes de cada examen.

La documentación de la prueba es eficaz si comunica información a los grupos de usuarios en un modo que resulte apropiado para la audiencia específica. Para adaptarse al nivel de capacitación de quienes usan las pruebas, se pueden redactar documentos separados o secciones de documentos para categorías identificables de usuarios como médicos, consultores, administradores, investigadores, educadores y, en ocasiones, examinandos. Por ejemplo, el usuario de una prueba que administre las pruebas e interprete los resultados necesita directrices para hacer estas tareas.

Los responsables de seleccionar las pruebas necesitan tener la capacidad de juzgar la idoneidad técnica de las pruebas y, por lo tanto, requieren una combinación de manuales técnicos, guías de usuario, manuales de la prueba, complementos de la prueba, kits de examen y conjuntos de muestras. Normalmente, estos documentos de respaldo se suministran a los usuarios potenciales o a revisores de la prueba, con suficiente información para permitirles evaluar la pertinencia e idoneidad técnica de una prueba. Los tipos de información presentados en estos documentos incluyen, por lo general, una descripción de la población de examinandos objetivo, el propósito declarado de la prueba, las especificaciones de la prueba, los formatos de ítems, los procedimientos de administración y calificación, los protocolos de seguridad de la prueba, los puntajes de corte u otros estándares, y una descripción del proceso de desarrollo de la prueba. Habitualmente, también se suministran resúmenes de datos técnicos como, por ejemplo, índices psicométricos de los ítems, evidencias de validez y confiabilidad/precisión, datos normativos, y puntajes de corte o reglas para la combinación de puntajes, incluyendo las reglas para las interpretaciones generadas por computadora.

Una característica esencial de la documentación para cualquier prueba son las deliberaciones de los usos comunes apropiados o inapropiados de los puntajes y un resumen de la evidencia que respalda las conclusiones. La inclusión de ejemplos de interpretaciones de puntajes coherentes con las aplicaciones previstas por los desarrolladores de la prueba resulta útil para que los usuarios puedan extraer inferencias precisas sobre la base de los puntajes. Cuando sea posible, los ejemplos de usos inapropiados de la prueba o de interpretaciones inadecuadas de los puntajes resultarán útiles como salvaguarda ante usos indebidos de las pruebas o de sus puntajes. Cuando sea factible, se deben describir las consecuencias negativas comunes, no intencionadas, del uso de las pruebas (incluyendo las oportunidades perdidas) y se deben ofrecer sugerencias para evitar tales consecuencias.

Los documentos de la prueba deben incluir suficiente información para permitir que los usuarios y revisores de la prueba determinen la pertinencia de la prueba para los usos previstos. Se deben citar otros materiales que proporcionen más detalles sobre la investigación por parte del editor o de investigadores independientes (p. ej. las muestras en que se basa la investigación y los datos sumariales) y el usuario o revisor de la prueba debe poder conseguirlos fácilmente. Este material complementario se puede suministrar en cualquier tipo de modalidad de publicación o inédita, ya sea en formato papel o electrónico.

Además de la documentación técnica, en algunos contextos se requieren materiales descriptivos para informar a los examinandos y a otras partes interesadas de la naturaleza y contenido de la prueba. La cantidad y el tipo de información suministrada dependerán de la prueba y las aplicaciones específicas. Por ejemplo, en situaciones que requieren consentimiento informado, la información debe ser suficiente para que los examinandos (o sus representantes) puedan tener un criterio sólido sobre la prueba. Esta información debe formularse en lenguaje no técnico y debe contener información que sea coherente con el uso de los puntajes de la prueba, y debe ser suficiente para ayudar a que el usuario tome una decisión informada. Los materiales pueden incluir una descripción general y la justificación de la prueba, los usuarios previstos de los resultados de la prueba, ítems de muestra o pruebas con muestras completas, e información sobre las condiciones de administración, confidencialidad y retención de los resultados. Sin embargo, para algunas aplicaciones, el nombre y la finalidad verdaderos se ocultan o encubren deliberadamente para evitar la simulación o el sesgo de las respuestas. En estos casos, los examinandos podrían sentirse motivados a revelar más o menos de una característica que se pretende evaluar. El ocultamiento o encubrimiento de la verdadera naturaleza o finalidad de una prueba son aceptables siempre y cuando las acciones que comportan sean coherentes con los principios legales y los estándares éticos.

ESTÁNDARES PARA LA DOCUMENTACIÓN DE RESPALDO DE LAS PRUEBAS

Los estándares de este capítulo empiezan con un estándar general (con el número 7.0), diseñado para comunicar el propósito central o el enfoque principal del capítulo. El estándar general también se puede ver como el principio rector del capítulo y se aplica a todas las pruebas y a todos los usuarios de la prueba. Todos los estándares posteriores se han divididos en cuatro unidades temáticas, etiquetadas de la siguiente manera:

1. Contenido de documentos de la prueba: Uso apropiado
2. Contenido de documentos de la prueba: Desarrollo de la prueba
3. Contenido de documentos de la prueba: Administración y calificación de la prueba
4. Cumplimiento de los plazos de entrega de los documentos de la prueba

Estándar 7.0

La información relacionada con las pruebas se debe documentar claramente, de manera que quienes usen las pruebas puedan tomar decisiones informadas respecto de qué prueba usar para un propósito concreto, cómo administrar la prueba seleccionada y cómo interpretar los puntajes.

Comentario: Los desarrolladores y editores de la prueba deben proporcionar información general que ayude a los usuarios de la prueba y a los investigadores a determinar la pertinencia de un uso previsto de la prueba en un contexto específico. Cuando los desarrolladores y editores tienen conocimiento de un uso específico que no se puede justificar, deben indicar este hecho con claridad. También se debe proporcionar información general para los examinandos y representantes legales que deben dar su consentimiento antes de la administración de la prueba (véase el estándar 8.4 relativo al consentimiento informado). También es posible que los administradores, e incluso el público general, necesiten información general sobre la prueba y sus resultados, de forma que puedan interpretar correctamente los mismos.

Los documentos de la prueba deben estar completos, ser precisos y estar claramente redactados, de manera que la audiencia prevista pueda entender fácilmente el contenido. La documentación de la prueba se debe suministrar en un formato que sea accesible para la población a la que se dirige. En las pruebas usadas para fines de rendición de cuentas educativa, la documentación debe estar disponible públicamente en un formato y lenguaje accesible a los usuarios potenciales, incluyendo el personal de la escuela, los padres, los estudiantes de todos los subgrupos relevantes de examinandos previstos y los miembros de la comunidad (p. ej., a través de Internet). La documentación de la prueba en contextos educativos también podría incluir orientación sobre la forma en que los usuarios pueden usar los materiales y resultados de la prueba para mejorar su instrucción.

Los documentos de la prueba deben proporcionar los detalles suficientes para permitir que los revisores e investigadores evalúen los análisis relevantes publicados en el manual o en el reporte técnico de la prueba. Por ejemplo, reportar matrices de correlación en el documento de la prueba puede permitir que el usuario de la prueba juzgue los datos en los que se basan las decisiones y conclusiones. De forma similar, la descripción detallada de la muestra y la naturaleza del análisis de factores que se llevó a cabo podría permitir al usuario de la prueba replicar los estudios reportados.

La documentación de la prueba también ayudará a quienes se vean afectados por las interpretaciones de los puntajes para decidir su participación en el programa de pruebas o cómo participar si la participación no es opcional.

Unidad 1. Contenido de documentos de la prueba: Uso apropiado

Estándar 7.1

Se debe documentar la justificación de una prueba, los usos recomendados de una prueba,

el respaldo de dichos usos y la información que apoya la interpretación de los puntajes. Cuando se puede anticipar razonablemente el uso inadecuado de una prueba, se deben especificar las precauciones contra tales usos.

Comentario: Los editores de la prueba deben hacer todo lo necesario para prevenir a los usuarios de la prueba contra los usos inadecuados. Sin embargo, los editores no pueden anticipar todos los usos inadecuados. Si los editores tienen conocimiento del uso inadecuado persistente por parte de un usuario de la prueba, es posible que resulten apropiadas acciones educativas adicionales, incluyendo facilitar información sobre los perjuicios potenciales para el individuo, la organización o la sociedad.

Estándar 7.2

Se debe documentar la población a la que se destina una prueba y las especificaciones de la prueba. Si se proporcionan datos normativos, se deben explicar los procedimientos para recopilar los datos, se debe describir la población de normalización en términos de variables demográficas relevantes y se debe informar sobre los años en que se recopilaron los datos.

Comentario: En los documentos de una prueba, se deben definir claramente las limitaciones conocidas de la prueba para determinadas poblaciones. Por ejemplo, es posible que una prueba utilizada para evaluar los progresos no sea apropiada para la selección de empleados en el comercio o la industria.

El usuario puede usar otro tipo de documentación para identificar la información normativa apropiada que se debe usar para una interpretación apropiada de los puntajes. Por ejemplo, el momento del año en que se recopilaron los datos normativos puede ser relevante en algunos contextos educativos. En contextos organizativos, la información sobre el contexto en que se reunieron los datos normativos (p. ej., en estudios concurrentes o predictivos; para fines de desarrollo o selección) también puede repercutir en lo que respecta a las normas apropiadas para el uso operativo.

Estándar 7.3

Cuando la información está disponible y se comparte de manera apropiada, los documentos de la prueba deben mencionar un conjunto representativo de los estudios concernientes a los usos específicos y generales de la prueba.

Comentario: Si un estudio citado por el editor de la prueba no ha sido publicado, el editor debe poner a disposición resúmenes a petición del usuario de la prueba y de los investigadores.

Unidad 2. Contenido de documentos de la prueba: Desarrollo de la prueba

Estándar 7.4

La documentación de la prueba debe resumir los procedimientos de desarrollo de la prueba, incluyendo descripciones y los resultados de los análisis estadísticos que se usaron en el desarrollo de la prueba, evidencia de la confiabilidad/precisión de los puntajes y la validez de sus interpretaciones recomendadas, y los métodos para establecer los puntajes de corte para el desempeño.

Comentario: Cuando corresponda, los documentos de la prueba deben incluir descripciones de los procedimientos usados para desarrollar los ítems y crear los conjuntos de ítems, para crear las pruebas o los formularios de las pruebas, para establecer escalas para los puntajes reportados y para determinar los estándares y reglas para los puntajes de corte o la combinación de puntajes. Los documentos de la prueba también deben proporcionar información que permita al usuario evaluar el sesgo o la imparcialidad para todos los grupos relevantes de examinandos previstos cuando sea relevante y factible llevar a cabo tales estudios. Además, se deben proporcionar otros datos estadísticos cuando sea apropiado, por ejemplo, información de nivel de ítem, información sobre los efectos de varios puntajes de corte (p. ej., número de candidatos que aprueban puntajes de corte potenciales, nivel de impacto adverso en puntajes

de corte potenciales), información sobre puntajes brutos y puntajes repetidos, datos normativos, los errores estándar de medida y una descripción de los procedimientos usados para equiparar diversos formularios (véase los capítulos 3 y 4 para obtener más información sobre evaluación de la imparcialidad y sobre los procedimientos y estadísticas de uso común en el desarrollo de pruebas).

Estándar 7.5

Los documentos de la prueba deben registrar las características relevantes de los individuos o grupos de individuos que participan en los trabajos de recolección de datos asociados con el desarrollo o la validación de la prueba (p. ej., información demográfica, situación laboral, nivel de grado), la naturaleza de los datos aportados (p. ej., datos de pronóstico, datos de criterio), la naturaleza de los juicios hechos por expertos en la materia (p. ej., vinculaciones de validación de contenido), las instrucciones que se proporcionaron a los participantes en los trabajos de recolección de datos para tareas específicas, y las condiciones bajo las cuales se recolectaron los datos del estudio de validez.

Comentario: Los desarrolladores de la prueba deben describir las características relevantes de quienes participan en las diferentes fases del proceso de desarrollo de la prueba y qué tareas realizó cada persona o grupo. Por ejemplo, se debe documentar quiénes son los participantes que determinaron los puntajes de corte y sus experiencias pertinentes. En función del uso de los resultados de la prueba, las características relevantes de los participantes pueden incluir la raza/origen étnico, género, edad, situación laboral, educación, situación de discapacidad e idioma principal. Las descripciones de las tareas y las instrucciones específicas proporcionadas a los participantes pueden ser útiles para que los futuros usuarios de la prueba seleccionen, y posteriormente usen, la prueba de manera apropiada. Las condiciones de las pruebas (por ejemplo, la extensión de la monitorización en el estudio de validez) pueden tener implicaciones para la generalización de los resultados y, por lo tanto, se debe documentar. También se debe documentar cualquier cambio en las condiciones estandarizas de las pruebas, por ejemplo, las adecuaciones y modificaciones hechas en las pruebas o en la administración de la prueba. Cuando se facilite la documentación requerida por este estándar, los desarrolladores y usuarios deben prestar atención al cumplimiento de los requisitos legales vigentes y de los estándares profesionales relacionados con la privacidad y seguridad de los datos.

Estándar 7.6

Cuando una prueba está disponible en más de un idioma, la documentación de la prueba debe proporcionar información sobre los procedimientos que se emplearon para traducir y adaptar la prueba. Cuando sea factible, también se deberá suministrar información relacionada con la evidencia de confiabilidad/precisión y validez.

Comentario: Además de proporcionar información sobre los procedimientos de traducción y adaptación, los documentos de la prueba deben incluir aspectos demográficos de los traductores y muestras de examinandos usadas en el proceso de adaptación, así como información sobre los problemas de interpretación de puntajes en cada uno de los idiomas a los que la prueba se haya traducido y adaptado. Cuando sea factible, se deberá proporcionar la evidencia de confiabilidad/precisión, validez y comparabilidad de los puntajes traducidos y adaptados (véase el estándar 3.14, en el capítulo 3, para más información sobre las traducciones).

Unidad 3. Contenido de documentos de la prueba: Administración y calificación de la prueba

Estándar 7.7

Los documentos de la prueba deben especificar las cualificaciones de usuario que se requieren para administrar y calificar una prueba, así como

las cualificaciones de usuario necesarias para interpretar con precisión los puntajes.

Comentario: Las declaraciones de las cualificaciones de usuario deben especificar la capacitación, la certificación, las competencias y la experiencia necesarias para permitir el acceso a una prueba o a los puntajes obtenidos con la misma. Cuando las cualificaciones se expresan en términos de conocimientos, competencias, capacidades y otras características requeridas para administrar, calificar e interpretar una prueba, la documentación de la prueba debe definir claramente los requisitos, de manera que el usuario pueda evaluar adecuadamente la competencia de los administradores.

Estándar 7.8

La documentación de la prueba debe incluir instrucciones detalladas sobre la administración y calificación de una prueba.

Comentario: Independientemente de que vaya a ser administrada en formato de papel y lápiz, formato computarizado u oralmente, o de que la prueba se base en el desempeño, la documentación de la prueba debe incluir las instrucciones de administración. Cuando proceda, estas instrucciones deberán incluir todos los factores relacionados con la administración de la prueba, incluyendo las cualificaciones, competencias y capacitación de los administradores de la prueba; los equipos necesarios; los protocolos para los administradores; las instrucciones de cronometraje y los procedimientos para la implementación de las adecuaciones de la prueba. Cuando estén disponibles, la documentación de la prueba incluirá estimaciones del tiempo requerido para administrar la prueba a poblaciones clínicas, poblaciones con discapacidades u otras poblaciones especiales con las que se prevé usar la prueba, basándose en los datos obtenidos de estos grupos durante la normalización de la prueba. Además, los usuarios de la prueba necesitan instrucciones sobre cómo calificar una prueba y qué puntajes de corte usar (o si se deben usar puntajes de corte) en la interpretación de puntajes. Si el usuario de la prueba no califica la prueba, se deben dar instrucciones sobre la forma de obtener una prueba calificada. Finalmente, la documentación de administración de una prueba debe incluir instrucciones para tratar con las irregularidades en la administración de la prueba y orientación sobre la forma de documentarlas.

Si una prueba está diseñada de manera que se puede usar más de un método para la administración o para el registro de las respuestas (por ejemplo, dar las respuestas en un cuadernillo, en una hoja separada o mediante computadora), el manual debe documentar claramente el grado en que los puntajes que proceden de la aplicación de estos métodos son intercambiables. Si los puntajes no son intercambiables, se debe reportar este hecho y se proporcionará orientación sobre la comparabilidad de los puntajes obtenidos bajo las diversas condiciones o métodos de administración.

Estándar 7.9

Si la seguridad de la prueba es crítica para la interpretación de los puntajes, la documentación debe explicar los pasos necesarios para proteger los materiales de la prueba y para evitar el intercambio inapropiado de información durante la sesión de administración.

Comentario: Cuando la interpretación correcta de los puntajes asume que el examinando no se ha visto expuesto al contenido de la prueba ni ha recibido asistencia ilícita, las instrucciones deben incluir procedimientos para garantizar la seguridad del proceso de evaluación y de todos los materiales de la prueba en todo momento. Los procedimientos de seguridad pueden incluir orientación para el almacenamiento y la distribución de los materiales de la prueba, así como instrucciones para mantener un proceso de evaluación seguro (por ejemplo, la identificación de los examinandos y la colocación de estos para evitar el intercambio de información). Los usuarios de la prueba deben ser conscientes de que las leyes, normativas y políticas federales y estatales pueden afectar a los procedimientos de seguridad.

En muchas situaciones, también se debe mantener la seguridad de los puntajes de las pruebas. Por ejemplo, en las pruebas de ascenso de algunos contextos laborales, solo el candidato y el personal de contratación tienen autorización para ver los puntajes, y el supervisor actual del candidato tiene expresamente prohibido hacerlo. La documentación puede incluir información sobre el almacenamiento de los puntajes y sobre las personas autorizadas para verlos.

Estándar 7.10

Las pruebas diseñadas para ser calificadas e interpretadas por examinandos deben incluir instrucciones de calificación y materiales interpretativos escritos en un idioma que los examinandos comprendan y que les ayuden a entender los puntajes de las pruebas.

Comentario: Si una prueba está diseñada para ser calificada por examinandos o para que sus puntajes sean interpretados por los mismos, el editor y desarrollador de la prueba deberá desarrollar procedimientos que faciliten la calificación e interpretación precisas. El material interpretativo puede incluir información como, por ejemplo, el constructo que se ha medido, los resultados del examinando y el grupo de comparación. El idioma apropiado para los procedimientos de calificación y los materiales interpretativos es el que satisfaga las necesidades específicas del examinando. Por lo tanto, es posible que los puntajes y materiales interpretativos tengan que proporcionarse en el idioma nativo del examinando para que puedan ser comprendidos.

Estándar 7.11

Los materiales interpretativos para las pruebas que incluyen estudios de caso deben proporcionar ejemplos que ilustren la diversidad de los posibles examinandos.

Comentario: Cuando los estudios de caso puedan ayudar al usuario en la interpretación de los puntajes y perfiles de la prueba, se deben incluir los estudios de caso en la documentación de la prueba y representar a miembros de los subgrupos para los que la prueba resulte pertinente. Para ilustrar la diversidad de los posibles examinandos, los estudios de caso pueden citar ejemplos donde participen mujeres y hombres de edades diversas, individuos que difieren en su orientación sexual, personas que representen varios grupos raciales o étnicos, e individuos con discapacidades. Los desarrolladores de la prueba pueden tal vez informar a los usuarios de que la inclusión de tales ejemplos tiene como fin ilustrar la diversidad de los examinandos potenciales y no promover la interpretación de los puntajes de manera que pueda entrar en conflicto con requisitos legales como la normalización del origen étnico o el género en contextos de empleo.

Estándar 7.12

Cuando los puntajes de las pruebas se usan para hacer predicciones sobre el comportamiento futuro, se debe proporcionar al usuario de la prueba la evidencia que respalda esas predicciones.

Comentario: Se debe informar al usuario de la prueba sobre cualquier puntaje de corte o regla para la combinación de puntajes brutos o reportados que sean necesarios para entender las interpretaciones de los puntajes. Se debe proporcionar una descripción de los grupos de jueces que establecen los puntajes de corte y de los métodos usados para obtener los puntajes de corte. Cuando se requiere la retención de los puntajes de corte o de las reglas para combinar puntajes por motivos de seguridad o propiedad, los propietarios de la propiedad intelectual serán responsables de documentar la evidencia que respalda la validez de las interpretaciones para los usos previstos. Estas evidencias se facilitarán, por ejemplo, mediante el reporte de los hallazgos de una revisión independiente de los algoritmos por profesionales cualificados. Cuando se proporcionen interpretaciones de los puntajes, incluyendo interpretaciones generadas por computadora, se facilitará un resumen de la evidencia que respalda las interpretaciones, así como las reglas y directrices usadas en la formulación de las interpretaciones.

Unidad 4. Cumplimiento de los plazos de entrega de los documentos de la prueba

Estándar 7.13

Los documentos de respaldo (p. ej., manuales de la prueba, manuales técnicos, guías de usuario y material complementario) deben estar disponibles para las personas apropiadas en el momento adecuado.

Comentario: Los documentos de respaldo deben suministrarse de forma oportuna. Algunos documentos (p. ej., instrucciones de administración, guías de usuario, pruebas o ítems de ejemplo) deben estar disponibles antes de la primera administración de la prueba. Otros documentos (p. ej., manuales técnicos que contienen información basada en datos de la primera administración) no se pueden suministrar antes de esa administración; no obstante, estos documentos deberán crearse rápidamente.

El desarrollador o editor de la prueba deberá ponderar cuidadosamente qué información se debe incluir en las primeras ediciones del manual de la prueba, el manual técnico o la guía de usuario, y qué información puede suministrarse de forma complementaria. Para pruebas inéditas de bajo volumen, la documentación puede ser relativamente breve. Cuando el desarrollador y el usuario son el mismo, la documentación y los resúmenes seguirán siendo necesarios.

Estándar 7.14

Cuando se hagan cambios importantes en una prueba, la documentación de la prueba se debe enmendar, complementar o revisar para mantener actualizada la información para los usuarios y para proporcionar información o precauciones adicionales útiles.

Comentario: Los documentos de respaldo deben indicar claramente la fecha de su publicación, así como el nombre o versión de la prueba para la que son relevantes. Cuando se realizan cambios importantes en los ítems o la calificación, la documentación de la prueba debe incluir información sobre el grado en que los antiguos y nuevos puntajes pueden ser intercambiables.

En ocasiones es necesario cambiar una prueba o los procedimientos de una prueba para eliminar la varianza irrelevante de constructo que pueda presentarse debido a características de un individuo que no están relacionadas con el constructo que se mide (p. ej., en pruebas con individuos con discapacidades). Cuando se altera una prueba o los procedimientos de una prueba, la documentación deberá incluir las deliberaciones sobre el modo en que la alteración puede afectar a la validez y comparabilidad de los puntajes de la prueba, y se deben proporcionar evidencias para demostrar el efecto de la alteración en los puntajes obtenidos de la prueba o procedimientos alterados, si el tamaño de la muestra lo permite.

8. DERECHOS Y RESPONSABILIDADES DE LOS EXAMINANDOS

ANTECEDENTES

Este capítulo examina los problemas de imparcialidad desde el punto de vista del examinando individual. La mayoría de los aspectos de imparcialidad afectan a la validez de las interpretaciones de los puntajes para los usos previstos. Los estándares de este capítulo abordan los derechos y responsabilidades de los examinandos con respecto a la seguridad de la prueba, su acceso a los resultados de la prueba y sus derechos cuando se reclaman irregularidades en el proceso de evaluación. En el capítulo 3 ("Imparcialidad en las pruebas") se tratan otros temas sobre la imparcialidad. En el capítulo 6 ("Administración, calificación, presentación de reportes e interpretación de pruebas") se incluyen consideraciones generales relacionadas con los reportes de los resultados de las pruebas. En el capítulo 10 ("Pruebas y evaluación psicológicas") se analizan los problemas relacionados con los derechos y responsabilidad de los examinandos en contextos clínicos o individuales.

Los estándares de este capítulo se dirigen a los proveedores de pruebas, no a los examinandos. Es responsabilidad compartida del desarrollador, administrador, monitor (si existe) y usuario de la prueba proporcionar a los examinandos la información sobre sus derechos y sus propias responsabilidades. La responsabilidad de informar al examinando se deberá distribuir de acuerdo con las circunstancias específicas.

Los examinandos tienen el derecho de ser evaluados con pruebas que cumplan los actuales estándares profesionales, incluyendo los estándares de calidad técnica, tratamiento coherente, imparcialidad, condiciones de administración y presentación de reportes de resultados. Los capítulos de la Parte I, "Fundamentos" y de la Parte II, "Operaciones", tratan de forma específica el diseño imparcial y apropiado, el desarrollo, la administración, la calificación y la presentación de reportes de las pruebas. Además, los examinandos tienen el derecho de disponer de información básica sobre la prueba y la forma en que se utilizarán los resultados. En la mayoría de las situaciones, el tratamiento imparcial y equitativo de los examinandos comporta proporcionar con antelación información sobre la naturaleza general de la prueba, el uso previsto de los puntajes y la confidencialidad de los resultados. Cuando no sea apropiada la divulgación completa de esta información (como en el caso de algunas pruebas psicológicas o de empleo), la información que se facilite debe ser uniforme para todos los examinandos. Los examinandos, o sus representantes legales cuando sea apropiado, necesitan suficiente información sobre la prueba y el uso previsto de los resultados para tomar una decisión informada sobre su participación.

En algunos casos, las leyes o estándares de práctica profesional (por ejemplo, las que rigen la investigación sobre sujetos humanos) exigen un consentimiento informado formal para realizar las pruebas. En otros casos, (p. ej., las pruebas de empleo), el consentimiento informado está implícito en otras acciones (p. ej., el envío de una solicitud de empleo) y no se requiere de un consentimiento formal. Cuanto mayores son las consecuencias para el examinando, mayor es la importancia de garantizar que el examinando cuenta con toda la información sobre la prueba y que su participación se realiza voluntariamente, excepto cuando la ley permite las pruebas sin consentimiento (p. ej., cuando la participación en una prueba es una exigencia legal o ha sido ordenada por mandato judicial). Si una prueba es opcional, el examinando tiene el derecho de saber las consecuencias de realizar o no realizar la prueba. En la mayoría de los casos, el examinando tiene el derecho de hacer preguntas o formular dudas y debe recibir una respuesta oportuna a las consultas legítimas.

Habitualmente, cuando sea coherente con los propósitos y la naturaleza de la evaluación, la

información general proporcionará el contenido y los propósitos de la prueba. En algunos programas, en interés de la imparcialidad, se les facilitan a todos los examinandos materiales útiles, como guías de estudio, preguntas de ejemplo o pruebas completas de ejemplo, cuando tal información no comprometa la validez de las interpretaciones de los resultados de futuras administraciones de pruebas. Los materiales de práctica deben tener la misma apariencia y formato que la prueba real. Por ejemplo, una prueba de práctica para una evaluación basada en la Web debe estar disponible a través de una computadora. Los programas de selección de personal pueden proporcionar, de forma legítima, más capacitación a determinados tipos de examinandos (p. ej., candidatos internos) que a otros (p. ej., candidatos externos). Por ejemplo, en el contexto de un programa de desarrollo de personal, una organización puede capacitar a los empleados actuales en competencias que se miden en las pruebas de empleo y no ofrecer esa capacitación a los candidatos externos. También se puede facilitar asesoría sobre las estrategias de los examinandos, incluyendo la gestión del tiempo y la conveniencia de omitir una respuesta en un ítem (cuando se admite la omisión de respuestas). También se proporciona al examinando información sobre las diversas políticas de evaluación, por ejemplo, sobre la disponibilidad de las adecuaciones y la determinación de la idoneidad de las adecuaciones para determinados individuos. Además, las comunicaciones a los examinandos deben incluir las políticas de contraprueba cuando se producen graves alteraciones en la administración de la prueba, cuando los examinandos creen que el desempeño actual no refleja apropiadamente sus capacidades reales, o cuando el examinando mejora sus conocimientos, competencias, capacidades u otras características subyacentes,

Como participantes de una evaluación, los examinandos tienen responsabilidades además de derechos. Sus responsabilidades incluyen estar preparados para realizar la prueba, seguir las instrucciones del administrador de la prueba, reflejarse a sí mismos con honestidad en la prueba y proteger la seguridad de los materiales de la prueba. La solicitud de adecuaciones o modificaciones son responsabilidad del examinando, o en el caso de menores de edad, del tutor del examinando. En situaciones de pruebas de grupo, los examinandos no deben interferir con el desempeño de otros examinandos. En algunos programas de pruebas, también se espera que los examinandos informen a las personas apropiadas, y de manera oportuna, si encuentran motivos para creer que sus resultados no reflejarán sus capacidades verdaderas.

La validez de las interpretaciones de los puntajes se basa en el supuesto de que un examinando ha obtenido honestamente un puntaje o decisión categórica específicos, como "aprobado" o "reprobado." La mayoría de los comportamientos fraudulentos o engañosos pueden reducir la validez de las interpretaciones de los puntajes y causar perjuicios a otros examinandos, sobre todo en situaciones competitivas donde se comparan los puntajes de los examinandos. Hay muchas formas de comportamiento que afectan a los puntajes de una prueba, por ejemplo, el uso de ayudas prohibidas o la suplantación de examinandos. De forma similar, hay muchas formas de comportamiento que comprometen la seguridad de los materiales de la prueba, incluyendo comunicar de antemano el contenido específico de la prueba a otros examinandos. El examinando está obligado a respetar los derechos de autor de los materiales de la prueba y no puede reproducir los materiales sin autorización ni divulgar de ninguna forma material de naturaleza similar a la prueba. Los examinandos, así como los administradores de la prueba, tienen la responsabilidad de proteger la seguridad de la prueba negándose a divulgar cualquier detalle del contenido de una prueba, a menos que la prueba concreta se haya diseñado para estar disponible con antelación. No cumplir con estas responsabilidades puede poner en riesgo la validez de las interpretaciones de los puntajes tanto para el examinando como para los demás. Los grupos externos que desarrollan ítems para la preparación de pruebas deben basar esos ítems en la información divulgada públicamente y no en información que los examinandos hayan compartido de manera inapropiada.

A veces, los programas de pruebas usan puntajes especiales, indicadores estadísticos y otros datos indirectos sobre irregularidades en las pruebas para examinar si los puntajes de una prueba se han obtenido limpiamente. Patrones inusuales de respuestas, grandes cambios en los puntajes de prueba y contraprueba, la velocidad de las respuestas e indicadores similares pueden acarrear un escrutinio detallado de determinados protocolos de evaluación y puntajes de pruebas. Por lo general, los detalles de los procedimientos para la detección de problemas se mantienen confidenciales para evitar comprometer su uso. Sin embargo, se debe informar a los examinandos de que, en circunstancias especiales (como anomalías en las respuestas o en los puntajes de una prueba), sus respuestas pueden someterse a un escrutinio especial. Se debe informar a los examinandos de que, si se detectan evidencias de irregularidad o fraude, sus puntajes podrían anularse o tomarse otras medidas.

ESTÁNDARES PARA LOS DERECHOS Y RESPONSABILIDADES DE LOS EXAMINANDOS

Los estándares de este capítulo empiezan con un estándar general (con el número 8.0), diseñado para comunicar el propósito central o el enfoque principal del capítulo. El estándar general también se puede ver como el principio rector del capítulo y se aplica a todas las pruebas y a todos los usuarios de la prueba. Todos los estándares posteriores se han dividido en cuatro unidades temáticas, etiquetadas de la siguiente manera:

1. Derechos de los examinandos a disponer de información antes de la prueba
2. Derechos de los examinandos a obtener acceso a los resultados de sus pruebas y a la protección frente a usos no autorizados de estos resultados
3. Derechos de los examinandos a reportes de puntajes imparciales y precisos
4. Responsabilidades de comportamiento de los examinandos a lo largo de todo el proceso de administración de la prueba

Estándar 8.0

Los examinandos tienen el derecho de disponer de información adecuada que les ayude a prepararse apropiadamente para una prueba, de manera que los resultados reflejen con exactitud su situación en el constructo que se evalúa y lleven a interpretaciones precisas e imparciales. También tienen el derecho a la protección de los resultados con identificación personal frente al acceso, uso o divulgación no autorizados. Además, los examinandos tienen la responsabilidad de reflejarse a sí mismos con precisión en el proceso de la prueba y de respetar los derechos de autor de los materiales de la prueba.

Comentario: A continuación, se describen los estándares específicos para los derechos y responsabilidades de los examinandos. Estos incluyen estándares para los tipos de información que se debe proporcionar a los examinandos antes de la prueba, de modo que puedan prepararse apropiadamente para realizar la prueba y que los resultados reflejen con exactitud su situación en el constructo que se evalúa. Los estándares también incluyen el acceso de los examinandos a los resultados de sus pruebas; la protección de los resultados frente al acceso, uso o divulgación no autorizados por parte de terceros, y los derechos de los examinandos a reportes de puntajes precisos e imparciales. Además, los estándares de este capítulo abordan la responsabilidad de los examinandos de reflejarse a sí mismos con precisión y de forma imparcial durante el proceso de la prueba, y de respetar la confidencialidad de los derechos de autor de los materiales de la prueba.

Unidad 1. Derechos de los examinandos a disponer de información antes de la prueba

Estándar 8.1

La información sobre el contenido y el propósito de la prueba que esté disponible para cualquier examinando antes de la prueba debe estar disponible para todos los examinandos. La información compartida debe estar disponible de forma gratuita y en formatos accesibles.

Comentario: El objetivo de este estándar es el tratamiento equitativo de todos los examinandos con respecto al acceso a información básica sobre un evento de prueba, por ejemplo, cuándo y dónde se llevará a cabo, qué materiales se deben llevar, cuál es el propósito de la prueba y cómo se utilizarán los resultados. Cuando corresponda, estas ofertas se harán a todos los examinandos y, en la medida de lo posible, deben estar en formatos accesibles a todos los examinandos. La accesibilidad de los formatos también se aplica a la información que se podría proporcionar en un sitio web público. Por ejemplo, en función del formato de la información, se pueden hacer conversiones para que las personas con discapacidades visuales

puedan acceder al material gráfico o textual. Es posible que el suministro de estos materiales en formatos accesible sea un imperativo legal en el caso de examinandos con discapacidades.

Cabe señalar que, aunque la información general sobre el contenido y el propósito de la prueba debe estar disponible para todos los examinandos, algunas organizaciones pueden complementar esta información con capacitación u orientación adicional. Por ejemplo, algunos empleadores pueden impartir competencias básicas a los trabajadores para ayudarles a cualificarse para puestos de mayor nivel. De forma similar, un profesor de escuela puede decidir entrenar a los estudiantes en un tema que se examinará, mientras que otros profesores se centran en otros temas.

Estándar 8.2

Se debe proporcionar con antelación a los examinandos tanta información sobre la prueba, el proceso de evaluación, el uso previsto de la prueba, los criterios de calificación, la política de evaluación, la disponibilidad de adecuaciones y la protección de la confidencialidad como sea compatible con la obtención de respuestas válidas y la formulación de interpretaciones apropiadas de los puntajes de la prueba.

Comentario: Cuando proceda, se debe informar a los examinados con antelación sobre el contenido de la prueba, incluyendo el área temática, los temas incluidos y los formatos de ítems. Se debe proporcionar orientación general sobre las estrategias de ejecución de una prueba. Por ejemplo, normalmente se debe informar a los examinandos sobre la conveniencia de omitir respuestas y advertirles de los límites de tiempo impuestos, de manera que puedan gestionar el tiempo de forma adecuada. Para administraciones por computadora, se debe mostrar a los examinandos ejemplos de la interfaz que se tiene previsto usar durante la prueba y se les debe dar la oportunidad de practicar con esas herramientas y dominar su uso antes de que empiece la prueba. Además, se les debe informar sobre las posibilidades de revisar los ítems que se han respondido u omitido anteriormente.

En la mayoría de las situaciones de pruebas, se deberá informar a los examinandos sobre el uso previsto de los puntajes y el grado de confidencialidad de estos , y se les debe comunicar si tendrán acceso a esos resultados y en qué momento. Las excepciones se producen cuando el conocimiento de los propósitos o de los usos previstos de los puntajes viola la integridad de su interpretación, por ejemplo, cuando la prueba se dirige a detectar simulaciones. Si un registro de la sesión de la prueba se guarda en formato escrito, audio, vídeo o cualquier otro, o si se guardan los registros asociados con el evento de la prueba (por ejemplo, la información de calificación), los examinandos tienen derecho a saber qué información de la prueba se divulgará y para qué finalidad se utilizarán los resultados. En algunos casos, se aplican los estándares legales a la información sobre el uso y la confidencialidad de (y el acceso de los examinandos a) los puntajes de la prueba. También se debe informar sobre las políticas relacionadas con las contrapruebas. Se debe advertir a los examinandos contra el comportamiento inadecuado y hacerles saber las consecuencias de las conductas indebidas (por ejemplo, la copia engañosa), que pueden tener como resultado la prohibición de completar la prueba o de recibir los puntajes de la prueba, o que podrían acarrear otras sanciones. Se debe informar a los examinandos, al menos de forma general, si habrá un escrutinio especial de los protocolos de la prueba o de los patrones de puntaje para detectar vulnerabilidades de seguridad, engaños u otros comportamientos inapropiados.

Estándar 8.3

Cuando se ofrece al examinando la opción de elegir el formato de la prueba, se debe proporcionar información sobre las características de cada formato.

Comentario: En ocasiones, los examinandos pueden elegir entre la administración de una prueba con papel y lápiz y la administración computarizada. Algunas pruebas se facilitan en diferentes idiomas. A veces, se ofrece una evaluación

alternativa. Los examinandos deben conocer las características de cada alternativa que esté disponible, de manera que puedan tomar una decisión informada.

Estándar 8.4

Se debe obtener el consentimiento informado de los examinandos, o de sus representantes legales si procede, antes de comenzar la prueba, excepto (a) cuando la evaluación sin consentimiento sea obligatoria por ley o normativa gubernamental, (b) cuando la evaluación se lleve a cabo como parte ordinaria de las actividades escolares, o (c) cuando el consentimiento sea claramente implícito, por ejemplo, en contextos de empleo. Es posible que la ley y los estándares profesionales vigentes requieran el consentimiento informado.

Comentario: El consentimiento informado conlleva que los examinandos o sus representantes tengan conocimiento, en un idioma que puedan comprender, de las razones de la evaluación, los tipos de pruebas que se van a usar, los usos previstos de los resultados de los examinandos u otra información, y de las diversas consecuencias materiales del uso previsto. En general, se recomienda que se solicite directamente a las personas que den su consentimiento formal en lugar de solicitarles únicamente que indiquen si deniegan su consentimiento.

No se requiere de consentimiento cuando la prueba es una obligación legal, como en el caso de una evaluación psicológica por mandato judicial, aunque pueden existir requisitos legales para suministrar información sobre los resultados de la sesión de la prueba a los examinandos. Por lo general, tampoco se requiere el consentimiento en contextos educativos para las pruebas administradas a todos los alumnos. Cuando se requiere una evaluación por motivos de empleo, acreditación o admisiones educativas, los solicitantes otorgan implícitamente su consentimiento al enviar su solicitud. Cuando sea factible, la persona que explique las razones de una prueba debe tener experiencia en la comunicación con los individuos de la población objetivo de la prueba (p. ej., personas con discapacidades o de diferentes procedencias lingüísticas).

Unidad 2. Derechos de los examinandos a obtener acceso a los resultados de sus pruebas y a la protección frente a usos no autorizados de estos resultados

Estándar 8.5

Se deben considerar cuidadosamente las políticas de publicación de puntajes de las pruebas que contienen datos de identificación y comunicarse claramente a quienes tengan acceso a los puntajes. Las políticas deben garantizar que los resultados de las pruebas que contengan nombres de examinandos individuales u otros datos de identificación personal solo se divulguen a quienes tengan un interés profesional legítimo en los examinandos y disponga de autorización para acceder a dicha información bajo las leyes de privacidad vigentes, a quienes se encuentren amparados por documentos de consentimiento informado de los examinandos o a quienes cuente con los permisos legales para obtener acceso a los resultados.

Comentario: Se debe mantener la confidencialidad de los resultados de las pruebas de individuos identificados por el nombre o por algún otro dato que permita identificar fácilmente a una persona, o identificarla rápidamente cuando esa información se combina con otra información. En algunos casos, la información se puede suministrar con carácter confidencial a otros profesionales con un interés legítimo en el caso específico, de modo coherente con las consideraciones legales y éticas, incluyendo, si corresponde, las leyes de privacidad. La información podría facilitarse a investigadores si se cumplen todas las siguientes condiciones: (a) se mantiene la confidencialidad de todos los examinandos, (b) el uso previsto es compatible con la práctica de investigación aceptada, (c) el uso

se lleva a cabo de conformidad con los actuales requisitos legales e institucionales para los derechos del sujeto y con las leyes de privacidad vigentes, y (d) el uso es coherente con los documentos archivados de consentimiento informado del examinando o con las condiciones de consentimiento implícito que sean apropiadas en algunos contextos.

Estándar 8.6

Los datos de la prueba que se mantienen o transmiten en archivos de datos, incluyendo toda la información de identificación personal (no solo los resultados), deben protegerse adecuadamente contra el acceso, uso o divulgación indebidos. Esto incluye protecciones físicas, técnicas y administrativas según sea apropiado para el conjunto específico de datos y sus riesgos, de conformidad con los requisitos legales vigentes. El uso de transmisión por telefax, redes computarizadas, bancos de datos u otros sistemas de procesamiento o transmisión electrónica de datos se deberá restringir a situaciones donde la confidencialidad se pueda garantizar razonablemente. Los usuarios deben desarrollar o seguir políticas, coherentes con los requisitos legales, que especifiquen si los examinandos pueden revisar y corregir la información personal y los métodos para hacerlo.

Comentario: El riesgo se reduce evitando los números o códigos de identificación que están asociados con los individuos y que se usan para otros fines (p. ej., números de la Seguridad Social o identificadores de empleados). Se deben adoptar todas las disposiciones razonables (como el cifrado de los datos) para mantener la confidencialidad de la información si se usa la comunicación por telefax o computadora para transmitir las respuestas de la prueba a otro centro para la calificación o si los puntajes se transmiten de forma similar. En algunos casos, las leyes vigentes de seguridad de datos pueden exigir que se tomen medidas específicas para proteger los datos. En la mayoría de los casos, el propietario de los datos desarrollará estas políticas.

Unidad 3. Derechos de los examinandos a reportes de puntajes imparciales y precisos

Estándar 8.7

Cuando los puntajes de examinandos individuales se asignan en categorías para la presentación de reportes, las etiquetas asignadas a las categorías se deben elegir de forma que reflejen las inferencias previstas y se deben describir con exactitud.

Comentario: Cuando se asocian etiquetas con los resultados de la prueba, se debe prestar atención para evitar etiquetas que puedan tener derivaciones innecesariamente estigmatizantes. Por ejemplo, etiquetas descriptivas como "básico", "competente" y "avanzado" llevarían interpretaciones menos estigmatizantes que términos como "deficiente" o "insatisfactorio". Además, se debe proporcionar información relacionada con la precisión de las clasificaciones de los puntajes (p. ej., la precisión de la decisión y la coherencia de la decisión).

Estándar 8.8

Cuando los puntajes de la prueba se usen para tomar decisiones sobre un examinando o para hacer recomendaciones a un examinando o a un tercero, el examinando debe disponer de acceso oportuno a una copia de cualquier reporte de puntajes e interpretación de la prueba, a menos que se haya renunciado explícitamente a ese derecho en el documento de consentimiento informado del examinando o implícitamente a través del procedimiento de solicitud en evaluaciones educativas, de acreditación o empleo, o esté prohibido por ley o mandato judicial.

Comentario: En algunos casos, es posible que un examinando disponga de la información adecuada cuando el reporte de la prueba se envía a un tercero pertinente (p. ej., el psicólogo o psiquiatra de un tratamiento), quien puede interpretar los resultados del examinando. Cuando se proporciona

al examinando una copia del reporte de la prueba y hay un motivo aparente para creer que los puntajes pueden tener una interpretación incorrecta, el examinador o un tercero informado debe estar disponible para interpretarlos, incluso si el reporte está redactado con claridad, ya que el examinando podría malinterpretar o tener preguntas que el reporte no responda de manera específica. En situaciones de pruebas de empleo, cuando los resultados se usan exclusivamente para decisiones de selección, las renuncias al acceso suelen ser una condición de las solicitudes de empleo, aunque el acceso a la información de la prueba pueda, con frecuencia, exigirse apropiadamente en otras circunstancias.

Unidad 4. Responsabilidades de comportamiento de los examinandos a lo largo de todo el proceso de administración de la prueba

Estándar 8.9

Los examinandos deben entender que la suplantación de examinandos para realizar la prueba, la divulgación del material de la prueba o la participación en cualquier forma de engaño son acciones inaceptables y que tales comportamientos pueden acarrear sanciones.

Comentario: Aunque los Estándares no puedan regular el comportamiento de los examinandos, los examinandos deben ser conscientes de sus responsabilidades personales y legales. Disponer la suplantación del examinando por otra persona constituye un fraude. En las pruebas diseñadas para medir el pensamiento independiente de un examinando, proporcionar respuestas que usen el trabajo de otras personas sin atribución o que hayan sido preparadas por alguien distinto al examinando constituye plagio. La divulgación de material confidencial de la prueba con la finalidad de dar a otros examinandos un conocimiento previo interfiere con la validez de las interpretaciones de los puntajes, y la circulación de ítems de la prueba en formato impreso o electrónico puede constituir una infracción a los derechos de autor. En pruebas de certificación o licenciamiento, tales acciones pueden poner en peligro la salud y la seguridad públicas. En general, la validez de las interpretaciones de los puntajes se verá cuestionada por la divulgación inapropiada de la prueba.

Estándar 8.10

En programas de pruebas educativas y de acreditación, cuando se espera que un reporte de puntaje individual se retrase de forma considerable más allá de un breve periodo de investigación debido a posibles irregularidades (por ejemplo, una posible conducta indebida), se debe notificar al examinando y dar el motivo de la investigación. Se deben tomar las medidas razonables para facilitar la revisión y para proteger el interés del examinando. Una vez finalizada la investigación, se deberá notificar al examinando sobre la resolución.

Estándar 8.11

En programas de pruebas educativas y de acreditación, cuando se considere necesario cancelar o retener el puntaje de un examinando debido a posibles irregularidades en la prueba, incluyendo una posible conducta indebida, se deberá explicar el tipo de evidencia y los procedimientos generales que se usarán para investigar la irregularidad a todos los examinandos cuyos puntajes se vean directamente afectados por la decisión. Se proporcionará a los examinandos una oportunidad razonable para que aporten evidencias de que el puntaje no se debería cancelar o retener. Las evidencias tomadas en consideración para decidir la acción final deberán estar disponibles para el examinando, a petición.

Comentario: Cualquier forma de engaño o comportamiento que reduzca la validez e imparcialidad de las interpretaciones de los resultados de la prueba se deberá investigar con rapidez, adoptando las medidas apropiadas. El puntaje de una prueba se puede cancelar o retener debido a

una posible conducta indebida del examinando o por anomalías que involucren a otras personas, como el robo o contratiempos administrativos. Debe estar disponible un recurso de apelación y se debe comunicárselo a los candidatos cuyos puntajes se podrían enmendar o retener. Algunas organizaciones de evaluación ofrecen la opción de una contraprueba rápida y gratuita o el arbitraje de disputas. La información proporcionada a los examinandos deberá ser lo suficientemente específica para entender la evidencia que se usa para respaldar la alegación de irregularidades en la prueba, pero no tan específica como para divulgar los secretos comerciales o facilitar el engaño.

Estándar 8.12

En programas de pruebas educativas y de acreditación, un examinando tiene derecho a un tratamiento imparcial y a un proceso de resolución razonable, apropiado a las circunstancias específicas, con respecto a los cargos asociados con las irregularidades de la prueba o a las cuestiones planteadas por el examinando relacionadas con la precisión de la calificación o de la clave de calificación. Los examinandos tienen derecho a recibir información sobre cualquier medio de recurso disponible.

Comentario: Cuando se cuestiona o invalida el puntaje de un examinando, o cuando un examinando busca una revisión o reconsideración de su puntaje o de algún otro aspecto del proceso de prueba, calificación o presentación de reportes, el examinando tiene derecho a un proceso metódico para un debate o revisión eficaz de la toma de decisiones del administrador o usuario de la prueba. En función de la magnitud de las consecuencias asociadas con la prueba, este proceso puede incluir desde una revisión interna de todos los datos pertinentes por un administrador hasta una conversación informal con un examinando o una extensa audiencia administrativa. Cuanto mayores sean las consecuencias, mayor será el grado de protecciones procesales que deberán estar disponibles. Los examinandos también deberán conocer los procedimientos para el recurso, las posibles tasas asociadas con los procedimientos de recurso, el tiempo previsto de resolución y cualquier otro asunto importante relacionado, incluyendo las consecuencias para el examinando. Algunos programas de pruebas pueden recomendar que el examinando esté representado por un abogado, aunque posiblemente con gastos a cargo del examinando. En función de las circunstancias y el contexto, los principios de los procedimientos reglamentarios podrían ser pertinentes para el proceso aplicado a los examinandos.

9. DERECHOS Y RESPONSABILIDADES DE LOS USUARIOS DE LA PRUEBA

ANTECEDENTES

Los capítulos anteriores han examinado principalmente las responsabilidades de quienes desarrollan, promueven, evalúan o encargan la administración de pruebas y las responsabilidades de los examinandos. El presente capítulo se centra en las responsabilidades de quienes se pueden considerar los usuarios de la prueba. Los usuarios de la prueba son profesionales que seleccionan los instrumentos específicos o supervisan la administración de la prueba (bajo su propia autoridad o a instancias de otros), así como los demás profesionales que participan activamente en la interpretación y uso de los resultados de la prueba. Esto incluye psicólogos, educadores, empleadores, desarrolladores de pruebas, editores de pruebas y otros profesionales. Dada la dependencia de los resultados de las pruebas en muchos contextos, habitualmente se ha ejercido presión sobre los usuarios de la prueba para que expliquen las decisiones basadas en la prueba y las prácticas de evaluación; en muchos casos, los usuarios de la prueba tienen la obligación legal de documentar la validez y la imparcialidad de esas decisiones y prácticas. Los estándares de este capítulo proporcionan orientación con respecto a los procedimientos de administración de la prueba y la toma de decisiones donde las pruebas juegan un papel. Por lo tanto, el presente capítulo incluye estándares de naturaleza general que se aplican en casi todos los contextos de pruebas.

Estos *Estándares* asumen que un propósito legítimo educativo, psicológico, de acreditación o de empleo justifica el tiempo y el gasto de la administración de la prueba. En la mayoría de los contextos, el usuario comunica este propósito a quienes tienen un legítimo interés en el proceso de medida y posteriormente trasmite las consecuencias del desempeño del examinando a quienes están facultados para recibir la información. En función del contexto de la medición, este grupo puede incluir a examinandos individuales, padres y tutores, educadores, empleados, responsables de las políticas, tribunales o el público general.

La validez y confiabilidad son consideraciones críticas en la selección y uso de las pruebas, y los usuarios de la prueba deben considerar la evidencia de (a) la validez de la interpretación para los usos previstos de los puntajes; (b) la confiabilidad/precisión de los puntajes; (c) la aplicabilidad de los datos normativos disponibles en el manual de la prueba, y (d) las consecuencias potenciales positivas y negativas del uso. También se debe tener en cuenta la literatura de investigación acumulada y, si procede, las características demográficas (p. ej. raza/origen étnico, género, edad, ingresos, antecedentes socioeconómicos, culturales y lingüísticos, educación y otras variables socioeconómicas) del grupo para el cual se elaboró originalmente la prueba y para el cual están disponibles los datos normativos. Los usuarios de la prueba también pueden consultar a los profesionales de medición. El nombre de la prueba por sí solo nunca proporciona información adecuada para decidir su selección.

En algunos casos, la selección de pruebas e inventarios se individualiza para un cliente específico. En otros contextos, todos los participantes realizan una batería predeterminada de pruebas. En ambos casos, los usuarios de la prueba deberán conocer bien los procedimientos administrativos apropiados y serán responsables de entender la evidencia de validación y confiabilidad, y de articular esa evidencia si se presentara la necesidad. Los usuarios de la prueba que supervisan la evaluación y las pruebas son responsables de garantizar que los administradores que administren y califiquen la prueba hayan recibido la capacitación y entrenamiento adecuados para llevar a cabo estas tareas. Se requiere que el usuario de la prueba que interpreta los puntajes e integra las inferencias obtenidas de los puntajes y otros datos relevantes tenga un alto nivel de competencia.

Idealmente, los puntajes de la prueba se interpretan a la luz de los datos disponibles, las propiedades psicométricas de los puntajes, los indicadores de esfuerzo y los efectos de las variables moderadoras y las características demográficas sobre los resultados de la prueba. Debido a que los ítems o tareas de una prueba que fue diseñada para un grupo específico puede introducir varianza irrelevante de constructo cuando se usa con otros grupos, es importante seleccionar una prueba con grupos de referencia demográficamente apropiados para la generabilidad de la inferencia que trata de formular el usuario de la prueba. Cuando una prueba desarrollada y normalizada para un grupo se aplica a otros grupos, las interpretaciones de los puntajes deberán calificarse y presentarse como hipótesis y no como conclusiones. Además, se deberá evaluar la idoneidad de los análisis estadísticos realizados en un solo grupo cuando se generalizan a otras poblaciones de examinandos. El usuario de la prueba debe basarse en cualquier evidencia de investigación existente de la prueba para extraer inferencias apropiadas y debe conocer los requisitos que restringen determinadas prácticas (p. ej. normalización por raza o género en algunos contextos).

Por otra parte, cuando proceda, una interpretación de los puntajes de los examinandos debe tener en cuenta no solo la relación probada entre los puntajes y los criterios, sino también la idoneidad de estos últimos. Los criterios deben someterse a un examen similar al examen de los predictores si se desea entender el grado de congruencia de los constructos subyacentes con las inferencias bajo consideración. Es importante que se reconozcan los datos que no respaldan la inferencia y se concilien o anoten como límites a la confianza que se puede tener en las inferencias. En general, la educación y la experiencia necesarias para interpretar pruebas de grupo son menos estrictas que las cualificaciones necesarias para interpretar pruebas administradas individualmente.

Los usuarios de la prueba deben seguir los procedimientos estandarizados de administración indicados por los desarrolladores de la prueba. La administración computarizada de pruebas también debe seguir los procedimientos estandarizados y se debe proporcionar suficiente supervisión para garantizar la integridad de los resultados. Cuando se requieran procedimientos no estándar, estos se deben describir y justificar. Los usuarios de la prueba también son responsables de facilitar condiciones de evaluación apropiadas. Por ejemplo, es posible que el usuario de la prueba tenga que determinar si un examinando es capaz de leer en el nivel requerido y si un examinando con discapacidad visual, auditiva o neurológica dispone de las adaptaciones adecuadas. El capítulo 3 ("Imparcialidad en las pruebas") trata en detalle las consideraciones y estándares de acceso igualitario.

Cuando se exige la administración de pruebas o el uso de datos de pruebas para una población específica por parte de autoridades gubernamentales, instituciones educativas, consejos de licencias o empleadores, el desarrollador y usuario de un instrumento puede ser básicamente el mismo. A menudo, en tales contextos, no existe una clara separación en términos de responsabilidades profesionales entre quienes desarrollan el instrumento y quienes los administran e interpretan los resultados. Por otra parte, los instrumentos producidos por editores independientes presentan un cuadro diferente. Habitualmente, los utilizarán diferentes usuarios de la prueba con una variedad de poblaciones y para diversos propósitos.

El desarrollador escrupuloso de una prueba estandarizada intentará controlar quién tiene acceso a la prueba y educar a los usuarios potenciales. Asimismo, la mayoría de los editores y promotores de pruebas trabajan para evitar el uso indebido de medidas estandarizadas y la interpretación errónea de puntajes individuales y medias de grupo. Los manuales de la prueba suelen ilustrar las interpretaciones y aplicaciones viables y no viables. Algunos identifican prácticas específicas que no resultan apropiadas y que se deben desaconsejar. Sin embargo, a pesar de los esfuerzos de los desarrolladores de pruebas, es probable que el uso apropiado de una prueba y la interpretación correcta de los puntajes sigan siendo primordialmente responsabilidad del usuario de la prueba.

Los examinandos, padres y tutores, legisladores, responsables de políticas, medios de

comunicación, tribunales y el público en general suelen preferir interpretaciones inequívocas de los datos de una prueba. En particular, tienden a atribuir los resultados positivos o negativos (incluyendo las diferencias de grupos) a un solo factor o a las condiciones que prevalecen en una institución social —en la mayoría de los casos, el hogar o la escuela. Frecuentemente, estos consumidores de datos de pruebas presionan por obtener justificaciones basadas en los puntajes para decisiones que solo se basan parcialmente en los puntajes de las pruebas. Un usuario de la prueba sensato ayudará a todas las partes interesadas a comprender que las decisiones correctas relacionadas con el uso de una prueba y la interpretación de los puntajes incluyen un elemento de juicio profesional. No siempre resulta evidente para los consumidores que la elección de diversos procedimientos de recolección de información implica una experiencia que no se puede cuantificar o verbalizar con facilidad. El usuario puede ayudar a que los consumidores reconozcan el hecho de que la ponderación de datos cuantitativos, la información educativa u ocupacional, las observaciones conductuales, los reportes anecdóticos y otros datos relevantes no siempre se pueden especificar con precisión. No obstante, los usuarios de la prueba deben proporcionar reportes e interpretaciones de los datos de la prueba que sean claros y comprensibles.

Debido a que frecuentemente los resultados de una prueba se reportan de forma numérica, suelen tener una apariencia de precisión, y a veces se tolera que los datos de la prueba anulen otras fuentes de evidencia sobre los examinandos. Hay circunstancias en las que una selección basada exclusivamente en los puntajes de una prueba puede resultar apropiada (p. ej., en el cribado laboral previo). Sin embargo, en contextos educativos, psicológicos, forenses y algunos de empleo, se recomienda a los usuarios de la prueba (y podría ser legalmente obligatorio) que consideren otras fuentes relevantes de información sobre los examinandos y no solo los puntajes. En estas situaciones, los psicólogos, educadores u otros profesionales familiarizados con el contexto local y con los examinandos locales suelen estar mejor cualificados para integrar esta diversa información de manera eficaz.

No es apropiado que estos estándares dicten niveles mínimos de correlación de criterios de pruebas, precisión de clasificación o confiabilidad/precisión para un propósito determinado. Tales niveles dependen de factores como la naturaleza del constructo medido, la edad de los individuos sometidos a la prueba y si las decisiones se deben tomar inmediatamente en base a la mejor evidencia disponible, aunque sea escasa, o si se pueden retrasar hasta que esté disponible una evidencia mejor. Sin embargo, resulta apropiado que los usuarios se cercioren de las alternativas existentes, la calidad y las consecuencias de estas alternativas, y de si un retraso en la toma de decisiones resultaría beneficioso. Como suele pasar en el desarrollo de pruebas, los equilibrios costo-beneficio resultan necesarios en el uso de pruebas. No obstante, en algunos contextos, los requisitos legales pueden establecer límites al grado de tales equilibrios. Como pasa con los estándares para las diversas fases del desarrollo de pruebas, cuando los estándares pertinentes no se cumplen en el uso de la prueba, los motivos deben ser convincentes. Cuanto mayor sea el impacto potencial sobre los examinandos, para bien o para mal, mayor será la necesidad de identificar y satisfacer los estándares pertinentes.

En la selección de una prueba y la interpretación del puntaje, se espera que el usuario de la prueba tenga un conocimiento claro del propósito de la prueba y de sus consecuencias probables. El usuario informado tendrá ideas definidas sobre cómo conseguir estos propósitos y cómo evitar la parcialidad y las consecuencias no deseables. Al suscribir estos *Estándares*, los editores de la prueba y los organismos que encargan el uso de la prueba aceptan proporcionar información sobre los puntos fuertes y débiles de sus instrumentos. Aceptan la responsabilidad de advertir de posibles interpretaciones incorrectas por intérpretes no sofisticados de puntajes individuales o datos agregados. Sin embargo, la responsabilidad última del uso y la interpretación correctos recae principalmente en el usuario de la prueba. Al asumir esta responsabilidad, el usuario deberá adquirir conocimientos

sobre los usos apropiados de la prueba y las poblaciones para las cuales resulta idónea. El usuario de la prueba deberá estar preparado para desarrollar un análisis lógico que respalde las diversas facetas de la evaluación y las inferencias extraídas de los resultados de la evaluación. Los usuarios de la prueba en todos los contextos (p. ej. clínico, de orientación, de acreditación, educativos, empleo, forense, psicológico) también deberán convertirse en expertos en comunicar las implicaciones de los resultados de la prueba a quienes estén facultados para recibir esta información.

En algunos casos, es posible que los usuarios tengan la obligación de recopilar evidencias adicionales sobre la calidad técnica de la prueba. Por ejemplo, si las evaluaciones de desempeño se califican localmente, se podría requerir evidencia del grado de concordancia entre evaluadores. Los usuarios también deben estar atentos a las probables consecuencias locales del uso de la prueba, sobre todo en el caso de programas de pruebas a gran escala. Si se usa el mismo material de pruebas en años sucesivos, los usuarios deberán supervisar activamente el programa para determinar si la reutilización ha puesto en riesgo la integridad de los resultados.

Algunos de los estándares siguientes reiteran ideas incluidas en otros capítulos, sobre todo el capítulo 3 (“Imparcialidad en las pruebas”), el capítulo 6 (“Administración, calificación, presentación de reportes e interpretación de pruebas”), el capítulo 8 (“Derechos y responsabilidades de los examinandos”), el capítulo 10 (“Pruebas y evaluación psicológicas”), el capítulo 11 (“Pruebas y acreditación en el centro de trabajo”) y el capítulo 12 (“Pruebas y evaluación educativas”). La repetición es intencional. Permite la enumeración en un capítulo de las principales obligaciones que debe asumir el administrador y usuario de la prueba, aunque estas responsabilidades pueden hacer referencia a temas que se tratan con mayor detalle en otros capítulos.

ESTÁNDARES PARA LOS DERECHOS Y RESPONSABILIDADES DE LOS USUARIOS DE LA PRUEBA

Los estándares de este capítulo empiezan con un estándar general (con el número 9.0), diseñado para comunicar el propósito central o el enfoque principal del capítulo. El estándar general también se puede ver como el principio rector del capítulo y se aplica a todas las pruebas y a todos los usuarios de la prueba. Todos los estándares posteriores se han dividido en tres unidades temáticas, etiquetadas de la siguiente manera:

1. Validez de las interpretaciones
2. Diseminación de la información
3. Seguridad de la prueba y protección de los derechos de autor

Estándar 9.0

Los usuarios de la prueba son responsables de conocer la evidencia de validación que respalda las interpretaciones previstas de los puntajes de las pruebas que usan, desde la selección de la prueba hasta el uso de puntajes, así como las consecuencias comunes positivas o negativas del uso de la prueba. Los usuarios de la prueba también tienen la responsabilidad legal y ética de proteger la seguridad del contenido de la prueba y la privacidad de los examinandos, y deben proporcionar información pertinente y oportuna a los examinandos y a otros usuarios de la prueba con quienes comparten los puntajes.

Comentario: Los usuarios de la prueba son profesionales que pueden dividirse en varias categorías, incluyendo quienes administran las pruebas y quienes interpretan y usan los resultados de las pruebas. Los usuarios de la prueba que interpretan y usan los resultados son responsables de cerciorarse de que existe una evidencia de validación apropiada que respalde las interpretaciones y usos de los resultados de la prueba. En algunos casos, los usuarios de la prueba también son legalmente responsables de cerciorarse del efecto de sus prácticas de evaluación sobre los subgrupos relevantes y de considerar las medidas apropiadas si se dan consecuencias negativas. Además, aunque a menudo se exige que los usuarios de la prueba compartan los resultados con los examinandos y otros grupos de usuarios de la prueba, deben recordar que se debe proteger el contenido de la prueba para mantener la integridad de los puntajes, y que los examinandos tienen expectativas razonables de privacidad, las cuales podrían estar especificadas en determinadas leyes y normativas estatales y federales.

Unidad 1. Validez de las interpretaciones

Estándar 9.1

La responsabilidad por el uso de la prueba se debe asumir por, o delegar a, aquellas personas que tengan la capacitación, las acreditaciones profesionales o la experiencia necesarias para gestionar esta responsabilidad. Se deben satisfacer todas las cualificaciones especiales para la administración o interpretación especificadas en el manual de la prueba.

Comentario: Los usuarios de la prueba solo deben interpretar los puntajes de los examinandos cuyas necesidades o características especiales están dentro del ámbito de las cualificaciones de los usuarios de la prueba. Este estándar tiene una importancia especial en áreas como las pruebas clínicas, forenses y de personalidad, las pruebas de educación especial, las pruebas de personas con discapacidades o con exposición limitada a la cultura dominante, las pruebas de estudiantes de inglés y en otras situaciones donde el impacto potencial es muy significativo. Cuando la situación o el grupo de examinandos quedan fuera de la experiencia del usuario, se debe obtener asistencia. Un número de organizaciones profesionales cuentan con códigos de ética que especifican las cualificaciones necesarias de

quienes administran pruebas e interpretan puntajes dentro del ámbito de práctica de esas organizaciones. En última instancia, el profesional es el responsable de garantizar que se cumplen los requisitos de capacitación clínica, los códigos éticos y los estándares legales para la administración e interpretación de pruebas.

Estándar 9.2

Antes de la adopción y uso de una prueba publicada, el usuario de la prueba deberá estudiar y evaluar los materiales suministrados por el desarrollador de la prueba. Son de especial importancia los materiales que resumen los propósitos de la prueba, especifican los procedimientos de la administración, definen las poblaciones objetivo de examinandos y examinan las interpretaciones de los puntajes con datos de validez y confiabilidad/precisión disponibles.

Comentario: Una premisa para el uso correcto de la prueba es el conocimiento de los materiales que complementan el instrumento. Como mínimo, esto incluye los materiales suministrados por el desarrollador de la prueba. Idealmente, el usuario debe estar familiarizado con los estudios pertinentes recogidos en la literatura profesional y debe tener la capacidad de discriminar entre pruebas apropiadas e inapropiadas para el uso previsto con la población objetivo. El nivel de confiabilidad/precisión del puntaje y los tipos de evidencia de validación requeridos para las interpretaciones de puntaje idóneas depende de la función de la prueba en el proceso de evaluación y el impacto potencial del proceso en las personas participantes. El usuario de la prueba debe conocer las restricciones legales que pueden limitar el uso de la prueba. En ocasiones, el juicio profesional puede llevar al uso de instrumentos con escasa evidencia de validación de las interpretaciones de los puntajes para el uso elegido. En estas situaciones, el usuario no debería suponer que los puntajes, decisiones o inferencias se basan en una evidencia bien documentada con respecto a la confiabilidad o validez.

Estándar 9.3

El usuario de la prueba debe tener una justificación clara para los usos previstos de un procedimiento de prueba o evaluación en términos de validez de las interpretaciones basadas en los puntajes y la contribución que hagan los puntajes al proceso de evaluación y toma de decisiones.

Comentario: El usuario de la prueba debe ser claro en lo que respecta a los motivos por los que administra una prueba. En otras palabras, la justificación de la función de cada instrumento en la selección, diagnóstico, clasificación y toma de decisiones debe presentarse antes, y no después, de la administración de la prueba. En algunos casos, los argumentos de referencia proporcionan la justificación de la elección de las pruebas, inventarios y procedimientos de diagnóstico que se van a utilizar, y la justificación también puede estar respaldada por materiales impresos preparados por el editor de la prueba. Asimismo, la justificación puede provenir de otras fuentes, por ejemplo, la literatura empírica.

Estándar 9.4

Cuando una prueba se va a utilizar para un propósito que tiene poca o ninguna evidencia de validación disponible, el usuario es responsable de documentar la justificación de la selección de la prueba y de obtener la evidencia de confiabilidad/precisión de los puntajes de la prueba y la validez de las interpretaciones que respaldan el uso de los puntajes para ese propósito.

Comentario: La persona que use los puntajes de la prueba para propósitos que no han sido específicamente recomendados por el desarrollador de la prueba es responsable de recopilar la evidencia de validación necesaria. En ocasiones, el respaldo de tales usos puede encontrarse en la literatura profesional. Si no es suficiente una evidencia anterior, se deberán recopilar datos adicionales a lo largo del tiempo a medida que se use la prueba. Las disposiciones de este estándar no se deben considerar como una prohibición de la generación de hipótesis sobre los datos de la prueba. No

obstante, estas hipótesis se deben etiquetar claramente como provisionales. Las partes interesadas deben tener conocimiento de las limitaciones potenciales de los puntajes de la prueba en tales situaciones.

Estándar 9.5

Los usuarios de la prueba deben estar atentos a la posibilidad de errores de puntaje y deben tomar las medidas apropiadas cuando se sospeche la existencia de errores.

Comentario: Los costos de los errores de puntaje son altos, sobre todo en los programas de pruebas de alto riesgo. En algunos casos, el examinando puede solicitar una nueva calificación. Si este derecho del examinando se reconoce en materiales publicados, debe respetarse. Sin embargo, los usuarios de la prueba no deben depender de que los examinandos sean quienes los alerten de la posibilidad de errores de puntaje. Cuando sea factible, la supervisión de la precisión de los puntajes deberá ser una responsabilidad de rutina de los administradores de un programa de pruebas y se debe llevar a cabo una recalificación cuando se sospeche la existencia de errores.

Estándar 9.6

Los usuarios de la prueba deben estar atentos a potenciales interpretaciones erróneas de los puntajes de la prueba; deberán adoptar medidas para minimizar o evitar las interpretaciones erróneas previsibles y los usos inapropiados de los puntajes.

Comentario: Audiencias no capacitadas pueden adoptar interpretaciones simplistas de los resultados de una prueba o pueden atribuir los puntajes altos, bajos o promedios a factores causales únicos. A veces, los usuarios de la prueba pueden anticipar tales interpretaciones erróneas y deben intentar evitarlas. Por supuesto, no es posible anticipar todas las interpretaciones no deseadas y pueden producirse consecuencias negativas imprevistas. Lo que se requiere es un esfuerzo razonable para propiciar interpretaciones y usos válidos y para solventar cualquier consecuencia negativa que se pueda producir.

Estándar 9.7

Los usuarios de la prueba deben verificar periódicamente que sus interpretaciones de los datos de la prueba siguen siendo apropiadas frente a cualquier cambio relevante en la población de examinandos, los métodos de administración o los propósitos de la evaluación.

Comentario: A lo largo del tiempo, un cambio gradual de las características de una población de examinandos puede afectar de forma significativa a la precisión de las inferencias extraídas de medias grupales. Las modificaciones en la administración de la prueba en respuesta a circunstancias imprevistas también pueden afectar a las interpretaciones.

Estándar 9.8

Cuando los resultados de una prueba se comunican al público o a los responsables de políticas, los responsables de la comunicación deben proporcionar y explicar cualquier información complementaria que pueda minimizar posibles interpretaciones erróneas de los datos.

Comentario: Los usuarios de la prueba tienen la responsabilidad de reportar los resultados de manera que faciliten las interpretaciones previstas para los usos propuestos de los puntajes, y esta responsabilidad se extiende más allá del examinando individual y llega a los grupos o individuos a quienes se proporciona los puntajes. Los usuarios de la prueba en situaciones de evaluaciones grupales son responsables de garantizar que los individuos que usan los resultados de la prueba estén capacitados para interpretar correctamente los puntajes. Presentaciones preliminares antes de la publicación de los resultados pueden dar a periodistas, responsables de políticas o miembros el público la oportunidad de asimilar los datos fundamentales. A menudo, la interpretación errónea puede ser el resultado de una presentación inadecuada de la información relevante para la interpretación de los puntajes.

Estándar 9.9

Cuando el usuario de una prueba considera una alteración en el formato, el modo de administración, las instrucciones o el idioma utilizado para administrar una prueba, el usuario debe disponer de una justificación sólida y de evidencia empírica, cuando sea posible, para concluir que la confiabilidad/precisión de los puntajes y la validez de las interpretaciones basadas en los puntajes no se verán afectados.

Comentario: En algunos casos, puede preverse razonablemente, sin evidencia, que cambios menores en el formato o el modo de administración van a tener poco o ningún efecto en los puntajes de las pruebas, las decisiones de clasificación o la idoneidad de las normas. Sin embargo, en otros casos, los cambios en el formato o en los procedimientos administrativos pueden tener efectos considerables en la validez de las interpretaciones de los puntajes, esto es, que modifiquen o cambien el constructo bajo evaluación. Si se generaliza una determinada modificación, se deberá recopilar la evidencia de validación; si resulta apropiado, también deberán desarrollarse las normas bajo las condiciones modificadas.

Estándar 9.10

Los usuarios de la prueba no deben depender exclusivamente de interpretaciones generadas por computadora de los resultados de la prueba.

Comentario: El usuario de servicios de calificación y presentación de reportes generados automáticamente tiene la obligación de familiarizarse con los principios que sirven de base a esas interpretaciones. Todos los usuarios que formulan inferencias y toman decisiones sobre la base de esos reportes deben tener la capacidad de evaluar una interpretación de puntaje generada por computadora a la luz de otra evidencia pertinente de un examinando. Los reportes narrativos automatizados no sustituyen al juicio profesional sólido y pueden ser equívocos si se usan de forma aislada.

Estándar 9.11

Cuando las circunstancias requieren que una prueba se administre en el mismo idioma para todos los examinandos de una población lingüísticamente diversa, el usuario de la prueba debe investigar la validez de las interpretaciones de los puntajes de los examinandos con competencia limitada en el idioma de la prueba.

Comentario: la prueba podría medir erróneamente el rendimiento, las competencias y las características de los examinandos que no tienen el idioma de la prueba como lengua principal, incluso si la administración de una prueba alternativa es legalmente aceptable. La práctica correcta requiere una evaluación continua de los datos para proporcionar evidencia que respalde el uso de la prueba con todos los grupos lingüísticos o evidencia para cuestionar el uso de la prueba cuando la competencia en el idioma no es relevante.

Estándar 9.12

Cuando uno de los propósitos principales de la prueba es describir el estado de una población local, regional o específica de examinandos, deben seguirse de forma estricta los criterios de inclusión o exclusión de los individuos.

Comentario: Pueden darse resultados sesgados por la exclusión de subgrupos específicos de examinandos. Por lo tanto, las decisiones de exclusión o inclusión de examinandos deben basarse en la representación apropiada de la población.

Estándar 9.13

En contextos educativos, clínicos o de orientación, el puntaje de un examinando no se debe interpretar de forma aislada; se debe considerar otro tipo de información pertinente que pueda llevar a explicaciones alternativas del desempeño del examinando en la prueba.

Comentario: No es factible ni necesario realizar una revisión intensiva de los puntajes de cada uno de los examinandos. En algunos contextos, la

información colateral puede ser escasa o no existir en absoluto. Sin embargo, en contextos de orientación, clínicos o educativos, a veces se encuentra disponible abundante información pertinente. Las explicaciones alternativas evidentes de puntajes bajos pueden incluir una baja motivación, fluidez limitada en el idioma de la prueba, oportunidad limitada de aprendizaje, escasa familiaridad con conceptos culturales en los que se basan los ítems y discapacidad perceptual o motora. El usuario de la prueba corrobora los resultados de la evaluación con información adicional de una variedad de fuentes, por ejemplo, entrevistas y resultados de otras pruebas (p. ej., para examinar el concepto de confiabilidad del desempeño a lo largo del tiempo o de varias pruebas). Cuando una inferencia se basa en un solo estudio o en estudios con muestras no representativas de los examinandos, el usuario de la prueba deberá tener mayor cautela con respecto a la inferencia formulada. En contextos clínicos o de orientación, el usuario de la prueba no debe pasar por alto el grado de funcionamiento del examinando en la vida diaria. Si las pruebas se administran mediante computadoras y otros dispositivos electrónicos o a través de Internet, los usuarios de la prueba seguirán teniendo la responsabilidad de proporcionar respaldo a la interpretación de los puntajes, incluyendo consideraciones de explicaciones alternativas cuando sea apropiado.

Estándar 9.14

Los usuarios de la prueba deben informar a los individuos que puede necesitar adecuaciones en la administración (p. ej. adultos mayores, examinandos con discapacidades o estudiantes de inglés) sobre la disponibilidad de las adecuaciones y, cuando se requieran, deben asegurarse de que estas adecuaciones estén disponibles de forma apropiada.

Comentario: Las adecuaciones apropiadas dependen de la naturaleza de la prueba y las necesidades de los examinandos, y deben estar en consonancia con la documentación proporcionada con la prueba. Los usuarios de la prueba deben informar a los examinandos sobre la disponibilidad de las adecuaciones. La responsabilidad de solicitar las adecuaciones y de proporcionar documentación que respalde sus solicitudes puede recaer entonces en los examinandos o en sus tutores. Los usuarios de la prueba deberán tener la capacidad de especificar la información o evidencia (p. ej., manual de la prueba, estudio de investigación) usada para optar por una adecuación apropiada.

Unidad 2. Diseminación de la información

Estándar 9.15

Se debe informar a quienes tienen un interés legítimo en una evaluación sobre los propósitos de esta, cómo se administrarán las pruebas, los factores considerados en la calificación de las respuestas de los examinandos, cómo se usarán los puntajes, durante cuánto tiempo de retendrán los registros y a quién y bajo qué condiciones se divulgarán.

Comentario: Los individuos con un interés legítimo en los resultados de la evaluación incluyen, entre otros, a los examinandos, los padres o tutores, los educadores y los magistrados. Este estándar tiene un mayor grado de relevancia y aplicación para la evaluación educativa y clínica que para la evaluación laboral. En la mayoría de los usos de las pruebas para la selección de solicitantes de empleo y programas educativos, para la concesión de licencias profesionales y de acreditaciones, o para medir el rendimiento, los propósitos de la evaluación y los usos previstos de los puntajes resultan evidentes para los examinandos. Sin embargo, se recomienda comunicar esta información al menos brevemente en estos contextos. No obstante, en algunas situaciones, es posible que la justificación de la evaluación solo quede clara para relativamente pocos examinandos. En tales contextos, puede ser necesario un análisis más detallado y explícito. La retención de registros, los requisitos de seguridad y la privacidad de los registros suelen regirse por requisitos legales

o prácticas institucionales, incluso en situaciones donde la divulgación de registros sería claramente beneficiosa para los examinandos. Antes de la evaluación, cuando proceda, el usuario de la prueba deberá comunicar al examinando quien va a tener acceso a los resultados y al reporte escrito, de qué manera se compartirán los resultados con el examinando y si los resultados se van a compartir con un tercero o el público, y en qué condiciones (p. ej. en procesos judiciales).

Estándar 9.16

A menos que las circunstancias demanden claramente que los resultados de la prueba se retengan, el usuario de la prueba tiene la obligación de proporcionar un reporte oportuno de los resultados al examinando y a otros facultados para recibir esta información.

Comentario: Con frecuencia, la naturaleza de los reportes de puntajes viene dictada por consideraciones prácticas. En algunos casos (p. ej., con algunas certificaciones o pruebas de empleo), solo puede ser factible un breve reporte escrito. En otros casos, podría ser deseable facilitar tanto un reporte oral como un reporte escrito. La interpretación debe variar de acuerdo con el nivel de sofisticación del destinatario. Cuando el examinando es un niño, son los padres o tutores quienes suelen recibir una explicación de los resultados. Cuando las pruebas se administran para selección o promoción de personal, o en otras circunstancias específicas, no siempre se suministra un comentario en forma de reporte o interpretación del puntaje. En algunos casos, las leyes de privacidad estatales o federales pueden regir el alcance y los destinatarios de la información divulgada.

Estándar 9.17

Si un examinando o usuario de la prueba tiene dudas sobre la integridad de los puntajes de los examinandos, el usuario de la prueba debe informar al examinando de sus derechos pertinentes, incluyendo la posibilidad de apelación y representación letrada.

Comentario: Es posible que los monitores de los programas de pruebas de admisión o licenciamiento reporten irregularidades en el proceso de administración de la prueba que se traduzcan en cuestionamientos por parte de los examinandos (p. ej., alarma de incendios en un edificio o fallo temporal del acceso a Internet). Cuando los puntajes de las pruebas sean manifiestamente incoherentes con la información de otros candidatos, los usuarios de la prueba (p. ej., funcionarios de admisión a la universidad) podrían plantear otros cuestionamientos. Los examinandos deberán ser informados de sus derechos en tales situaciones, si los hubiere.

Estándar 9.18

Los usuarios de la prueba deben explicar a los examinandos las oportunidades, si las hubiere, para repetir un examen; los usuarios también deben indicar si se reportará algún puntaje previo o posterior a las personas facultadas para recibir los reportes de puntajes.

Comentario: Algunos programas de pruebas permiten a los examinandos repetir un examen varias veces, cancelar los puntajes u ocultar los puntajes a destinatarios potenciales. Se debe informar a los examinandos y a otros destinatarios de puntajes de tales privilegios, si los hubiese, y de las condiciones bajo las que se aplican.

Estándar 9.19

Los usuarios de la prueba tienen la obligación de proteger la privacidad de los examinandos y las instituciones que participan en un programa de pruebas, a menos que se acuerde la divulgación de información privada o esté autorizada por la ley de manera específica.

Comentario: La protección de la privacidad de los examinandos individuales es un principio bien establecido en la medición psicológica y educativa. El almacenamiento y la transmisión de este tipo de información deben cumplir los estándares legales y profesionales vigentes, y se debe extremar la precaución para proteger la confidencialidad

de los puntajes y la información complementaria (p. ej., la condición de discapacidad). En algunos casos, los usuarios de la prueba y los organismos de evaluación pueden adoptar restricciones más estrictas de las que dicta la ley con respecto a la comunicación y uso compartido de los resultados de las pruebas. Es posible que se apliquen las leyes de privacidad a determinados tipos de información y, a veces, los códigos de ética adoptados por organizaciones profesionales pueden contener estándares similares o más estrictos. En algunos programas de pruebas las condiciones para la divulgación se indican al examinando antes de la evaluación, y hacer la prueba puede constituir la aceptación de la divulgación de los puntajes correspondientes, de la manera en que se especifique. En otros programas, el examinando (o sus padres o tutores) deberán aceptar formalmente cualquier divulgación de información de la prueba a individuos u organismos que no se hayan especificado en la literatura publicada del administrador. Es posible que las leyes de privacidad vigentes, si las hubiere, rijan y permitan (como en el caso de los distritos escolares para fines de rendición de cuentas) o prohíban (como en contextos clínicos) la divulgación de la información de la prueba. Se debe señalar que, con frecuencia, la ley garantiza el derecho del público y los medios a examinar los resultados agregados de las pruebas del sistema público de educación. Esto suele incluir los puntajes de las pruebas desagregados por subgrupos demográficos cuando los números son suficientes para generar resultados estadísticamente válidos y para evitar la identificación de los examinandos individuales.

Estándar 9.20

En situaciones donde los resultados de la prueba se comparten con el público, los usuarios de la prueba deben formular y compartir la política establecida relacionada con la publicación de los resultados (p. ej., pertinencia temporal, nivel de detalle) y aplicar esa política a lo largo del tiempo de forma sistemática.

Comentario: Los desarrolladores y usuarios de la prueba deben considerar las prácticas de las comunidades a las que sirven y facilitar la creación de políticas comunes relacionadas con la publicación de los resultados. Por ejemplo, en muchos estados, la publicación de datos de pruebas educativas a gran escala suele ser una exigencia legal. Sin embargo, incluso cuando no se requiere la publicación de datos, pero se realiza de forma rutinaria, los usuarios de la prueba deben tener políticas claras que rijan los procedimientos de publicación. Diferentes políticas sin las justificaciones apropiadas pueden confundir al público y causar controversias innecesarias.

Unidad 3. Seguridad de la prueba y protección de los derechos de autor

Estándar 9.21

Los usuarios de la prueba tienen la responsabilidad de proteger la seguridad de las pruebas, incluyendo la de ediciones anteriores.

Comentario: Cuando las pruebas se usan para fines de selección, acreditación, rendición de cuentas en el ámbito educativo, o para diagnóstico, tratamiento y monitorización clínicos, resulta esencial la protección rigurosa de la seguridad de la prueba por motivos relacionados con la validez de las inferencias extraídas, la protección de los derechos de propiedad intelectual y los costos asociados con el desarrollo de pruebas. Los desarrolladores y editores de pruebas, y los individuos titulares de los derechos de autor de las pruebas, proporcionarán directrices específicas sobre la seguridad de la prueba y la eliminación de los materiales de la prueba. El usuario de la prueba tiene la responsabilidad de garantizar la seguridad de los materiales de la prueba de acuerdo con las directrices profesionales establecidas para la prueba, así como con los estándares legales vigentes. La reventa de materiales protegidos por derechos de autor en foros abiertos es una violación de este estándar, y las grabaciones de audio y vídeo para fines de entrenamiento se deben gestionar de manera que no se divulguen al público. Estas prohibiciones también se aplican a las ediciones anteriores de

la prueba; los usuarios de la prueba deberán ayudar a garantizar que los materiales se eliminen de forma segura cuando ya no estén en uso (p. ej., en el momento de su retirada o después de la compra de una nueva edición). En tales situaciones, la coherencia y la claridad de la definición de prácticas aceptables y no aceptables resultan esenciales. Cuando las pruebas se vean involucradas en litigios, se debe restringir la inspección de los instrumentos (en la medida en que lo permita la ley) a quienes tengan la obligación de salvaguardar la seguridad de la prueba por imperativo legal o ética profesional.

Estándar 9.22

Los usuarios de la prueba tienen la responsabilidad de respetar los derechos de autor de la prueba, incluyendo los derechos de autor de pruebas que se administren mediante dispositivos electrónicos.

Comentario: Por ley y ética, los usuarios de la prueba no pueden reproducir o crear versiones electrónicas de materiales protegidos por derechos de autor para usos rutinarios sin el consentimiento del titular de los derechos de autor. Estos materiales (tanto en formato papel como electrónico) incluyen ítems de la prueba, protocolos de la prueba, formularios complementarios como hojas de respuestas y formularios de perfiles, plantillas de calificación, tablas de conversión de puntajes brutos a puntajes reportados y tablas de normas. El almacenamiento y transmisión de la información de la prueba debe cumplir los estándares legales y profesionales vigentes.

Estándar 9.23

Los usuarios de la prueba deben recordar a todos los examinandos, incluyendo a aquellos que realizan pruebas administradas electrónicamente, y a otras personas que puedan tener acceso a los materiales de la prueba, que las políticas y normativas sobre derechos de autor pueden prohibir la divulgación de los ítems de la prueba sin autorización específica.

Comentario: En algunos casos, la información sobre los derechos de autor y las prohibiciones sobre la divulgación de los ítems de la prueba se proporcionan en formato escrito o verbal como parte del procedimiento previo al inicio de una prueba o como parte de los procedimientos de administración. No obstante, incluso en los casos en que esta información no es parte formal de la administración de la prueba, si los materiales están protegidos por derechos de autor, los usuarios de la prueba deberán informar a los examinandos de sus responsabilidades en esta área.

PARTE III

III

Aplicaciones de las pruebas

10. PRUEBAS Y EVALUACIÓN PSICOLÓGICAS

ANTECEDENTES

Este capítulo aborda temas importantes para los profesionales que usan pruebas psicológicas para la evaluación de individuos. Los temas que se tratan en este capítulo incluyen la selección y administración de pruebas, la interpretación de puntajes de pruebas, el uso de información colateral en la evaluación psicológica, los tipos de pruebas y los propósitos de la evaluación psicológica. En este capítulo, se revisan pruebas psicológicas de tipo cognitivo y neuropsicológico, de conducta problemática, de familias y parejas, de comportamientos sociales y de adaptación, de personalidad y vocacionales. Además, el capítulo incluye una descripción general de cinco usos comunes de pruebas psicológicas: diagnóstico; evaluación psicológica; evaluación neuropsicológica; planificación de intervención y evaluación de resultados; decisiones judiciales y gubernamentales; y conciencia personal, identidad social, y salud, desarrollo y acción psicológicos. Los estándares de este capítulo se aplican a contextos donde se llevan a cabo evaluaciones en profundidad de personas, ya sea de forma individual o grupal. Las pruebas psicológicas se usan también en otros contextos, sobre todo en contextos educativos y de empleo. Las pruebas diseñadas para medir características específicas de candidatos relacionadas con el trabajo para fines de selección se tratan en el texto y los estándares del capítulo 11; las pruebas usadas en contextos educativos se abordan en profundidad en el capítulo 12.

Es crucial que los profesionales que usen pruebas para realizar evaluaciones de individuos tengan conocimiento de los factores educativos, lingüísticos, nacionales y culturales, así como capacidades físicas que influyen en (a) el desarrollo de un examinando, (b) los métodos para obtener y transmitir información, y (c) la planificación e implementación de las intervenciones. Por lo tanto, se recomienda a los lectores a revisar el capítulo 3, que trata de la imparcialidad en las pruebas; el capítulo 8, que se centra en los derechos de los examinandos; y el capítulo 9, que se centra en los derechos y responsabilidades de los usuarios de la prueba. En los capítulos 1, 2, 4, 5, 6 y 7, los lectores encontrarán detalles adicionales importantes sobre la validez; la confiabilidad y la precisión; el desarrollo de pruebas; el escalamiento y la equiparación; la administración, calificación, presentación de reportes e interpretación; y sobre la documentación de respaldo.

El uso de pruebas psicológicas proporciona un método para la recolección de información dentro de un marco más amplio de evaluación psicológica de un individuo. Por lo general, las evaluaciones psicológicas implican una interacción entre un profesional capacitado y con experiencia en evaluaciones, el examinando y un cliente que puede ser el propio examinando o un tercero. El examinando puede ser un niño, un adolescente o un adulto. Habitualmente, el cliente es la persona u organismo que organiza la evaluación. Los clientes pueden ser pacientes, personas bajo orientación, padres, niños, empleados, empleadores, abogados, estudiantes, organismos del gobierno u otras partes responsables. Los contextos donde se usan pruebas o inventarios psicológicos incluyen, entre otros, jardines de infancia, escuelas infantiles, primarias y secundarias, facultades y universidades, contextos de preselección de empleo, hospitales, prisiones, clínicas de salud y de salud mental, y otros centros profesionales.

Las tareas que comportan una evaluación psicológica (recopilación, evaluación, integración y reporte de información saliente significativa para los aspectos del funcionamiento del examinando sometido a examen) incluyen un conjunto de actividades profesionales complejas y sofisticadas. Una evaluación psicológica se lleva a cabo para responder preguntas específicas sobre el funcionamiento psicológico o el comportamiento de un examinando durante un intervalo de tiempo

concreto o para predecir un aspecto del funcionamiento psicológico o el comportamiento en el futuro. Debido a que típicamente los puntajes de las pruebas se interpretan en el contexto de otra información sobre el examinando, la evaluación psicológica de un individuo también incluye, por lo general, entrevistas al examinando, la observación de su comportamiento en el contexto apropiado, la revisión de sus registros educativos, sanitarios, psicológicos y otros que resulten pertinentes, y la integración de estos hallazgos con otro tipo de información que terceros podrían proporcionar. Los resultados de las pruebas e inventarios usados en las evaluaciones psicológicas pueden ser útiles para que los profesionales entiendan más cabalmente a los examinandos y formulen hipótesis, inferencias y decisiones más informadas sobre los aspectos del funcionamiento psicológico de los examinandos o las intervenciones apropiadas.

La interpretación de los puntajes de las pruebas e inventarios puede ser una parte valiosa del proceso de evaluación y, si se utiliza adecuadamente, puede aportar información útil a los examinandos y a otros usuarios de la interpretación. Por ejemplo, los resultados de las pruebas e inventarios se pueden usar para evaluar el funcionamiento psicológico de un individuo; asignar una clasificación de diagnóstico; detectar y caracterizar deterioros neuropsicológicos, retrasos de desarrollo y discapacidades de aprendizaje; para determinar la validez de un síntoma; evaluar los puntos fuertes cognitivos o de personalidad, o los problemas de salud mental o de comportamiento emocional; para evaluar intereses y valores; determinar estadios de desarrollo o valorar los resultados de un tratamiento. Los resultados de las pruebas también pueden proporcionar información que se use para la toma de decisiones que tienen un impacto significativo y duradero en las vidas de las personas (p. ej., decisiones educativas o vocacionales, diagnóstico, planes de tratamiento, incluyendo planes de intervenciones psicofarmacológicas, evaluaciones de intervenciones y resultados, decisiones de libertad condicional, compromiso civil, custodia infantil, competencia para ser juzgado, litigios por daños personales y decisiones de pena de muerte).

Selección y administración de pruebas

La selección y administración de pruebas e inventarios psicológicos suele ser individualizada para cada participante. Sin embargo, en algunos contextos se pueden realizar pruebas predeterminadas para todos los participantes, y las interpretaciones de los resultados se podrían proporcionar en un escenario de grupo.

El proceso de evaluación empieza clarificando, en la medida de lo posible, los motivos por los cuales se evalúa al examinando. Estos motivos y otros intereses pertinentes guían la selección de las pruebas, inventarios y procedimientos de diagnóstico que se van a usar, así como la identificación de otras fuentes de información necesarias para la evaluación del examinando. Las conclusiones preliminares pueden llevar a la selección de pruebas adicionales. El profesional tiene la responsabilidad de familiarizarse con la evidencia de validación de los usos previstos de los puntajes de las pruebas e inventarios seleccionados, incluyendo las pruebas online o administradas por computadora. Durante la selección de la prueba, también se debe considerar la evidencia de confiabilidad/ precisión de los puntajes y la disponibilidad de los datos normativos aplicables en la literatura de investigación acumulada de la prueba. En el caso de pruebas que hayan sido revisadas, generalmente se deben seleccionar las ediciones que tienen el respaldo actual del editor. En ocasiones, resulta apropiado el uso de una edición anterior de un instrumento (p. ej., cuando se lleva a cabo una investigación longitudinal o cuando una edición previa contiene subpruebas pertinentes no incluidas en una edición posterior). Además, los profesionales tienen la responsabilidad de vigilar la dependencia respecto de puntajes de pruebas que sean obsoletas; en tales casos, resultarán adecuadas las contrapruebas. En usos internacionales, es especialmente importante verificar que el constructo que se evalúa tiene un significado equivalente en los distintos contextos culturales y por encima de las fronteras nacionales.

Las consideraciones de validez y confiabilidad/ precisión son esenciales, pero las características demográficas de los grupos para los cuales la

prueba se diseñó originalmente y que tienen disponibles datos normativos iniciales y posteriores son también consideraciones importantes para la selección de pruebas. Seleccionar una prueba con grupos normativos apropiados demográfica y clínicamente, pertinentes para el examinando y para el propósito de la evaluación, es importante para la generabilidad de las inferencias que los profesionales tratan de formular. Es posible que no sea apropiado aplicar a otros grupos una prueba construida para un grupo concreto. Si la prueba se usa, las interpretaciones de los puntajes se deberán clasificar y presentar como hipótesis y no como conclusiones.

Las pruebas y los inventarios que cumplen los exigentes estándares técnicos de calidad son una condición necesaria, pero no suficiente, para una administración y calificación de pruebas responsable, y para la interpretación y uso de los puntajes. Un profesional que lleva a cabo una evaluación psicológica debe disponer de una capacitación y entrenamiento completos y apropiados, adquirir las acreditaciones adecuadas, adherirse a las directrices éticas profesionales y tener un alto grado de juicio profesional y de conocimientos científicos.

Los profesionales que supervisan las pruebas y la evaluación deben ser expertos en los procedimientos correctos de administración de pruebas. Son los responsables de garantizar que todas las personas que administran y califican las pruebas hayan recibido la capacitación y entrenamiento adecuados para llevar a cabo las tareas asignadas. Los administradores de pruebas deben administrar las pruebas tal como se indica en los manuales de las pruebas y deben adherirse a los estándares éticos y profesionales. Por lo general, la educación y la experiencia necesarias para administrar pruebas de grupo o para monitorizar pruebas administradas por computadora son menos extensas que las cualificaciones necesarias para administrar e interpretar puntajes de pruebas administradas individualmente, que requieren interacciones entre el examinando y el administrador de la prueba. En muchas situaciones donde se requieren observaciones de conducta complejas, es posible que no sea apropiado el uso de no profesionales para administrar o calificar las pruebas. Antes de comenzar el proceso de evaluación, el usuario de la prueba o la parte responsable (p. ej., el padre, o tutor legal) deberá saber quién va a tener acceso a los resultados de la prueba y al reporte escrito, de qué manera se compartirán los resultados con el examinando, y si las decisiones que se basan en los resultados se van a compartir con el examinando, un tercero o el público y cuándo (p. ej. en procesos judiciales).

Los administradores de pruebas deben ser conscientes de las limitaciones personales que afectan a su capacidad de administrar y calificar una prueba de manera precisa e imparcial. Estas limitaciones pueden incluir factores físicos, perceptuales y cognitivos. Algunas pruebas presentan exigencias considerables a los administradores (p. ej., registrar las respuestas rápidamente, manipulación de equipos o ejecución de ítems complejos durante la administración). Los administradores de pruebas que no pueden cumplir de forma cómoda estas exigencias no deben administrar tales pruebas. Para las pruebas que requieren instrucciones orales antes o durante su administración, los administradores deberán asegurarse de que no existen barreras a una clara comprensión por parte de los examinandos.

Cuando se usan baterías de pruebas, el profesional debe determinar el orden apropiado de las pruebas que se administran. Por ejemplo, cuando se administren pruebas cognitivas o neuropsicológicas, algunos profesionales administran primero las pruebas que evalúan dominios básicos (p. ej., la atención) y termina con pruebas que evalúan dominios más complejos (p. ej., funciones ejecutivas). Los profesionales también tienen la responsabilidad de establecer las condiciones de la evaluación que sean apropiadas para las necesidades y capacidades de los examinandos. Por ejemplo, es posible que el examinador tenga que determinar si un examinando es capaz de leer en el nivel requerido y si las discapacidades visuales, auditivas, psicomotoras o clínicas o los déficits neurológicos cuentan con las adecuaciones correctas. El capítulo 3 trata en detalle las consideraciones y estándares relacionados con el acceso.

La administración estandarizada no es necesaria para todas las pruebas, pero es importante

para la interpretación de los puntajes en muchas pruebas y propósitos. En esas situaciones, se deben seguir los procedimientos de administración estandarizada de pruebas. Cuando se requieran o admitan procedimientos de administración no estándar, estos se deben describir y justificar. Si la prueba no estaba monitorizada o si se administró bajo procedimientos no estandarizados, se debe informar al intérprete de los resultados. En algunos casos, la administración de la prueba puede proporcionar la oportunidad para que examinadores especializados observen atentamente el desempeño de los examinandos bajo condiciones estandarizadas. Por ejemplo, las observaciones de los administradores de la prueba les pueden permitir el registro de los comportamientos que se evalúan, entender la manera en que los examinandos llegan a las respuestas, identificar los puntos fuertes y débiles de los examinandos, y hacer modificaciones en el proceso de evaluación. Si las pruebas se administran por computadora u otros dispositivos técnicos, el profesional tiene la responsabilidad de determinar si el propósito de la evaluación y las capacidades del examinando requieren la presencia de un monitor o personal de respaldo (p. ej., para ayudar con el uso de las computadoras o el software). Asimismo, algunas pruebas administradas por computadora pueden requerir que se dé al examinando la oportunidad de recibir instrucciones y practicar antes de la administración de la prueba. Los capítulos 4 y 6 proporcionan detalles adicionales sobre las pruebas administradas por medios tecnológicos.

Esfuerzos inapropiados por parte de la persona que está en evaluación podrían afectar los resultados de la evaluación psicológica e introducir errores en la medida del constructo en cuestión. Por lo tanto, en algunos casos, se deberá explicar al examinando la importancia de emplear los esfuerzos apropiados cuando se lleva a cabo una prueba. En muchas pruebas, la medida del esfuerzo se puede deducir de pruebas independientes o de respuestas incorporadas en un procedimiento de evaluación estándar (p. ej., número elevado de errores, respuestas no coherentes y respuestas inusuales correspondientes a patrones de síntomas) y el esfuerzo se puede medir a lo largo del proceso de evaluación. Cuando son evidentes unos bajos niveles de esfuerzo y motivación durante la administración de la prueba, seguir con la evaluación podría traducirse en interpretaciones incorrectas de los puntajes.

Los profesionales tienen la responsabilidad de proteger la confidencialidad y seguridad de los resultados y materiales de las pruebas. El almacenamiento y la transmisión de este tipo de información deberán cumplir los estándares legales y profesionales.

Interpretación de los puntajes de las pruebas

Idealmente, los puntajes usados en la evaluación psicológica se interpretan a la luz de un número de factores, incluyendo los datos normativos disponibles apropiados para las características del examinando, las propiedades psicométricas de la prueba, los indicadores de esfuerzo, las circunstancias del examinando en el momento de ejecutar la prueba, la estabilidad temporal de los constructos que se miden, y los efectos de las variables moderadoras y las características demográficas en los resultados de la prueba. Es poco frecuente que el profesional tenga los recursos disponibles para realizar personalmente la investigación o para recopilar las normas representativas que, en algunos tipos de evaluación, serían necesarias para hacer inferencias sobre el funcionamiento pasado, presente y futuro de cada examinando. Por lo tanto, es posible que el profesional tenga que basarse en la investigación y el corpus de conocimientos científicos disponibles para la prueba que respalda las inferencias apropiadas. La presentación de la evidencia de validación y confiabilidad/precisión no suele ser necesaria en el informe escrito que resume las conclusiones de la evaluación, pero el profesional debe hacer todos los esfuerzos necesarios para conocer (y estar preparado para articular) tal evidencia si fuese necesario.

Cuando se deducen características y se hacen inferencias sobre los comportamientos pasados, presentes y futuros de un examinando a partir de los puntajes de una prueba, el profesional debe considerar otros datos disponibles que respalden

o cuestionen las inferencias. Por ejemplo, el profesional deberá revisar el historial y la información de comportamientos pasados del examinando, así como la literatura pertinente, para familiarizarse con la evidencia de respaldo. En ocasiones, el profesional también deberá corroborar los resultados de una sesión de evaluación con los resultados de otras pruebas y sesiones de evaluación para examinar la confiabilidad/precisión y la validez de las inferencias formuladas sobre el desempeño de un examinando a lo largo del tiempo o de varias pruebas. La triangulación de varias fuentes de información, incluyendo comportamientos estilísticos y de ejecución que se deducen de la observación durante la administración de la prueba, puede reforzar la confianza en la inferencia. Es importante que se reconozcan los datos que no respaldan las inferencias y bien conciliarse con otra información o anotarse como limitación de la confianza puesta en la inferencia. Cuando hay una sólida evidencia para la confiabilidad/precisión y la validez de los puntajes para los usos previstos, y una sólida evidencia de la idoneidad de la prueba para el examinando que se evalúa, aumentará la competencia del profesional para extraer inferencias apropiadas. Cuando una inferencia se basa en un solo estudio o se basa en varios estudios cuyas muestras tienen una generabilidad limitada respecto del examinando, el profesional deberá ser más cauteloso con la inferencia y deberá anotar en el reporte las limitaciones relacionadas con las conclusiones extraídas de la inferencia.

La definición clara de la forma en que se van a utilizar pruebas psicológicas concretas minimizará los riesgos para la interpretabilidad de los puntajes obtenidos. Estos riesgos se producen como resultados de la varianza irrelevante de constructo (es decir, aspectos de la prueba y del proceso de evaluación que no son pertinentes para el propósito de los puntajes de la prueba) y la subrepresentación del constructo (es decir, la incapacidad de la prueba de representar importantes aspectos para el propósito de la evaluación). El sesgo de las respuestas y la simulación son ejemplos de componentes irrelevantes de constructo que pueden desviar considerablemente los puntajes obtenidos, traduciéndose posiblemente en interpretaciones inexactas o equívocas. En situaciones donde se anticipa el sesgo de las respuestas o la simulación, los profesionales pueden seleccionar una prueba que tenga escalas (p. ej., porcentaje de "sí", porcentaje de "no"; "simulación positiva" "simulación negativa") que aclaren los riesgos para la validez. De este modo, los profesionales podrían evaluar el grado de tolerancia de los examinandos a las demandas percibidas del administrador de la prueba o los intentos de presentarse a sí mismos como discapacitados (con "simulación negativa") o funcionales ("simulación positiva").

Con frecuencia, para algunos fines (incluida la orientación profesional y la evaluación neuropsicológica), se usan baterías de pruebas. Por ejemplo, las baterías de orientación profesional podrían incluir pruebas de capacidades, valores, intereses y personalidad. Las baterías neuropsicológicas podrían incluir medidas de orientación, atención, habilidades comunicativas, función ejecutiva, fluidez, habilidades motoras visuales y visuales-espaciales, resolución de problemas, organización, memoria, inteligencia, rendimiento académico y/o personalidad, junto con baterías de esfuerzo. Con frecuencia, cuando las baterías de pruebas psicológicas incorporan varios métodos y puntajes, los patrones de resultados de las pruebas se interpretan como el reflejo de un constructo o incluso de una interacción entre constructos que subyace en el desempeño de las pruebas. Basándose en los patrones de los puntajes de las pruebas, se podrían postular las interacciones entre los constructos que subyacen en las configuraciones de resultados de la prueba Cuando sea posible, se debe identificar la literatura que reporta evidencia de confiabilidad/precisión y validez de las configuraciones de los puntajes que respaldan las interpretaciones propuestas. Sin embargo, se entiende que existe poca o ninguna literatura que describa la validez de las interpretaciones de los puntajes de baterías de pruebas flexibles o altamente personalizadas. El profesional debe reconocer que es habitual que se produzca variabilidad de los puntajes en distintas pruebas de una batería en la población general y, si están disponibles, debe usar datos de valoración de referencia para determinar si la variabilidad observada es excepcional.

Si la literatura es incompleta, es posible que las inferencias resultantes se presenten con la clasificación de hipótesis para una futura verificación y no como enunciados probabilísticos relativos a la probabilidad de un comportamiento que implique alguna evidencia de validación conocida.

Información colateral usada en pruebas y evaluación psicológicas

Los puntajes de las pruebas que se usan como parte de evaluaciones psicológicas se interpretan mejor en el contexto del historial personal y otros rasgos y características personales pertinentes del examinando. A menudo, la calidad de las interpretaciones formuladas a partir de pruebas y evaluaciones psicológicas se mejora con la obtención de información colateral plausible procedente de terceras fuentes importantes, por ejemplo, profesores, profesionales sanitarios, registros escolares, judiciales, militares, profesionales y otros. La calidad de la información colateral se mejora con el uso de varios métodos de adquisición. Además de los puntajes de pruebas objetivos, se pueden usar otros métodos para minimizar la necesidad de que el evaluador dependa del juicio individual, por ejemplo, observaciones conductuales estructuradas, listas de comprobación, calificaciones y entrevistas. Por ejemplo, una evaluación de metas profesionales se puede mejorar mediante la obtención de un historial de empleo, así como mediante la administración de pruebas para evaluar las aptitudes y el rendimiento académico, los intereses vocacionales, los valores de trabajo, la personalidad y el temperamento. La disponibilidad de información sobre las diversas características y atributos, cuando se adquiere de distintas fuentes y a través de varios métodos, permite a los profesionales evaluar con mayor precisión el funcionamiento psicosocial de un individuo y facilita una toma de decisiones más eficaz. Cuando se usan datos colaterales, el profesional debe adoptar las medidas necesarias para verificar la precisión y la confiabilidad, sobre todo cuando los datos proceden de terceros que pueden tener un interés adquirido en los resultados de la evaluación.

Tipos de pruebas y evaluación psicológicas

Para los fines de este capítulo, los tipos de pruebas psicológicas se han dividido en seis categorías: pruebas cognitivas y neuropsicológicas, pruebas de comportamiento problemático, pruebas para familias y parejas, pruebas de comportamientos sociales y de adaptación, pruebas de personalidad y pruebas vocacionales.

Pruebas y evaluación cognitivas y neuropsicológicas

Por lo general, las pruebas se usan para evaluar varios tipos de funcionamiento cognitivo y neuropsicológico, incluyendo la inteligencia, dominios de capacidades generales y dominios más específicos (p. ej., razonamiento abstracto y pensamiento categórico, rendimiento académico, atención, capacidades cognitivas, función ejecutiva, lenguaje, aprendizaje y memoria, funciones motoras, sensomotoras y preferencias laterales, y percepción y organización/integración perceptual). Se puede producir una superposición en los constructos evaluados por pruebas de diferentes funciones o dominios. Al igual que otros tipos de pruebas, las pruebas cognitivas y neuropsicológicas requieren un nivel mínimamente suficiente de capacidad del examinando para mantener la atención, así como un esfuerzo apropiado. Por ejemplo, cuando se administran pruebas cognitivas o neuropsicológicas, algunos profesionales administran primero las pruebas que evalúan dominios básicos (p. ej., la atención) y termina con la administración de pruebas que evalúan dominios más complejos (p. ej., la función ejecutiva).

Razonamiento abstracto y pensamiento categórico. Las pruebas de razonamiento y pensamiento miden una amplia gama de habilidades y capacidades, incluyendo la capacidad de los examinandos para inferir relaciones, formar nuevos conceptos o estrategias, responder a circunstancias ambientales cambiantes, así como la capacidad de entender un problema o concepto, desarrollar una estrategia para resolver ese problema y, de ser necesario, alterar tales conceptos o estrategias a medida que las situaciones cambian.

Rendimiento académico. Las pruebas de rendimiento académico miden los conocimientos y la competencia que ha adquirido una persona en situaciones formales e informales de aprendizaje. Los dos principales tipos de pruebas de rendimiento académico son las baterías generales de rendimiento y las pruebas diagnósticas de rendimiento. Las baterías generales de rendimiento están diseñadas para evaluar el nivel de aprendizaje de una persona en varias áreas (p. ej., lectura, matemáticas y ortografía). Por el contrario, las pruebas diagnósticas de rendimiento se centran, por lo general, en un área temática (p. ej., la lectura) y evalúan una competencia académica con mayor detalle. Los resultados de las pruebas se usan para determinar los puntos fuertes de los examinandos y también para identificar fuentes de dificultades o deficiencias académicas. El capítulo 12 proporciona detalles adicionales sobre pruebas de rendimiento académico en contextos educativos.

Atención. La atención se refiere a un dominio que abarca los constructos de estimulación, creación de conjuntos, despliegue estratégico de atención, atención continua, atención dividida, atención concentrada, atención selectiva y vigilante. Las pruebas pueden medir (a) los niveles de alerta, orientación y localización; (b) la capacidad de centrar, desplazar y mantener la atención y de seguir uno o más estímulos bajo diversas condiciones; (c) la amplitud de la atención; y (d) el funcionamiento del almacenamiento de atención a corto plazo. Los puntajes de cada uno de los aspectos de la atención que se haya examinado se deben reportar individualmente, de manera que sea posible clarificar la naturaleza de un trastorno de atención.

Capacidad cognitiva. Entre las pruebas más extensamente administradas están las medidas diseñadas para cuantificar las capacidades cognitivas. La interpretación de resultados de una prueba de capacidad cognitiva se rige por los constructos teóricos usados para desarrollar la prueba. Algunas evaluaciones de la capacidad cognitiva se basan en resultados de baterías de pruebas multidimensionales diseñadas para acceder a una amplia gama de habilidades y capacidades. Los resultados de las pruebas se usan para formular inferencias sobre el nivel general de funcionamiento intelectual de una persona y sobre los puntos fuertes y débiles de varias capacidades cognitivas, y para diagnosticar trastornos cognitivos.

Función ejecutiva. Este tipo de funciones intervienen en los desempeños organizados (p. ej., flexibilidad cognitiva, control inhibitorio, multitarea) que se necesitan para la consecución independiente, deliberada y efectiva de objetivos en diversas situaciones sociales, el procesamiento cognitivo y la resolución de problemas. Algunas pruebas remarcan (a) los planes razonados de acción que anticipan las consecuencias de soluciones alternativas, (b) el desempeño motor en situaciones de resolución de problemas que requieren intenciones orientadas a los objetivos, y/o (c) la regulación del desempeño para conseguir un resultado deseado.

Lenguaje. Por lo general, las deficiencias de lenguaje se identifican con evaluaciones que se centran en la fonología, morfología, sintaxis, semántica, supralingüística y pragmática. Se pueden evaluar varias funciones, incluyendo las capacidades y habilidades de lectura, auditivas, y de lenguaje oral y escrito. Las evaluaciones de trastornos de lenguaje se centran en el habla funcional y la compresión verbal medidas a través de los modos orales, escritos o gestuales; el acceso y la elaboración léxicos; la repetición de lenguaje oral y la fluidez verbal asociativa. Si se evalúa a una persona multilingüe por un posible trastorno de lenguaje, se debe abordar el grado en que el trastorno se debe más directamente a problemas de desarrollo del lenguaje (p. ej., retrasos fonológicos, morfológicos, sintácticos, semánticos o pragmáticos; discapacidades intelectuales; deterioros periféricos, sensoriales o neurológicos centrales; condiciones psicológicas o trastornos sensoriales) que a una falta de destreza en un determinado idioma.

Aprendizaje y memoria. Este tipo de funciones incluye la adquisición, retención y recuperación

de información más allá de los requisitos de procesamiento y almacenamiento a corto plazo de la información. Estas pruebas pueden medir la adquisición de nueva información a través de varios canales sensoriales y mediante formatos de pruebas heterogéneos (p. ej., listas de palabras, texto en prosa, figuras geométricas, tableros de figuras, dígitos y melodías musicales). Es posible que las pruebas de memoria también requieran retención y recuperación de información antigua (p. ej., datos personales, así como datos y habilidades de aprendizaje común). Además, las pruebas de reconocimiento de información almacenada se pueden usar en la comprensión de déficits de memoria.

Funciones motoras, funciones sensomotoras y preferencias laterales. Con frecuencia, las funciones motoras (p. ej., dar golpes con el dedo) y las funciones sensoriales (p. ej., estimulación táctil) se miden como parte de una extensa evaluación neuropsicológica. Las pruebas motoras evalúan varios aspectos del movimiento como, por ejemplo, la velocidad, destreza, coordinación e intencionalidad. Las pruebas sensoriales evalúan la función de las áreas de visión, audición, tacto y, a veces, olfato. Las pruebas también se llevan a cabo para examinar la integración de las funciones perceptuales y motoras.

Percepción y organización/integración perceptual. Este tipo de funcionamiento involucra el razonamiento y juicio en la medida en que se relacionan con el procesamiento y elaboración de complejas combinaciones y entradas sensoriales. Las pruebas de percepción pueden hacer hincapié en el procesamiento perceptual inmediato, pero también requieren conceptualizaciones que implican algunos procesos de razonamiento y juicio. Algunas pruebas tienen componentes motores que van desde movimientos simples hasta la elaboración de construcciones complejas. Estas pruebas evalúan actividades que incluyen desde la velocidad perceptual y el tiempo de reacción de elección hasta el procesamiento de información compleja y el razonamiento visual-espacial.

Pruebas y evaluación de comportamientos problemáticos

Los comportamientos problemáticos incluyen dificultades de ajuste que interfieren con el funcionamiento eficaz de una persona en situaciones de la vida cotidiana. Las pruebas se usan para evaluar el comportamiento y autopercepción del individuo para diagnósticos diferenciales y la clasificación educativa de una variedad de trastornos emocionales y de conducta, y para ayudar en el desarrollo de planes de tratamiento. En algunos casos (p. ej., evaluaciones para la pena de muerte), el análisis retrospectivo y distintas fuentes de información ayudan a obtener la evaluación más completa posible. A menudo, la observación de una persona en su entorno resulta valiosa para entender cabalmente las demandas específicas del entorno, no solo para ofrecer una evaluación más exhaustiva, sino también para proporcionar recomendaciones más útiles.

Pruebas y evaluación para familias y parejas

Las pruebas familiares abordan los problemas de la dinámica, cohesión y relaciones interpersonales entre los miembros de una familia, incluyendo parejas, padres, hijos y otros miembros de una familia extendida. Las pruebas desarrolladas para evaluar familias y parejas se caracterizan por medir los patrones de interacción de familias parciales o completas. En ambos casos se requiere el enfoque simultáneo en dos o más miembros de la familia en lo que respecta a sus transacciones. Las pruebas con parejas pueden abordar factores como problemas de intimidad, compatibilidad, intereses compartidos, confianza y creencias espirituales.

Pruebas y evaluación de comportamientos sociales y de adaptación

Las medidas de comportamientos sociales y de adaptación evalúan la motivación y capacidad para cuidar de uno mismo y en relación con otros. Los comportamientos sociales y de adaptación se basan en un repertorio de conocimientos, habilidades y capacidades que permiten a las personas satisfacer las demandas y expectativas cotidianas del entorno, por ejemplo, comer, vestirse,

trabajar, participar en actividades de ocio, usar el transporte, interactuar con compañeros, comunicarse con otros, hacer compras, gestionar el dinero, mantener un horario, vivir de forma independiente, ser socialmente receptivo y llevar a cabo conductas saludables.

Pruebas y evaluación de personalidad

La evaluación de personalidad requiere una síntesis de los aspectos del funcionamiento de un individuo que contribuyen a la formulación y expresión de pensamientos, actitudes, emociones y comportamientos. Algunos de estos aspectos son estables a lo largo del tiempo; otros cambian con la edad o son específicas de una situación. En la evaluación de un individuo, los funcionamientos cognitivos y emocionales se podrían considerar por separado, pero sus influencias están interrelacionadas. Por ejemplo, una persona cuyas percepciones tiene una alta precisión o que es relativamente estable en el ámbito emocional, podría controlar la suspicacia mejor que una persona cuyas percepciones son imprecisas o distorsionadas o que es emocionalmente inestable.

Los puntajes o descriptores de personalidad obtenidos de una prueba de personalidad se podrían considerar un reflejo de los constructos teóricos subyacentes, o de las escalas o factores derivados empíricamente que guiaron la construcción de la prueba. Los formatos de estímulo-respuesta de las pruebas de personalidad varían ampliamente. Algunos incluyen una serie de preguntas (p. ej., inventarios de autoevaluación) que el examinando debe responder mediante la elección de distintas opciones bien definidas; otros comportan enfrentar una situación novedosa donde la respuesta de los examinandos no está completamente estructurada (p. ej., responder a estímulos visuales, contar historias, debatir sobre imágenes o responder a estímulos proyectivos). Los resultados pueden incluir temas, patrones o indicadores diagnósticos, así como puntajes. Las respuestas se califican y combinan en dimensiones derivadas lógica o estadísticamente, establecidas por investigaciones previas.

Las pruebas de personalidad pueden estar diseñadas para evaluar actitudes, sentimientos, rasgos o características relacionadas normales o anormales. Las pruebas dirigidas a medir características normales de la personalidad se construyen para generar puntajes que reflejen el grado en que una persona manifiesta dimensiones de personalidad empíricamente identificadas e hipotesizadas como presentes en el comportamiento de la mayoría de los individuos. La configuración de puntajes de una persona en estas dimensiones se usa, por tanto, para inferir su comportamiento actual y su posible comportamiento ante nuevas situaciones. Los puntajes de las pruebas fuera del rango previsto se podrían considerar expresiones acentuadas de rasgos normales o indicar psicopatologías. Estos puntajes también podrían reflejar el funcionamiento normal de una persona dentro de una cultura diferente de la población en la que se basa la norma.

Otras pruebas de personalidad están diseñadas específicamente para medir los constructos que subyacen a funcionamientos anormales y psicopatologías. Los desarrolladores de algunas de estas pruebas usan individuos diagnosticados previamente para construir sus escalas y basan sus interpretaciones en la asociación entre los puntajes de la escala de la prueba, dentro de un rango determinado, y las correlaciones de conducta de personas que puntuaron dentro de ese rango, en comparación con muestras clínicas. Si las interpretaciones formuladas a partir de los puntajes van más allá de la teoría que guiaba la construcción de la prueba, se deberá recopilar y analizar la evidencia de validación de las interpretaciones a partir de los datos pertinentes adicionales.

Pruebas y evaluación vocacionales

Por lo general, las pruebas vocacionales incluyen la medición de intereses, necesidades y valores del trabajo, así como la consideración y evaluación de elementos relacionados con el desarrollo, la madurez y la indecisión profesionales. El rendimiento académico y las capacidades cognitivas, que se trataron anteriormente en la sección de capacidad cognitiva, son también componentes importantes en las pruebas y evaluaciones vocacionales. Los resultados de estas pruebas se suelen usar para mejorar el desarrollo y el conocimiento

personal, para la orientación profesional, el asesoramiento en reasignación y la toma de decisiones en el ámbito vocacional. Con frecuencia, estas intervenciones tienen lugar en el contexto de rehabilitación educativa y vocacional. No obstante, las pruebas vocacionales también se pueden usar en el centro de trabajo como parte de programas corporativos de desarrollo profesional.

Inventarios de intereses. La medición de intereses está diseñada para identificar las preferencias de una persona con respecto a diversas actividades. Los inventarios de autoevaluación de intereses son muy utilizados para evaluar las preferencias personales, incluyendo el agrado o aversión por diversos trabajos y actividades de ocio, áreas escolares, ocupaciones o tipos de personas. Los puntajes resultantes pueden proporcionar una mayor comprensión de los tipos y patrones de intereses en planes de estudio (p. ej., especialidades universitarias), diversos campos de trabajo (p. ej., ocupaciones específicas), o en áreas más básicas y generales de interés relacionadas con actividades concretas (p. ej., ventas, prácticas de oficina o actividades mecánicas).

Inventarios de valores del trabajo. La medición de valores del trabajo identifica las preferencias de una persona por los diversos reforzamientos que se pueden obtener de las actividades de trabajo. A veces estos valores se identifican como necesidades que esas personas tratan de satisfacer. Los valores o necesidades del trabajo se pueden categorizar como intrínsecas e importantes por el placer obtenido de la actividad (p. ej., ser independiente, usar las habilidades personales) o extrínsecas e importantes por las compensaciones que aportan (p. ej., pago, ascensos). En general, el formato de las pruebas de valores del trabajo incluye una autoclasificación de la importancia del valor asociado con las cualidades descritas por los ítems.

Medidas de desarrollo, madurez e indecisión profesionales. Áreas adicionales de la evaluación vocacional incluyen medidas del desarrollo y la madurez profesional, y medidas de la indecisión profesional. Habitualmente, los inventarios que miden el desarrollo y la madurez profesional recaban autodescripciones en respuesta a los ítems que preguntan sobre el conocimiento del individuo del mundo laboral; autovaloraciones de sus capacidades en la toma de decisiones; las actitudes hacia las profesiones y la elección de profesiones; y el grado actual de compromiso de los individuos en la planificación profesional. Por lo general, las medidas de la indecisión profesional se construyen y estandarizan para evaluar el nivel de indecisión profesional de un examinando y las razones o antecedentes de esa indecisión. Los resultados de estas pruebas se utilizan, con frecuencia, como guía para el diseño y el suministro de servicios profesionales a individuos y grupos, y para evaluar la eficacia de las intervenciones profesionales.

Propósitos de las pruebas y evaluación psicológicas

Para fines de este capítulo, el uso de las pruebas psicológicas se ha dividido en cinco categorías: pruebas de diagnóstico; pruebas de evaluaciones neuropsicológicas; pruebas de planificación de intervenciones y evaluación de resultados; pruebas para decisiones judiciales y gubernamentales; y pruebas de conciencia personal, identidad social, y salud, desarrollo y acción psicológicos. Sin embargo, no siempre estas categorías son mutuamente exclusivas.

Pruebas de diagnóstico

El *diagnóstico* se refiere a un proceso que incluye la recopilación e integración de los resultados de las pruebas con información previa y actual sobre una persona, junto con las condiciones contextuales pertinentes, para identificar características de funcionamiento psicológico saludable, así como trastornos psicológicos. Los trastornos se pueden manifestar por sí mismos en la información obtenida durante la evaluación de los atributos cognitivos, emocionales, de adaptación, de conducta, de personalidad, neuropsicológicos, físicos o sociales.

Las pruebas psicológicas resultan útiles para los profesionales involucrados en el diagnóstico

de la salud psicológica de un individuo. La evaluación se puede llevar a cabo para confirmar un diagnóstico hipotesizado o para descartar diagnósticos alternativos. El diagnóstico se complica por la prevalencia de la comorbilidad entre categorías de diagnósticos. Por ejemplo, un individuo diagnosticado con demencia podría ser diagnosticado simultáneamente como depresivo. O un niño diagnosticado con discapacidad de aprendizaje también podría ser diagnosticado con el trastorno por déficit de atención e hiperactividad. El objetivo del diagnóstico es proporcionar una breve descripción de la disfunción psicológica del examinando y ayudar a que cada examinando reciba las intervenciones apropiadas para la disfunción psicológica o conductual que el cliente, o un tercero, considera que afecta al funcionamiento previsto del cliente y/o a su calidad de vida. Cuando la intención de la evaluación es el diagnóstico diferencial, el profesional debe usar pruebas donde exista evidencia de que los puntajes discriminan entre dos o más grupos de diagnóstico. Las diferencias de medias entre grupos no proporcionan suficiente evidencia para la precisión del diagnóstico diferencial; los desarrolladores de pruebas también deben suministrar información adicional como los tamaños de los efectos o datos que indiquen el grado de superposición entre grupos de criterios. En el desarrollo de planes de tratamiento, los profesionales suelen usar descripciones de diagnóstico no categóricas del funcionamiento del cliente, junto con dimensiones pertinentes al tratamiento (p. ej., capacidad funcional, grado de ansiedad, grado de desconfianza, receptividad a interpretaciones, grado de conocimiento de los comportamientos y nivel de funcionamiento intelectual).

Los criterios de diagnóstico pueden variar de un sistema de nomenclatura a otro. Anotar el sistema de nomenclatura en uso es un paso inicial importante porque distintos sistemas diagnósticos pueden usar el mismo término para describir distintos síntomas. Incluso dentro de un sistema diagnóstico, los síntomas descritos por el mismo término pueden diferir entre distintas ediciones del manual. De forma similar, una prueba que usa un término de diagnóstico en el título puede diferir considerablemente de otra prueba que usa un título similar o de una subescala que usa el mismo término. Por ejemplo, algunos sistemas diagnósticos pueden definir la *depresión* por sintomatología conductual (p. ej., retardo psicomotor, perturbaciones en el apetito o el sueño), por sintomatología afectiva (p. ej., sentimientos disfóricos, monotonía emocional) o por sintomatología cognitiva (p. ej., pensamientos de desesperación, morbilidad). Además, los síntomas de las categorías diagnósticas raramente son mutuamente exclusivas. Por lo tanto, se puede prever que varias categorías diagnósticas puedan compartir un síntoma dado. Se podrían obtener inferencias formuladas con más información y precisión relacionadas con un diagnóstico de los puntajes de las pruebas si se diese una ponderación apropiada a los síntomas incluidos en la categoría diagnóstica y a la idoneidad de cada prueba para evaluar los síntomas. Por lo tanto, el primer paso en la evaluación de la idoneidad de una prueba para la obtención de puntajes o información indicativa de un síndrome específico de diagnóstico es comparar el constructo que la prueba tiene previsto medir con la sintomatología descrita en los criterios de diagnóstico.

Se pueden usar distintos métodos para evaluar categorías diagnósticas específicas. Algunos métodos se basan fundamentalmente en entrevistas estructuradas que usan un formato de "sí" / "no" o "verdadero" / "falso", donde el profesional se interesa en la presencia o ausencia de la sintomatología específica del diagnóstico. Otros métodos suelen basarse sobre todo en pruebas de personalidad o de funcionamiento cognitivo y usan configuraciones de los puntajes obtenidos. Estas configuraciones de puntajes indican el grado de similitud de las respuestas de los examinandos con respecto a las respuestas de individuos que pertenecen a un grupo de diagnóstico específico, de acuerdo con investigaciones previas.

Los diagnósticos hechos con la ayuda de puntajes de pruebas se suelen basar en relaciones empíricamente demostradas entre el puntaje de la prueba y la categoría diagnóstica. Actualmente, están actualmente estudios de validez que demuestran las relaciones entre los puntajes de las

pruebas y algunas categorías diagnósticas, aunque no todas. Muchos otros estudios acreditan la evidencia de validación para las relaciones entre los puntajes de las pruebas y varios subconjuntos de síntomas que contribuyen a una categoría diagnóstica. Aunque a menudo no resulta factible para los profesionales individuales llevar a cabo personalmente una investigación entre los puntajes obtenidos y las categorías diagnósticas es importante familiarizarse con la literatura investigativa que examina estas relaciones.

A menudo, el profesional puede mejorar las interpretaciones del diagnóstico que se derivan de los puntajes de las pruebas mediante la integración de esos resultados con inferencias formuladas a partir de otras fuentes de información sobre el funcionamiento del examinando, por ejemplo, información del historial de autoevaluaciones proporcionada por personas allegadas u observaciones sistemáticas en el entorno natural o en el contexto de evaluación. En el proceso para llegar a un diagnóstico, un profesional también debe buscar información que no corrobore el diagnóstico y, en algunos casos, determinar límites apropiados al grado de confianza que se da al diagnóstico. Cuando sea pertinente para una decisión de remisión, el profesional debe reconocer los diagnósticos alternativos que pueden requerir consideración. Se debe prestar especial atención a todos los datos disponibles pertinentes antes de concluir que un examinando encaja en una categoría diagnóstica. La competencia cultural resulta esencial para evitar los diagnósticos erróneos o patologizar en exceso un comportamiento, afecto o cognición culturalmente apropiada. Las pruebas también se usan para evaluar la idoneidad de continuar el diagnóstico inicial, especialmente después de un plan de tratamiento o si el funcionamiento psicológico del cliente ha cambiado a lo largo del tiempo.

Pruebas de evaluaciones neuropsicológicas

Las pruebas neuropsicológicas analizan el actual estado psicológico y conductual de un examinando, incluyendo manifestaciones de los cambios neurológicos, neuropatológicos o neuroquímicos que puedan surgir durante el desarrollo o de psicopatologías, lesiones corporales o cerebrales, o enfermedad. Por lo general, los propósitos de las pruebas neuropsicológicas incluyen, entre otros, los siguientes: diagnóstico diferencial asociado con las fuentes de disfunción cognitiva, perceptual y de personalidad; diagnóstico diferencial entre dos o más presuntas etiologías de disfunción cerebral; evaluación de funcionamiento deficiente secundario a un evento cortical o subcortical; establecimiento de medidas de referencia neuropsicológicas para el control de enfermedades cerebrales progresivas o los efectos de la recuperación; identificación de patrones de funciones y disfunciones corticales superiores para la formulación de estrategias de recuperación y para el diseño de procedimientos correctivos; y caracterización de las funciones de la conducta cerebral para ayudar en acciones judiciales civiles o penales.

Pruebas de planificación de intervenciones y evaluación de resultados

Con frecuencia, los profesionales se basan en los resultados de las pruebas para la asistencia en la planificación, ejecución y evaluación de intervenciones. Por lo tanto, es importante su nivel de conocimiento respecto de la información de validez que respalda o no las relaciones entre los resultados de las pruebas, las intervenciones prescritas y los resultados deseados. Las intervenciones se pueden usar para prevenir la aparición de uno o más síntomas, para remediar las deficiencias y para atender las necesidades psicológicas, físicas y sociales básicas de una persona con el fin de mejorar su calidad de vida. Por lo general, la planificación de una intervención se produce después de una evaluación de la naturaleza, evolución y severidad de un trastorno y de una revisión de las condiciones personales y contextuales que pueden afectar a su resolución. Podrían darse evaluaciones posteriores que requieran de la administración repetida de la misma prueba en un esfuerzo de diagnosticar la naturaleza y severidad del trastorno, para revisar los efectos de las intervenciones, para revisar las intervenciones según sea necesario y para cumplir los estándares éticos y legales.

Pruebas para decisiones judiciales y gubernamentales

Es posible que los clientes busquen voluntariamente evaluación psicológica para asistencia en cuestiones relacionadas con un tribunal u otro organismo gubernamental. Por otra parte, a veces los tribunales u otros organismos gubernamentales requieren que una persona se someta de forma no voluntaria a una evaluación psicológica que puede incluir una amplia variedad de pruebas. El objetivo de estas evaluaciones psicológicas es proporcionar información importante a un tercero (p. ej., el abogado del examinando, el abogado contrario, el juez o un órgano administrativo) sobre el funcionamiento psicológico del examinando relacionado con los asuntos legales en cuestión. En general, se debe obtener el consentimiento informado; para niños e individuos mentalmente incompetentes (p. ej., personas con demencia) se debe obtener el consentimiento informado de los tutores legales. Al iniciar la evaluación para decisiones judiciales y gubernamentales, el profesional debe explicar los propósitos previstos de la evaluación e identificar a quienes posiblemente tengan acceso a los resultados y al reporte de la prueba. A menudo, el profesional y el examinando no tienen pleno conocimiento de las cuestiones o parámetros legales que afectan a la evaluación, y si el examinando decide no continuar después de ser informado de la naturaleza y el propósito del examen, el profesional (cuando proceda) puede tratar de administrar la evaluación, posponerla, aconsejar al examinando que se ponga en contacto con su abogado, o notificar de la falta de disposición del examinando al individuo u organismo que solicita la evaluación.

La evaluación por motivos legales puede tener lugar como parte de un proceso civil (p. ej., compromiso involuntario, capacidad testamentaria, competencia para ser juzgado, conceder la custodia de un menor, lesiones personales, demandas legales), un proceso penal (p. ej., competencia para ser juzgado, declaración de no culpabilidad por motivos de enajenación, circunstancias atenuantes al pronunciar la sentencia), la determinación de adecuaciones razonables para empleados con discapacidades, o un proceso o decisión administrativa (p. ej., revocación de licencia, libertad condicional, compensación para un trabajador). El profesional tiene la responsabilidad de explicar los puntajes de la prueba y las interpretaciones que se derivan de los mismos en términos de los criterios legales en los que se basará el jurado, el juez o el órgano administrativo para tomar la decisión de la cuestión legal. En casos que involucren cuestiones legales, es importante evaluar la orientación de la realización de la prueba de los examinandos, incluyendo el sesgo de las respuestas, para asegurarse de que los procedimientos legales no se hayan visto afectados por las respuestas dadas. Por ejemplo, las personas que quieren obtener la máxima compensación económica por una lesión personal podrían tener la motivación para exagerar los síntomas cognitivos y emocionales, mientras que las personas que intentan evitar la pérdida de una licencia profesional pueden tratar de presentarse a sí mismos en la mejor posición posible, minimizando los síntomas o deficiencias. Al formar una opinión de evaluación, es necesario interpretar los puntajes de la prueba con el conocimiento informado relacionado con la evidencia de validación y confiabilidad disponible. A la hora de elaborar tales opiniones, también es necesario integrar los puntajes de un examinando con todas las otras fuentes de información que se refieren al estado actual del examinando, incluyendo fuentes psicológicas, sanitarias, educativas, ocupacionales, legales, socioculturales y otros registros colaterales pertinentes.

Algunas pruebas tienen como objetivo proporcionar información sobre el funcionamiento de un cliente que ayude a aclarar una cuestión legal determinada (p. ej., funcionamiento parental en un caso de custodia de menores o la capacidad de un acusado para entender los cargos en audiencias sobre la competencia para ser juzgado). Los manuales de algunas pruebas también ofrecen datos demográficos y actuariales de grupos normativos que son representativos de personas involucradas en el sistema legal. No obstante, muchas pruebas miden constructos que son, en general, pertinentes para las cuestiones legales, aun cuando pueden no estar disponibles normas específicas para

el contexto judicial o gubernamental. Se espera que los profesionales hagan todo lo posible para tener en cuenta la evidencia de validación y confiabilidad/precisión que respalda o no sus interpretaciones y para poner límites apropiados a las opiniones elaboradas. Se espera que los usuarios de la prueba que ejercen en contextos judiciales y gubernamentales tengan en cuenta los conflictos de interés que puede suponer un sesgo en la interpretación de los resultados.

La protección de la confidencialidad de los resultados de la prueba de un examinando y de los propios instrumentos de la prueba plantea retos específicos a los profesionales que trabajan con abogados, jueces, jurados y otros responsables legales. El examinando tiene derecho a esperar que los resultados de una prueba solo se comuniquen a las personas legalmente autorizadas para recibirlos y que no se reporte ninguna otra información de la sesión de pruebas que no sea pertinente para la evaluación. El profesional deberá ser informado de las posibles amenazas a la confidencialidad y seguridad de la prueba (p. ej., la divulgación a otro profesional cualificado de preguntas de la prueba, respuestas del examinando o puntajes brutos o estandarizados de las pruebas) y deberá buscar, si es necesario, recursos legales y profesionales apropiados.

Pruebas de conciencia personal, identidad social, y salud, desarrollo y acción psicológicos

Con frecuencia, las pruebas e inventarios se usan para proporcionar información que ayuda a las personas a entenderse a sí mismas, identificar sus propios puntos fuertes y débiles, y aclarar cuestiones importantes para su propio desarrollo. Por ejemplo, los resultados de los inventarios de personalidad pueden ser útiles para que los examinandos tengan una mejor comprensión de sí mismos y de sus interacciones con los demás. Las medidas de identidad étnica y aculturación (dos componentes de la identidad social) que evalúan los aspectos cognitivos, afectivos y de conducta de los modos en que las personas se identifican con sus trasfondos culturales, también pueden ser informativas.

En ocasiones, las pruebas psicológicas se usan para evaluar la capacidad de un individuo de entender y adaptarse a afecciones de salud. En estos casos, las observaciones y listas de comprobación, así como las pruebas, se usan para medir la comprensión que un individuo con una afección (p. ej., diabetes) tiene sobre el proceso de su enfermedad y sobre las técnicas de conducta y cognitivas aplicables para el mejoramiento o control de los síntomas del estado patológico.

Los resultados de inventarios de intereses y pruebas de capacidad pueden resultar útiles para las personas que deben tomar decisiones educativas o profesionales. Las pruebas cognitivas y neuropsicológicas apropiadas que han sido normalizadas o estandarizadas para niños pueden facilitar la supervisión del desarrollo y crecimiento durante los años formativos, cuando las intervenciones relevantes pueden ser más eficaces para el reconocimiento y la prevención de posibles dificultades incapacitantes de aprendizaje. Los puntajes de las pruebas para jóvenes adultos o niños en este tipo de medidas podrían cambiar en los años siguientes; por lo tanto, los usuarios de las pruebas deben ser precavidos respecto del exceso de confianza en resultados que pueden estar obsoletos.

Los resultados de las pruebas se pueden usar de diversas maneras para la autoexploración, el crecimiento y la toma de decisiones. Primero, los resultados pueden proporcionar a los individuos nueva información que les permita compararse con otros o evaluarse centrando su atención en autodescripciones o autocaracterizaciones. Los resultados de las pruebas también podrían servir para estimular las deliberaciones entre el examinando y el profesional, facilitar el análisis por parte del examinando, proporcionar instrucciones para futuras consideraciones de tratamiento, ayudar a que las personas identifiquen sus puntos fuertes y débiles, y ofrecer al profesional un marco general de organización e integración de la información sobre un individuo. Las pruebas para el crecimiento personal pueden llevarse a cabo en programas de capacitación y desarrollo, dentro de un plan de estudios educativo, durante la psicoterapia, en programas de rehabilitación como parte

de un proceso educativo o de desarrollo profesional, o en otras situaciones.

Resumen

El uso responsable de las pruebas en la práctica psicológica requiere que el profesional se comprometa a desarrollar y mantener los conocimientos y la competencia necesarios para seleccionar, administrar e interpretar las pruebas e inventarios como elementos cruciales del proceso de pruebas y evaluación psicológicas (véase el cap. 9). Los estándares de este capítulo ofrecen un marco de orientación para los profesionales para la consecución de relevancia y eficacia en el uso de pruebas psicológicas dentro de los límites definidos por los principios educativos, de experiencia y éticos del profesional. Los capítulos y estándares anteriores pertinentes para las pruebas y la evaluación psicológicas describen aspectos generales sobre la calidad (cap. 1 y 2), imparcialidad (cap. 3), diseño y desarrollo (cap. 4) y administración (cap. 6) de las pruebas. El capítulo 11 examina el uso de pruebas para el centro de trabajo, incluyendo la acreditación, y la importancia de la recopilación de datos que proporcionen evidencia de la precisión de una prueba para la predicción de desempeño en el trabajo; el capítulo 12 examina las aplicaciones educativas; y el capítulo 13 examina el uso de las pruebas en evaluación de programas y políticas públicas.

ESTÁNDARES PARA LAS PRUEBAS Y LA EVALUACIÓN PSICOLÓGICAS

Los estándares de este capítulo se han separado en cinco unidades temáticas denominadas de la siguiente manera:

1. Cualificaciones del usuario de la prueba
2. Selección de pruebas
3. Administración de pruebas
4. Interpretación de pruebas
5. Seguridad de pruebas

Unidad 1. Cualificaciones del usuario de la prueba

Estándar 10.1

Quienes usan pruebas psicológicas deben restringir sus actividades relacionadas con las pruebas y la evaluación a sus áreas de competencia, según se hayan demostrado mediante credenciales apropiadas de educación, capacitación y experiencia.

Comentario: El uso y la interpretación responsables de los puntajes de pruebas requieren los niveles apropiados de experiencia, un sólido juicio profesional y el conocimiento de los principios empíricos y teóricos de las pruebas. En muchas evaluaciones, la competencia también requiere de suficiente familiaridad con la población de la que forma parte el examinando para facilitar la selección de pruebas, la administración y la interpretación de los puntajes. Por ejemplo, cuando se administran pruebas de personalidad y neuropsicológicas como parte de la evaluación psicológica de un individuo, los puntajes de las pruebas se deben entender en el contexto del estado físico y psicológico, el desarrollo cultural y lingüístico, y los antecedentes educativos, sanitarios, ocupacionales y de género del individuo. La calificación también deberá tener en cuenta otras evidencias pertinentes para las pruebas utilizadas. La interpretación de los puntajes de las pruebas requiere que el juicio profesionalmente responsable se ejerza dentro de los límites de los conocimientos y capacidades que la educación, capacitación y experiencia supervisadas ha conferido al profesional, así como en el contexto donde se lleva a cabo la evaluación.

Estándar 10.2

Quienes seleccionan pruebas y formulan inferencias a partir de los puntajes de las pruebas deben estar familiarizados con la evidencia pertinente de validación y confiabilidad/precisión para los usos previstos de los puntajes y evaluaciones, y deben estar preparados para articular un análisis lógico que respalde todos los aspectos de la evaluación y las inferencias hechas a partir de esa evaluación.

Comentario: En general, no es necesaria la presentación y el análisis de la evidencia de validación y confiabilidad/precisión en un reporte que se proporcione al examinando o a un tercero, ya que es demasiado engorroso y de poco interés para la mayoría de los lectores. No obstante, en situaciones en que la selección de pruebas puede ser problemática (p. ej., subpruebas orales con examinandos sordos), se recomienda una breve descripción de la justificación del uso o no uso de medidas específicas.

Cuando las inferencias potenciales derivadas de los puntajes de pruebas psicológicas no tienen el respaldo de los datos actuales, pero presentan posibilidades de validez futura, el desarrollador y el usuario de la prueba podrían describirlas como hipótesis para una validez posterior en la interpretación de los puntajes. Se deberá advertir a quienes reciban las interpretaciones de tales resultados de que esas inferencias todavía no tienen una evidencia de validación demostrada adecuadamente y que no deben servir de base a decisiones diagnósticas o formulación de pronósticos.

Estándar 10.3

Los profesionales deben verificar que las personas bajo su supervisión tengan los conocimientos

y capacidades apropiadas para administrar y calificar las pruebas.

Comentario: Las personas que administran pruebas, pero no participan en su selección o interpretación deben estar supervisadas por un profesional. Deben tener conocimientos de (y experiencia con) los problemas existentes de los examinandos (p. ej. lesiones cerebrales) y los contextos de las pruebas (p. ej., clínicos, forenses).

Unidad 2. Selección de pruebas

Estándar 10.4

Las pruebas que se combinan para formar una batería de pruebas deben ser apropiadas para los propósitos de la evaluación.

Comentario: Por ejemplo, en una evaluación psicológica para obtener evidencia de una lesión en un área cerebral, es necesario seleccionar una combinación de pruebas con sensibilidad y especificidad diagnósticas conocidas respecto de las discapacidades que se deriven del trauma en regiones concretas del cerebro.

Estándar 10.5

Las pruebas seleccionadas para el uso en evaluaciones psicológicas deben ser idóneas para las características y los antecedentes del examinando.

Comentario: Por lo general, cuando las pruebas sean parte de una evaluación psicológica, el profesional deberá tener en cuenta las características del examinando individual, incluyendo la edad y el nivel de desarrollo, raza/origen étnico, género y características físicas y/o lingüísticas que puedan afectar a la capacidad del examinando de cumplir los requisitos de la prueba. El profesional también deberá tener en cuenta la disponibilidad de normas y la evidencia de validación para una población representativa del examinando. Si no hay estudios normativos o de validez disponibles para una población pertinente, las interpretaciones de una prueba se deben clasificar y presentar como hipótesis y no como conclusiones.

Estándar 10.6

Cuando se necesita un diagnóstico diferencial, el profesional debe elegir, si es posible, una prueba o pruebas para las que exista una evidencia plausible de que sus puntajes distingan entre dos o más grupos de diagnósticos de interés, y no que solamente puedan distinguir los casos anormales en una población general.

Comentario: Para los profesionales será especialmente útil que la evidencia de validación se encuentre en una forma que les permita determinar el nivel de confianza que se puede otorgar a las interpretaciones para un individuo. Las diferencias entre medias de grupo y su importancia estadística proporcionan información inadecuada respecto de la validez para propósitos de diagnóstico individual. La información adicional que se podría considerar incluye los tamaños de los efectos o una tabla que muestre el grado de superposición de las distribuciones de predictores entre distintos grupos de criterios.

Unidad 3. Administración de pruebas

Estándar 10.7

Antes de las pruebas, los profesionales y administradores deben proporcionar al examinando (o a otros terceros si corresponde) información introductoria que sea fácilmente comprensible.

Comentario: El objetivo de la administración óptima de pruebas es reducir el error en la medida del constructo. Por ejemplo, el examinando debe entender los parámetros relacionados con la prueba, como los límites de tiempo, las observaciones o la carencia de estas, y las oportunidades de hacer pausas. Además, el examinando debe tener conocimiento de los límites de la confidencialidad, quién tendrá acceso a los resultados de la prueba, si los resultados o decisiones que se basan en los puntajes se compartirán con el examinando y cuándo se compartirían, si el examinando tendrá la oportunidad de repetir la prueba y bajo qué circunstancias se produciría esta repetición.

Estándar 10.8

Los profesionales y administradores de las pruebas deben seguir las instrucciones de administración, incluyendo la calibración de equipos técnicos y la verificación de la exactitud y replicabilidad de los puntajes, y deben facilitar opciones para la evaluación que faciliten el desempeño de los examinandos.

Comentario: Debido a que los datos normativos con respecto a los cuales se evaluará el desempeño de un examinando se recopilaron bajo procedimientos informados estándar, el profesional deberá conocer y tener en cuenta el efecto que cualquier procedimiento no estándar podría tener en el puntaje obtenido de un examinando y en la interpretación de ese puntaje. Cuando se usan pruebas que emplean un formato de respuestas no estructurado (por ejemplo, algunas pruebas proyectivas), el profesional debe seguir las instrucciones de administración y aplicar criterios de calificación objetivos cuando sea apropiado y estén disponibles.

En algunos casos, las pruebas se pueden llevar a cabo en contextos realistas para determinar cómo responde el examinando en estos contextos. Por ejemplo, una evaluación de trastorno de atención se podría realizar en un entorno ruidoso o perturbador, en lugar de hacerlo en un entorno que proteja al examinando contra los riesgos externos para la eficiencia del desempeño.

Estándar 10.9

Cuando se decide sobre el uso de administración de pruebas basada en tecnología, los profesionales deben tener en cuenta el propósito de la evaluación, el constructo que se mide y las capacidades del examinando.

Comentario: El control de calidad debe ser parte integral de la administración de pruebas computarizadas o basadas en tecnología. Algunas pruebas basadas en tecnología pueden requerir que los examinandos tengan la oportunidad de recibir instrucción y practicar antes de la administración, a menos que el propósito de la prueba sea evaluar la capacidad del uso de tales equipos. El profesional tiene la responsabilidad de determinar si la administración basada en tecnología se debe monitorizar, o si se necesita personal de soporte técnico para ayudar con el uso de los equipos y el software de la prueba. Si la prueba no se monitorizó o si no había personal de soporte disponible, se debe informar al intérprete de los puntajes de la prueba.

Unidad 4. Interpretación de pruebas

Estándar 10.10

Quienes seleccionan pruebas o interpretan sus resultados no deben permitir que individuos o grupos con intereses adquiridos en los resultados de una evaluación tengan una influencia inapropiada en la interpretación de estos resultados.

Comentario: Los grupos o individuos con un interés adquirido en la relevancia o el significado de las conclusiones de evaluaciones psicológicas pueden incluir, entre otros, a empleadores, profesionales de la salud, representantes legales, personal de escuelas, terceros pagadores y miembros de la familia. En algunos casos, pueden existir requisitos legales que limiten la capacidad de un profesional para impedir que interpretaciones inapropiadas de las evaluaciones afecten a las decisiones, pero los profesionales tienen la obligación de documentar cualquier discrepancia en tales circunstancias.

Estándar 10.11

Cuando sea apropiado o lo exija la ley, los profesionales deben compartir los puntajes e interpretaciones de las pruebas con el examinando. Esta información se debe expresar en un lenguaje que el examinando (o, si corresponde, el representante legal del examinando) pueda comprender.

Comentario: Los puntajes e interpretaciones de las pruebas se deben expresar en términos que puedan ser entendidos fácilmente por el examinando u otros facultados para acceder a los resultados. En la mayoría de los casos, un reporte debe generarse y estar disponible para la fuente de referencia. Ese reporte deberá adherirse a los estándares requeridos

por la profesión y/o la fuente de referencia, y la información se deberá documentar de manera que sea comprensible para la fuente de referencia. En algunas situaciones clínicas, podría ser perjudicial compartir observaciones con el examinando. Se debe prestar atención para minimizar las consecuencias imprevistas de las observaciones de una prueba. Cualquier divulgación de los resultados de una prueba a un individuo o cualquier decisión de no divulgar tales resultados deberá ser coherente con los estándares legales vigentes, por ejemplo, con las leyes de privacidad.

Estándar 10.12

En la evaluación psicológica, la interpretación de puntajes de pruebas o de patrones de resultados de una batería de pruebas debe considerar otros factores que pueden influir en una determinada conclusión de la evaluación. Si procede, se debe incluir en el reporte una descripción de tales factores y un análisis de las hipótesis o explicaciones alternativas que pueden haber contribuido a los patrones de resultados.

Comentario: Existen muchos factores que pueden influir en los resultados de pruebas individuales o en las conclusiones generales de una evaluación psicológica (por ejemplo, la cultura, el género, la raza u origen étnico, el nivel educativo, el ser diestro o zurdo, el estado mental actual, el estado de salud, las preferencias lingüísticas y la situación de la prueba). Cuando se preparan las interpretaciones de los puntajes de una prueba y los reportes extraídos de una evaluación, los profesionales deben considerar el grado en que esos factores pueden introducir varianza irrelevante de constructo en los resultados de la prueba. Cuando sea posible o apropiado, también se debe informar la interpretación de los resultados de una prueba en el proceso de evaluación a través de un análisis de las características estilísticas y cualitativas del comportamiento en la realización de la prueba, que se puedan obtener de las observaciones, entrevistas e información histórica. La inclusión de información cualitativa puede ser útil para la comprensión de las conclusiones de las pruebas y evaluaciones. Además, a menudo se usan pruebas de esfuerzo o simulación para determinar la posibilidad de fraude o simulación.

Estándar 10.13

Cuando la validez de un diagnóstico se valora mediante la evaluación del nivel de concordancia entre interpretaciones de los puntajes de una prueba y el diagnóstico, los términos o categorías diagnósticas empleadas se deben definir o identificar cuidadosamente.

Comentario: Dos sistemas de diagnóstico que se usan de forma habitual son los relacionados con la psiquiatría (es decir, basados en el *Manual diagnóstico y estadístico de los trastornos mentales*) y con la salud (es decir, basados en la *Clasificación internacional de enfermedades*). Se debe anotar el sistema usado para diagnosticar al examinando, según sea pertinente. Algunos síndromes (p. ej., deterioro cognitivo leve, discapacidad del aprendizaje social) no aparecen en ninguno de los sistemas; para estos, se debe usar una descripción de las deficiencias, con el diagnóstico más cercano posible.

Estándar 10.14

Cuando los profesionales presenten recomendaciones o decisiones en términos de base actuarial, debe estar disponible la evidencia de validación relacionada con los criterios.

Comentario: Las interpretaciones de las pruebas no deben implicar la existencia de evidencia empírica de una relación entre los resultados de pruebas específicas, intervenciones prescritas y conclusiones deseadas, a menos que tal evidencia esté disponible para poblaciones similares a las representativas del examinando.

Estándar 10.15

La interpretación de los resultados de una prueba o de una batería de pruebas para fines diagnósticos se debe basar en varias fuentes de pruebas e información colateral, y en el conocimiento de los principios normativos, empíricos

y teóricos, así como en las limitaciones, de tales pruebas y datos.

Comentario: Un patrón determinado de desempeños en pruebas representa una vista transversal del individuo que se evalúa en un contexto específico. En tales contextos, las interpretaciones de las conclusiones derivadas de una compleja batería de pruebas requieren una educación apropiada en (experiencia supervisada con y conocimiento de) las limitaciones procedimentales, teóricas y empíricas de las pruebas y del procedimiento de evaluación.

Estándar 10.16

Si un editor sugiere que las pruebas se van a utilizar en combinación, el profesional debe revisar los procedimientos recomendados y la evidencia para la combinación de pruebas, y determinar si la justificación proporcionada por el editor es apropiada para la combinación específica de las pruebas y los usos previstos.

Comentario: Por ejemplo, si medidas de inteligencia se presentan con medidas de memoria, o si medidas de intereses y estilos de personalidad se presentan juntas, deberán estar disponibles los datos de respaldo de validación y confiabilidad/precisión para esas combinaciones de puntajes e interpretaciones.

Estándar 10.17

Quienes usan interpretaciones generadas por computadora de los datos de una prueba deben verificar que la calidad de la evidencia de validación es suficiente para las interpretaciones.

Comentario: Los esfuerzos para reducir un conjunto complejo de datos en las interpretaciones generadas por computadora de un determinado constructo podrían traducirse en análisis equívocos o simplificados en exceso de los significados de los puntajes, que a su vez pueden llevar a decisiones diagnósticas y pronósticos fallidos. Se deberá revisar la relevancia e idoneidad de las normas en las que se basan las interpretaciones.

Unidad 5. Seguridad de pruebas

Estándar 10.18

Los profesionales y otros que tengan acceso a los materiales y resultados de las pruebas deben mantener la confidencialidad de los resultados y materiales de la evaluación de manera coherente con los requisitos científicos, profesionales, legales y éticos. Las pruebas (incluyendo versiones obsoletas) no deben estar disponibles para el público ni revenderse a usuarios de la prueba no cualificados.

Comentario: Los profesionales deben tener amplios conocimientos sobre, y guardar conformidad con, el mantenimiento de registros y las directrices de confidencialidad exigidos por la ley federal vigente y en las jurisdicciones donde ejerzan, así como de las directrices de las organizaciones profesionales a las que pertenezcan. Es posible que el editor y el usuario de la prueba, el examinando y terceras partes (p. ej., escuela, juzgado, empleador) tengan diferentes niveles de comprensión o reconocimiento de la necesidad de confidencialidad de los materiales de las pruebas. En la medida de lo posible, el profesional que usa una prueba es responsable de gestionar la confidencialidad de la información de la prueba entre todas las partes. Es importante que el profesional tenga presente las posibles amenazas a la confidencialidad y los recursos legales y profesionales disponibles. Asimismo, los profesionales tienen la responsabilidad de mantener la seguridad de los materiales de la evaluación y respetar los derechos de autor de todas las pruebas. La distribución, presentación o reventa de materiales de una prueba (incluyendo ediciones obsoletas) a destinatarios no autorizados infringe los derechos de autor de los materiales y pone en riesgo la seguridad de la prueba. Cuando sea necesario revelar el contenido de una prueba en el proceso de explicación de los resultados o en un proceso judicial, se debe llevar a cabo en un entorno controlado. Siempre que sea posible, no se deben distribuir copias del contenido o se deben distribuir de manera que se proteja la seguridad de la prueba en la medida de lo posible.

11. PRUEBAS Y ACREDITACIÓN EN EL CENTRO DE TRABAJO

ANTECEDENTES

Las organizaciones usan las pruebas de empleo para muchos fines, incluyendo la selección, asignación y promoción de empleados. En general, la *selección* se refiere a las decisiones sobre las personas que entran a trabajar en la organización; la *asignación* se refiere a las decisiones sobre la asignación de personas a puestos dentro de la organización; y la *promoción* se refiere a las decisiones sobre el desarrollo profesional de las personas dentro de la organización. Los que estos tres fines tienen en común es que se centran en la predicción de futuras conductas en el trabajo, con el objetivo de influir en los resultados organizacionales como, por ejemplo, la eficiencia, el crecimiento, la productividad y la motivación y la satisfacción de los empleados.

Las pruebas que se usan en procesos de obtención de licencias y certificación, que denominaremos aquí *acreditación* de forma general, se centran en las capacidades o competencias actuales de la persona postulante en un dominio específico. En muchas ocupaciones, los profesionales individuales deben obtener una licencia de los organismos gubernamentales. En otras ocupaciones, son las sociedades profesionales, los empleadores u otras organizaciones quienes asumen la responsabilidad de la acreditación. Aunque la obtención de licencias suele implicar la entrega de una credencial para entrar en una profesión, los programas de acreditación pueden existir en varios niveles, de principiante a experto en un campo determinado. Por lo general, la certificación se busca voluntariamente, aunque las ocupaciones difieren en el grado en que la obtención de una certificación influye en la inserción o el avance laboral. El proceso de acreditación puede incluir evaluación y otros requisitos, como educación y experiencias supervisadas. Los *Estándares* se aplican al uso de pruebas como un componente de un proceso más amplio de acreditación.

Asimismo, en los centros de trabajo, las pruebas se utilizan para muchos otros fines distintos a las decisiones de personal y la acreditación. La evaluación como herramienta para el crecimiento personal puede ser parte de programas de capacitación y desarrollo, en los que los instrumentos que miden características personales, intereses, valores, preferencias y estilos de trabajo se usan con el objetivo de proporcionar autoconocimiento a los empleados. Las pruebas también pueden tener lugar en el contexto de evaluación de programas, como en el caso de estudios experimentales sobre la eficacia de un programa de capacitación, donde las pruebas se administran como pre y post medidas. Algunas evaluaciones realizadas en contextos de empleo (como las entrevistas de trabajo no estructuradas para las que no se hacen alegaciones de validez predictiva) son, por naturaleza, no estandarizadas y, por lo general, no resulta factible aplicar los estándares a tales evaluaciones. Sin embargo, el tema central de este capítulo es el uso específico de pruebas para las decisiones de personal y la acreditación. En otros capítulos se examinan muchos otros temas pertinentes al uso de las pruebas en contextos organizacionales: cuestiones técnicas en los capítulos 1, 2, 4 y 5; documentación en el capítulo 7; y evaluación individualizada psicológica y de personalidad de candidatos de empleo en el capítulo 10.

Como se describe en el capítulo 3, el ideal de imparcialidad en las pruebas se consigue si un determinado puntaje de prueba tiene el mismo significado para todos los individuos y no se ve influido de manera sustancial por barreras irrelevantes de constructo en el desempeño de los individuos. Por ejemplo, una persona con discapacidad visual puede tener dificultades para leer las preguntas en un inventario de personalidad u otras evaluaciones vocacionales que se presentan con una letra pequeña. Las personas jóvenes que acaban de incorporarse al personal podrían ser menos sofisticadas en estrategias de realización de pruebas que los postulantes de trabajo más

experimentados y, por lo tanto, sus puntajes se verían afectados. Una persona no familiarizada con la tecnología puede tener dificultades con la interfaz de usuario en una evaluación con simulaciones computarizadas. En cada uno de estos casos, el desempeño se ve obstaculizado por una fuente de varianza que no está relacionada con el constructo de interés. Una práctica de pruebas correcta supone una supervisión cuidadosa de todos los aspectos del proceso de evaluación y tomar las medidas apropiadas cuando se requieren para evitar ventajas o desventajas indebidas de algunos candidatos, causadas por factores no relacionados con el constructo que se evalúa.

Pruebas de empleo

La influencia del contexto en el uso de la prueba

Las pruebas de empleo comportan el uso de la información de la prueba como ayuda en la toma de decisiones sobre el personal. Tanto el contenido como el contexto de las pruebas de empleo pueden variar en gran medida. El contenido puede abarcar varios dominios de conocimientos, capacidades, habilidades, rasgos, actitudes, valores y otras características individuales. Algunas características contextuales representan elecciones hechas por la organización empleadora; otras representan restricciones que deben tenerse en cuenta por esa misma organización. Las decisiones sobre el diseño, la evaluación y la implementación del sistema de evaluación son específicas al contexto donde se va a usar el sistema. Entre las características contextuales importantes se incluye las siguientes:

Conjunto de candidatos internos vs. externos. En algunos casos, como en contextos de promoción, los candidatos que se someten a la prueba son ya empleados de la organización. En otros, se buscan solicitudes de individuos que no pertenezcan a la organización. También se puede dar el caso de que se busque una combinación de candidatos internos y externos.

Candidatos cualificados vs. no cualificados. En algunos casos, se buscan personas con poca cualificación en conocimientos o capacidades especializadas, ya que el trabajo no requiere estas especializaciones o porque la organización tiene previsto ofrecer capacitación una vez contratadas. En otros casos, se buscan trabajadores cualificados o con experiencia, con la expectativa de que puedan desempeñar de inmediato un trabajo especializado. Por lo tanto, un trabajo específico puede requerir sistemas de selección muy diferentes, en función de la contratación o promoción de individuos cualificados o no cualificados.

Corto plazo vs. largo plazo. En algunos casos, el objetivo del sistema de selección es predecir el desempeño inmediatamente o poco después de la contratación. En otros casos, el interés es el desempeño a largo plazo, como en el caso de predicciones que se refieren a la posibilidad de que los candidatos lleven a cabo satisfactoriamente una tarea asignada en el extranjero y a lo largo de varios años. Las cuestiones sobre el cambio de tareas y requisitos del trabajo también pueden llevar a centrarse en los conocimientos, capacidades, habilidades y otras características que se prevén necesarias para el desempeño del trabajo objetivo en el futuro, incluso si no son parte de la configuración actual del trabajo.

Cribado de inclusión vs. cribado de exclusión. En algunos casos, el objetivo del sistema de selección es cribar a los individuos que pueden ofrecer un alto desempeño en un conjunto de criterios de conducta o de resultados de interés para la organización. En otros, el objetivo es hacer una criba de exclusión de las personas que probablemente tendrían un desempeño deficiente. Por ejemplo, es posible que una organización quiera descartar a una pequeña proporción de individuos que presentan un alto riesgo de comportamiento patológico, anormal, contraproducente o criminal. La misma organización puede requerir un cribado de inclusión de personas con una alta probabilidad de desempeño óptimo.

Toma de decisiones mecánica vs. crítica. En algunos casos, la información de la prueba se usa de manera automatizada y mecánica. Este es el caso cuando los puntajes de una batería de pruebas se

combinan mediante fórmulas y los candidatos se seleccionan en un estricto orden descendente de clasificación, o cuando únicamente los candidatos con puntajes de corte específicos resultan elegibles para continuar con las fases posteriores de un sistema de selección. En otros casos, la información de una prueba se integra críticamente con la información de otras pruebas y con información externa a las pruebas para formar una evaluación general del candidato.

Uso continuo vs. uso puntual de una prueba. En algunos casos, una prueba se puede usar en una organización a lo largo de un periodo extenso, permitiendo la acumulación de datos y experiencias en el uso de la prueba en ese contexto. En otros casos, la preocupación sobre la seguridad de la prueba hace que el uso repetido no sea factible y se requiere una nueva prueba en cada administración. Por ejemplo, una prueba de trabajo para socorristas donde se requiera el rescate de un maniquí desde el fondo de una piscina no se ve afectada si los candidatos tienen un conocimiento detallado de la prueba con antelación. Por el contrario, una prueba escrita de conocimientos para agentes de policía puede verse seriamente afectada si algunos candidatos tienen acceso por adelantado a la prueba. La cuestión clave es si el conocimiento previo del contenido de una prueba afecta de forma indebida el desempeño de los candidatos y, en consecuencia, cambia el constructo medido por la prueba y la validez de las inferencias basadas en los puntajes.

Conjunto fijo vs. flujo continuo de candidatos. En algunos casos, se puede reunir un conjunto de candidatos antes del comienzo del proceso de selección, como sucede cuando la política de una organización es considerar a todos los candidatos que se presenten antes de una fecha específica. En otros casos, hay un flujo continuo de postulantes sobre los que se debe tomar decisiones de empleo de forma continuada. En el caso de un conjunto fijo, es posible una clasificación de los candidatos; en el caso de un flujo continuo, es posible que la decisión sobre cada candidato deba tomarse independientemente de la información de otros candidatos.

Tamaño de muestra pequeño vs. grande. El tamaño de una muestra afecta al grado de uso de distintas líneas de evidencia para el examen de la validez e imparcialidad de las interpretaciones de los puntajes para los usos previstos de las pruebas. Por ejemplo, para tamaños de muestra pequeños, no resulta técnicamente factible basarse en el contexto local para establecer relaciones empíricas entre la prueba y los puntajes de criterios. En pruebas de empleo, los tamaños de muestra suelen ser pequeños; el ejemplo extremo es un trabajo con un solo titular. En ocasiones, están disponibles tamaños de muestra grandes cuando hay varios titulares para el trabajo, cuando varios trabajos comparten requisitos similares y se pueden agrupar, o cuando organizaciones con trabajos similares colaboran para desarrollar un sistema de selección.

Un nuevo trabajo. Un caso especial del problema de un tamaño de muestra pequeño se produce cuando se crea un nuevo trabajo y no hay titulares para el mismo. A medida que surgen nuevos trabajos, los empleadores necesitan procedimientos de selección para cubrir los nuevos puestos. Se puede usar el juicio profesional para identificar pruebas de empleo apropiadas y proporcionar una justificación para el programa de selección, incluso si la variedad de métodos para documentar la validez presenta limitaciones. Aunque es raro que la evidencia de validación basada en estudios orientados a criterios se pueda recabar antes de la creación de un nuevo trabajo, es posible usar métodos para generalizar la evidencia de validación en situaciones con tamaños de muestra pequeños (véase el análisis en la página 192 sobre contextos con muestras pequeñas), así como estudios orientados a criterios que trabajan con expertos en la materia responsables de diseñar el trabajo.

Tamaño del conjunto de candidatos relativo al número de vacantes de trabajo. El tamaño del conjunto de candidatos puede limitar el tipo de sistema de evaluación viable. En el caso de trabajos atractivos, puede existir un alto número de candidatos y se podrían usar pequeñas pruebas de cribado para reducir el conjunto a un tamaño

práctico para la administración de pruebas más caras y prolongadas. Grandes conjuntos de candidatos también pueden comportar problemas de seguridad de la prueba, limitando a la organización a métodos de evaluación que permitan una administración simultánea a todos los candidatos.

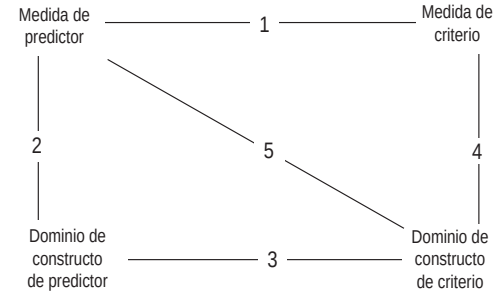
Por lo tanto, el uso de la prueba por parte de los empleadores está condicionado por las características contextuales. El conocimiento de estas características juega un papel importante en el juicio profesional que influirá en los tipos de sistemas de evaluación desarrollados y en las estrategias usadas para evaluar críticamente la validez de las interpretaciones de los puntajes para los usos previstos de la prueba.

El proceso de validación en las pruebas de empleo

Con frecuencia, el proceso de validación empieza con un análisis del trabajo en el que se recopila información sobre las funciones y las tareas, las responsabilidades, las características del trabajador y otros datos pertinentes. Esta información proporciona una base empírica para la articulación de lo que se entiende como *desempeño profesional* del trabajo en consideración, para el desarrollo de medidas del desempeño y para las características hipotesizadas de los individuos que pueden ser predictivas del desempeño.

En la mayoría de las aplicaciones de evaluación en contextos de empleo, la inferencia fundamental a extraer de los puntajes de las pruebas se relaciona con la predicción: el usuario de la prueba quiere formular una inferencia a partir de los resultados de la prueba para determinados comportamientos o resultados laborales futuros. Incluso cuando la estrategia de validez utilizada no implica relaciones empíricas con parámetros predictores, como en el caso de la evidencia de validación basada en el contenido de una prueba, existe un criterio implícito. Por tanto, aunque se podrían usar distintas estrategias para recopilar la evidencia, la inferencia que se respalda es que los puntajes de la prueba se puedan usar para predecir comportamientos de trabajo posteriores. El proceso de validación en contextos de empleo conlleva la recopilación y evaluación de la evidencia pertinente para sostener o cuestionar esta inferencia. Como se explicó anteriormente en el capítulo 1 (en la sección "Evidencia basada en relaciones con otras variables"), se puede usar una variedad de estrategias de validez para respaldar la inferencia.

Por lo tanto, establecer esta inferencia predictiva requiere prestar atención a dos dominios: el dominio de la prueba (el predictor) y el dominio del comportamiento o resultado de trabajo de interés (el criterio). Evaluar el uso de una prueba para una decisión de empleo se puede considerar como una evaluación de la hipótesis de la vinculación entre estos dominios. Operativamente, existen muchas formas de relacionar estos dominios, como ilustra el siguiente diagrama.



Vinculaciones alternativas entre las medidas de predictor y de criterio

El diagrama diferencia entre un dominio de constructo de predictor y una medida de predictor, y entre un dominio de constructo de criterio y una medida de criterio. Un *dominio de constructo de predictor* se define al especificar el conjunto de comportamientos, capacidades, habilidades, rasgos, actitudes y valores que se incluirán bajo etiquetas de constructo especificas (p. ej., razonamiento verbal, velocidad al teclear, diligencia). De forma similar, un *dominio de constructo de criterio* especifica un conjunto de comportamientos o resultados de trabajo que se incluirán bajo etiquetas de constructo específicas (p. ej., desempeño de tareas básicas, trabajo en equipo, concurrencia, volumen de ventas, desempeño general de trabajo). Las medidas de predictor y de criterio pretenden evaluar la situación de un individuo

respecto de las características evaluadas en esos dominios.

El diagrama enumera inferencias sobre un número de vinculaciones que suelen ser de interés. La primera vinculación (con la etiqueta 1 en el diagrama) se da entre los puntajes de una medida de predictor y los puntajes de una medida de criterio. Esta inferencia se prueba a través de exámenes empíricos de las relaciones entre las dos medidas. Las vinculaciones segunda y cuarta (con las etiquetas 2 y 4) son conceptualmente similares: Ambas examinan la relación de una medida operativa con el dominio de constructo de interés. Entre las formas de evidencia que se pueden examinar en la evaluación de estas vinculaciones están el análisis lógico, el juicio experto y la convergencia con (o la divergencia de) medidas conceptualmente similares o diferentes. La vinculación 3 implica la relación entre el dominio de constructo de predictor y el dominio de constructo de criterio. La vinculación inferida se establece sobre la base de un análisis teórico y lógico. Habitualmente, se basa en una evaluación sistemática del contenido del trabajo y el juicio experto sobre las características individuales relacionadas con un desempeño de trabajo óptimo. La vinculación 5 examina una relación directa de la medida de predictor con el dominio de constructo de criterio.

Algunas medidas de predictor están diseñadas explícitamente como ejemplos de dominios de constructo de criterio de interés; por lo tanto, el isomorfismo entre la medida y el dominio de constructo constituye una evidencia directa de la vinculación 5. Establecer la vinculación 5 de este modo es el signo característico de enfoques que dependen en gran medida de lo que los *Estándares* denominan *evidencia de validación basada en el contenido de la prueba*. Las pruebas donde los candidatos a puestos de socorristas realizan operaciones de rescate, o en las que candidatos a puestos de procesadores de datos escriben y editan textos, ofrecen ejemplos de contenido de pruebas que forman la base de validez.

Un requisito previo al uso de una medida de predictor en la selección de personal es que se establezcan las inferencias relativas a la vinculación entre la medida de predictor y el dominio de constructo de criterio. Como ilustra el diagrama, existen diversas estrategias para establecer esta vinculación esencial. Una estrategia es directa, a través de la vinculación 5; una segunda implica el emparejamiento de las vinculaciones 1 y 4; y una tercera comporta el emparejamiento de las vinculaciones 2 y 3.

Cuando la prueba está diseñada como un ejemplo del dominio de constructo de criterio, la evidencia de validación se puede establecer directamente a través de la vinculación 5. Otra estrategia para la relación entre una medida de predictor y el dominio de constructo de criterio se centra en las vinculaciones 1 y 4: emparejar una vinculación empírica entre el predictor y las medidas de criterio con la evidencia de idoneidad con la que la medida de criterio representa el dominio de constructo de criterio. La vinculación empírica entre la medida de predictor y la medida de criterio es parte de lo que los *Estándares* denominan *evidencia de validación basada en relaciones con otras variables*. La vinculación empírica de la prueba y la medida de criterio debe complementarse con la evidencia de relevancia de la medida de criterio para el dominio de constructo de criterio, a fin de completar la vinculación entre la prueba y el dominio de constructo de criterio. La evidencia de relevancia de la medida de criterio para el dominio de constructo de criterio se basa normalmente en el análisis del trabajo, aunque en algunos casos la vinculación entre el dominio y la medida es tan directa que la relevancia es aparente sin el análisis del trabajo (p. ej., cuando el constructo de criterio de interés es el absentismo o la rotación). Observe que esta estrategia no se basa necesariamente en un dominio de constructo de predictor bien desarrollado. Las medidas de predictor como, por ejemplo, medidas de biodatos afinados empíricamente, se construyen sobre la base de vinculaciones empíricas entre las respuestas a los ítems de la prueba y la medida de criterio de interés. Tales medidas pueden, en algunos casos, desarrollarse sin una concepción plenamente establecida del dominio de constructo de predictor; la base para su uso es la vinculación empírica directa entre las respuestas de la prueba y una medida de criterio pertinente. A menos que los tamaños de

muestra sean muy grandes, la capitalización aleatoria puede ser un problema, en cuyo caso se deben tomar las medidas apropiadas (p. ej., validez cruzada).

Asimismo, otra estrategia para vincular los puntajes de predictor y el dominio de constructo de predictor se centra en emparejar la evidencia de idoneidad con la que la medida de predictor representa al dominio de constructo de predictor (vinculación 2) con la evidencia de vinculación entre el dominio de constructo de predictor y el dominio de constructo de criterio (vinculación 3). Como se observó anteriormente, no existe una sola ruta directa para establecer estas vinculaciones. Estas implican líneas de evidencia integradas bajo la "validez de constructo" en conceptualizaciones anteriores del proceso de validez. Es posible que una combinación de líneas de evidencia (p. ej., juicio experto de las características predictivas del éxito en el trabajo, inferencias extraídas de un análisis de incidentes críticos de desempeño eficaz o ineficaz, y métodos de entrevista y observación) puedan respaldar inferencias sobre constructos de predictor vinculados con el dominio de constructo de criterio. Las medidas de estos constructos de predictor pueden entonces seleccionarse o desarrollarse, y se puede establecer una vinculación entre la medida de predictor y el dominio de constructo de predictor, con varias líneas de evidencia para la vinculación 2, como se mencionó anteriormente.

Las diversas estrategias para vincular puntajes de predictor con el dominio de constructo de criterio pueden diferir en su aplicabilidad potencial a un contexto de pruebas de empleo determinado. Aunque la disponibilidad de algunas líneas de evidencia pueda ser limitada, estas limitaciones no reducen la importancia de establecer un argumento de validez para la inferencia predictiva.

Por ejemplo, los métodos para establecer vinculaciones son más limitados en contextos donde solo están disponibles pequeñas muestras. En estas situaciones, la recopilación de evidencia local de relaciones predictor-criterio no es factible y pueden resultar más útiles los métodos para generalizar la evidencia a partir de otros contextos. Existen una variedad de métodos para la generalización de la evidencia de validación de la interpretación de la inferencia predictiva a partir de otros contextos. La evidencia de validación se puede transportar directamente desde otro contexto cuando exista evidencia sólida (p. ej., un cuidadoso análisis del trabajo) que indique que el trabajo local es altamente comparable con el trabajo para el cual se están importando los datos de validez. Estos métodos se pueden basar en la evidencia de las vinculaciones 1 y 4 que ya ha sido establecida en otros estudios, como en el caso del estudio de transportabilidad descrito anteriormente. También se podría establecer la evidencia de la vinculación 1 usando técnicas como el metaanálisis para combinar los resultados de varios estudios, y un cuidadoso análisis del trabajo podría establecer la evidencia de la vinculación 4, mostrando que el trabajo focal es similar a otros trabajos incluidos en el metaanálisis. En el caso extremo, se podría desarrollar un sistema de selección para un trabajo de nueva creación que, en la actualidad, no tenga interesados. Aquí, la generalización de la evidencia de otros contextos puede resultar especialmente útil.

En muchas aplicaciones de evaluación, existe un considerable y creciente corpus de investigación que trata sobre algunas de las inferencias examinadas anteriormente, si no de todas. Una integración meta-analítica de esta investigación puede formar parte integral de la estrategia para vincular la información de la prueba con el dominio de constructo de interés. El valor de la recopilación de datos de validez locales varía con la magnitud, relevancia y coherencia de las conclusiones de investigaciones que usan medidas de predictor similares y dominios de constructo de criterio similares para trabajos similares. En algunos casos, un registro de investigación acumulada, pequeña y no sistemática puede llevar a una estrategia de validez que se base en gran medida en los datos locales; en otros, una base de investigación extensa y sistemática puede hacer innecesaria la inversión en recursos para la recopilación de datos locales adicionales.

Por lo tanto, se pueden utilizar diversas fuentes de datos y diversas líneas de evidencia para evaluar la vinculación entre una medida de predictor

y el dominio de constructo de criterio de interés. No hay un solo método de investigación de preferencia para el establecimiento de esta vinculación. Más bien, el usuario de la prueba debe considerar las características específicas de la situación de evaluación y aplicar el juicio profesional para desarrollar una estrategia para probar la hipótesis de una vinculación entre la medida de predictor y el dominio de criterio.

Bases para evaluar el uso de la prueba de empleo

Aunque el objetivo principal de las pruebas de trabajo es la predicción precisa de los comportamientos o resultados de trabajo posteriores, es importante reconocer que hay límites en el grado de predicción de tales criterios. No se puede alcanzar una predicción perfecta. En primer lugar, el comportamiento en el contexto de trabajo se ve influido por una amplia variedad de factores organizacionales y extraorganizacionales, incluyendo el entrenamiento del supervisor y los colegas, la capacitación formal e informal, el diseño del trabajo, las estructuras y sistemas de la organización, y las responsabilidades familiares, entre otros. En segundo lugar, el comportamiento en el contexto de trabajo también se ve influido por una amplia variedad de características individuales, incluyendo los conocimientos, capacidades, habilidades, personalidad y actitudes de trabajo, entre otros. Por lo tanto, cualquier característica única solo será un predictor imperfecto, e incluso complejos sistemas de selección solo se centran en el conjunto de constructos considerados más críticos para el trabajo, y no en todas las características que pueden influir en el comportamiento de trabajo. En tercer lugar, siempre se producirán errores de medida, incluso en pruebas y medidas de criterio bien desarrolladas.

Por lo tanto, los sistemas de evaluación no se pueden juzgar con respecto a un estándar de predicción perfecta. En su lugar, se deben juzgar en términos comparativos con métodos de selección alternativos disponibles. El juicio profesional, informado por el conocimiento de la literatura de investigación sobre el grado de precisión predictiva relacionado con las alternativas disponibles, influye en las decisiones sobre el uso de la prueba.

A menudo, las decisiones sobre el uso de la prueba se ven influidas por consideraciones adicionales, incluyendo la utilidad (es decir, la relación costo-beneficio) y el retorno de inversión, los juicios de valor sobre la importancia relativa de seleccionar un dominio de criterio en vez de otros, las preocupaciones sobre las reacciones de los postulantes ante el contenido y los procesos de la prueba, la disponibilidad e idoneidad de métodos de selección alternativos, y los requisitos legales o reglamentarios que rigen el uso, la imparcialidad y los objetivos de la política de la prueba, por ejemplo, la diversidad del personal. Sin duda, los valores organizacionales juegan un papel en las decisiones sobre el uso de la prueba; así, incluso organizaciones con evidencia comparable que respalda una inferencia prevista extraída de los puntajes de una prueba pueden alcanzar diferentes conclusiones sobre si se debe usar o no una prueba específica.

Pruebas en la acreditación profesional y ocupacional

Las pruebas son ampliamente utilizadas en la acreditación de personas para muchas actividades y profesiones. Los gobiernos federales, estatales y locales imponen requisitos legales para garantizar que quienes obtienen una licencia tiene los conocimientos y capacidades suficientes para realizar actividades profesionales importantes con seguridad y eficacia. La certificación juega un papel similar en muchas actividades no reguladas por los gobiernos y, con frecuencia, es un precursor necesario para la promoción. Asimismo, la certificación se ha usado también en gran medida para indicar que una persona tiene unas habilidades específicas (p. ej., manejo de equipos especializados de reparación automotriz), las cuales pueden ocupar solo una parte de sus tareas profesionales. En general, la obtención de licencias y la certificación se denominarán aquí *acreditación*.

Las pruebas usadas en la acreditación tienen como objetivo proporcionar al público,

incluyendo empleados y organismos gubernamentales, un mecanismo confiable para identificar profesionales que hayan cumplido estándares específicos. Los estándares pueden ser estrictos, pero no tan exigentes como para limitar indebidamente el derecho de individuos cualificados a ofrecer sus servicios al público. La acreditación también sirve para proteger al público mediante la exclusión de personas que se considera no cualificadas para realizar las tareas de una profesión u actividad. Las cualificaciones para la acreditación, por lo general, incluyen requisitos educativos, un determinado grado de experiencia supervisada y otros criterios específicos, así como la obtención de un puntaje aprobatorio en uno o más exámenes. Las pruebas se usan en la acreditación de un amplio espectro de profesiones y actividades, incluyendo la medicina, la profesión jurídica, la enseñanza, la arquitectura, la profesión inmobiliaria y la cosmetología. En algunas de estas actividades, como la ciencia actuarial, la neuropsicología clínica o las especialidades médicas, las pruebas se usan también para certificar los niveles avanzados de especialización. En algunas profesiones o actividades también es necesaria la renovación de licencias o la recertificación periódica.

Las pruebas usadas en la acreditación están diseñadas para determinar si el candidato domina los conocimientos y habilidades esenciales. El enfoque se pone en los estándares de competencia necesarios para un desempeño eficaz (p. ej., en la obtención de licencias esto se refiere a un desempeño práctico seguro y efectivo). Normalmente, el diseño de la prueba empieza con una definición adecuada de la actividad o especialidad, de manera que se pueda identificar claramente que las personas se dedican a esa actividad. A continuación, se describen la naturaleza y los requisitos de la actividad, en su forma actual. Para identificar los conocimientos y habilidades necesarias para un ejercicio competente, es importante llevar a cabo un análisis del trabajo real y documentar las tareas y responsabilidades esenciales de la actividad o profesión de interés. Se puede usar una amplia variedad de métodos empíricos, incluyendo la técnica de incidentes críticos, el análisis del trabajo, la evaluación de necesidades de capacitación o los estudios o encuestas de prácticas de los profesionales en ejercicio. A menudo, paneles de expertos en el área trabajan en colaboración con expertos en medición para definir las especificaciones de la prueba, incluyendo los conocimientos y habilidades necesarios para un desempeño seguro y eficaz, y la forma apropiada de evaluarlos. Los *Estándares* se aplican a todos los formatos de pruebas, incluyendo las pruebas tradicionales de selección múltiple y de selección de respuestas, porfolios, tareas de juicio situacional o exámenes orales. En la evaluación de estos componentes de práctica también se utilizan tareas de desempeño más elaboradas, a veces mediante simulaciones por computadora como, por ejemplo, en diagnóstico de pacientes o planificación del tratamiento. También es posible que se usen tareas de desempeño práctico (p. ej., el manejo de una grúa o una curación dental), con la observación y evaluación de uno o más examinadores.

Las pruebas de acreditación pueden abarcar un número de áreas de conocimiento o capacidad relacionadas, pero diferentes. El diseño del programa de pruebas incluye decidir qué áreas se van a incluir, si se usará una prueba o una serie de pruebas, y cómo se combinarán los puntajes de las pruebas para llegar a una decisión global. En algunos casos, se permite que los puntajes altos de una prueba desplacen (es decir, compensen) los puntajes bajos, de manera que resulta apropiada una combinación aditiva. En otros casos, se usa un modelo de decisión conjuntiva que requiere un desempeño aceptable en cada una de las pruebas de una serie de exámenes. Se debe considerar cuidadosamente el tipo de modelo de decisiones aprobado-reprobado para un programa de acreditación, y se debe articular la base conceptual o empírica de ese modelo de decisiones.

La validez de las pruebas de acreditación depende principalmente de la evidencia relacionada con el contenido, a menudo en la forma de juicios sobre la idoneidad de la prueba para representar el dominio de contenido asociado con la actividad o especialidad que se considera. Tal evidencia se puede complementar con otras formas de evidencia externas a la prueba. Por ejemplo, se podría proporcionar información sobre el proceso

mediante el cual se desarrollaron las especificaciones del contenido y la especialización de las personas que han emitido juicios sobre el dominio de contenido. La evidencia relacionada con los criterios es de aplicabilidad limitada porque los exámenes de acreditación no tienen como objetivo predecir el desempeño individual en un trabajo específico, sino más bien proporcionar evidencia de que los candidatos han adquirido los conocimientos, habilidades y juicios necesarios para un desempeño eficaz, a menudo en una amplia variedad de trabajos o contextos (usamos el término *juicio* para referirnos a la aplicación de conocimientos o capacidades en situaciones específicas). Además, normalmente las medidas de desempeño en ejercicio no están disponibles para quienes no obtienen una acreditación.

La definición del nivel mínimo de conocimientos y capacidades que requiere la obtención de licencias o la certificación es una de las tareas más importantes y difíciles para los responsables de la acreditación. La validez de la interpretación de los puntajes de las pruebas depende de que el estándar para aprobar haga una distinción apropiada entre el desempeño correcto e incorrecto. A menudo, se usan paneles de expertos para especificar el nivel de desempeño que se establece como requisito. Los estándares deben ser lo bastante altos como para garantizar que el público, los empleadores y los organismos gubernamentales obtengan un servicio adecuado, pero no tan altos que se conviertan en limitaciones injustificadas. La verificación de la idoneidad de los puntajes de corte o de los puntajes de una prueba para la obtención de licencias o la certificación es un elemento crucial en el proceso de validez. El capítulo 5 ofrece un análisis general sobre la fijación de puntajes de corte (véase los Estándares 5.21—5.23 para ver temas específicos sobre los puntajes de corte).

En ocasiones, los órganos legislativos intentan legislar sobre un puntaje de corte, por ejemplo, un 70% de ítems de prueba respondidos correctamente. Los puntajes de cortes que se establecen de este modo tan arbitrario pueden ser perjudiciales por dos razones. En primer lugar, sin información detallada sobre la prueba, los requisitos del trabajo y su relación, es imposible establecer un estándar correcto. En segundo lugar, sin información detallada sobre el formato de la prueba y la dificultad de los ítems, estos puntajes de corte arbitrarios carecen de significado.

Los puntajes de las pruebas de acreditación deben ser precisos en las inmediaciones del puntaje de corte. Es posible que no necesiten ser tan precisos para los examinandos que aprueban o reprueban con claridad. Las pruebas de destreza basadas en computadora pueden incluir una disposición para finalizar la prueba cuando resulta evidente que se puede tomar una decisión sobre el desempeño de los candidatos, lo que se traduce en una prueba más corta para los candidatos cuyo rendimiento claramente supera o está por debajo del desempeño mínimo requerido para un puntaje de aprobación. Debido a que las pruebas de destreza pueden no estar diseñadas para proporcionar resultados exactos para el rango completo de puntajes, muchas de estas pruebas reportan los resultados simplemente como "aprobado" o "reprobado". Cuando los candidatos reciben comentarios sobre su desempeño, se requiere precisión para todo el rango de puntajes. Los errores estándar de medida condicional, examinados en el capítulo 2, proporcionan información sobre la precisión de puntajes específicos.

Los candidatos que reprueben pueden encontrar útil la información sobre las áreas en las tuvieron un desempeño especialmente deficiente. Esta es la razón por la que a veces se suministran subpuntajes. A menudo, los subpuntajes se basan en un número relativamente pequeño de ítems y pueden ser mucho más confiables que el puntaje total. Además, las diferencias entre los subpuntajes pueden reflejar simplemente un error de medida. Por estas razones, la decisión de proporcionar subpuntajes a los candidatos debe considerarse cuidadosamente, y se debe facilitar la información para una interpretación apropiada. En el capítulo 2 y el estándar 2.3 se trata la importancia de la confiabilidad de los subpuntajes.

Debido a que la acreditación suele acarrear riesgos altos y es un proceso continuo, con pruebas aplicadas mediante un programa regular, normalmente no es deseable usar el mismo formulario

de prueba repetidamente. Por lo tanto, generalmente se necesitan nuevos formularios o versiones de una prueba de forma periódica. Desde una perspectiva técnica, todos los formularios de una prueba se deben preparar con las mismas especificaciones, evaluar el mismo contenido y usar la misma ponderación de componentes o temas.

Los formularios de pruebas alternativos deben tener la misma escala de puntajes, de manera que estos puedan retener su significado. Se pueden usar varios métodos de vinculación o equiparación de formularios alternativos para garantizar que el estándar para la aprobación representa el mismo nivel de desempeño en todos los formularios. Observe que la divulgación de formularios de pruebas anteriores puede poner en riesgo el grado de comparabilidad de distintos formularios de pruebas.

La práctica de profesiones y actividades suele cambiar a lo largo del tiempo. Las restricciones legales cambiantes, el progreso en áreas científicas y el refinamiento de las técnicas pueden implicar la necesidad de cambios en el contenido de la prueba. Cada profesión o actividad debe reevaluar periódicamente los conocimientos y habilidades medidos en los exámenes que se utilizaron para cumplir los requisitos de la acreditación. Cuando el cambio es sustancial, será necesario revisar la definición de la profesión y el contenido de la prueba para reflejar las circunstancias cambiantes. Estos cambios en la prueba podrían alterar el significado de la escala de puntajes. Cuando se hacen revisiones importantes en la prueba o cuando cambia la escala de puntajes, se debe volver a establecer el puntaje de corte.

Algunos grupos de acreditación consideran necesario, como asunto práctico, ajustar periódicamente el puntaje de aprobación u otros criterios para regular el número de candidatos acreditados que acceden a la profesión. Este procedimiento es cuestionable y presenta graves problemas para la calidad técnica de los puntajes de las pruebas, y amenaza la validez de la interpretación de un puntaje de aprobación como indicador de competencias de nivel básico. Ajustar periódicamente los puntajes de corte también implica que los estándares tendrán un nivel más alto en unos años que en otros, una práctica que es difícil de justificar en términos de calidad del desempeño. A veces, la escala de puntajes se ajusta de manera que un número determinado o una proporción de candidatos alcancen el puntaje de aprobación. Este método, aunque menos obvio para los candidatos que cambiar el puntaje de corte, también es técnicamente inapropiado ya que cambia el significado de los puntajes de un año a otro. Aprobar un examen de acreditación debe significar que el candidato cumple con los estándares de conocimientos y capacidades establecidos por el órgano de acreditación para garantizar un ejercicio eficaz.

Los problemas de engaño y seguridad de la prueba son de especial importancia en la realización de pruebas de acreditación. Los problemas de seguridad se tratan en los capítulos 6 y 9. Los problemas de engaño por parte de los examinandos se examinan en el capítulo 8 (véase los Estándares 8.9-8.12, que tratan sobre las irregularidades en las pruebas).

La imparcialidad y el acceso, temas del capítulo 3, son importantes para las pruebas de obtención de licencias y certificación. La evaluación de una adecuación o modificación de una prueba de acreditación deberá tener en cuenta las funciones críticas que se realizan en el trabajo de interés de la prueba. En el caso de las pruebas de acreditación, la criticalidad de las funciones del trabajo se basa en el interés público, así como en la propia naturaleza del trabajo. Cuando una condición limita la capacidad de un individuo para realizar una función crítica de un trabajo, es posible que no resulte apropiado adecuar o modificar el examen de obtención de licencia o certificación (es decir, algunos cambios pueden alterar sustancialmente factores que el examen tiene previsto medir para la protección de la seguridad, el bienestar y la salud pública).

ESTÁNDARES PARA PRUEBAS Y ACREDITACIÓN EN EL CENTRO DE TRABAJO

Los estándares de este capítulo se han separado en tres unidades temáticas denominadas de la siguiente manera:

1. Estándares aplicables con carácter general a las pruebas y la acreditación en el centro de trabajo
2. Estándares para las pruebas de empleo
3. Estándares para la acreditación

Unidad 1. Estándares aplicables con carácter general a las pruebas y la acreditación en el centro de trabajo

Estándar 11.1

Antes del desarrollo e implementación de una prueba de empleo o acreditación, se debe hacer una clara formulación de las interpretaciones previstas de los puntajes para los usos especificados. Las tareas de validez posterior se deben diseñar para determinar en qué medida se ha logrado esto para todos los subgrupos pertinentes.

Comentario: Los objetivos de las pruebas de empleo y acreditación pueden variar considerablemente. Algunas pruebas de empleo se usan para descartar a las personas menos capacitadas para el trabajo en cuestión, en tanto que otras están diseñadas para identificar a las personas más idóneas para ese trabajo. Las pruebas de empleo también varían en aspectos del comportamiento de trabajo que trata de predecir, lo que puede incluir la cantidad y calidad del trabajo, conductas contraproducentes, trabajo en equipo, etc. Las pruebas de acreditación y algunas pruebas de empleo están diseñadas para identificar candidatos que satisfagan un determinado nivel de competencia en un dominio objetivo de conocimientos, habilidades y juicios.

Estándar 11.2

La evidencia de validación basada en el contenido de la prueba requiere una definición exhaustiva y explícita del dominio de contenido de interés.

Comentario: En general, el dominio de contenido de un trabajo para una prueba de empleo se debe describir en términos de las tareas que se realizan y los conocimientos, capacidades, habilidades y otras características que el trabajo requiere. Se deben definir claramente de manera que se puedan asociar con el contenido de la prueba. Los conocimientos, habilidades, capacidades y otras características incluidas en el dominio de contenido deben ser aquellos que los postulantes cualificados ya tengan cuando se les considera para el trabajo en cuestión. Además, no se debe esperar que la importancia de estas características para el trabajo en consideración cambie sustancialmente a lo largo del tiempo.

Para pruebas de acreditación, el dominio de contenido objetivo consiste, por lo general, en conocimientos, habilidades y juicios necesarios para un desempeño eficaz. El dominio de contenido objetivo se debe definir claramente de manera que se pueda asociar con el contenido de la prueba.

Estándar 11.3

Cuando el contenido de la prueba es una fuente primaria de evidencia de validación que respalda la interpretación del uso de una prueba en decisiones de empleo o acreditación, se debe demostrar una estrecha relación entre el contenido de la prueba y el trabajo o los requisitos profesionales/ocupacionales.

Comentario: Por ejemplo, si el contenido de la prueba muestra las tareas del trabajo con una fidelidad considerable (p. ej., con ejemplos de trabajo reales como el manejo de máquinas) o, a juicio de

los expertos, simula correctamente el contenido de las tareas del trabajo (p. ej., con determinados ejercicios de evaluación del centro) o si la prueba muestrea los conocimientos específicos del trabajo (p. ej., información necesaria para realizar algunas tareas) o las habilidades que se requieren para un desempeño competente, se puede ofrecer evidencia relacionada con el contenido como forma principal de evidencia de validación. Si la relación entre el contenido de la prueba y el contenido del trabajo no es clara y directa, cobrarán importancia otras líneas de evidencia de validación.

Cuando se presenta una evidencia de validación basada en el contenido de la prueba para un trabajo o tipo de trabajos, la evidencia debe incluir una descripción de las principales características del trabajo que la prueba se propone muestrear. A menudo resulta útil incluir información sobre la frecuencia relativa, la importancia y la criticalidad de los elementos. En un examen de acreditación, la evidencia deberá incluir una descripción de las principales responsabilidades, tareas o actividades desempeñadas por los profesionales que la prueba quiere muestrear, así como los conocimientos y capacidades fundamentales y necesarias para desempeñar esas responsabilidades, tareas o actividades.

Estándar 11.4

Cuando se integran varios puntajes de pruebas (o se integra información de puntajes e información que no procede de las pruebas) con la finalidad de tomar una decisión, se debe explicar claramente el papel que juega cada componente, y la inferencia formulada a partir de cada fuente de información se debe respaldar mediante evidencia de validación.

Comentario: En la acreditación, es posible que se exija a los candidatos obtener un puntaje igual o superior a un mínimo especificado en cada una de las pruebas (p. ej., un examen práctico basado en habilidades y una prueba de conocimientos de selección múltiple), o igual o superior a un puntaje de corte respecto de un puntaje agregado total. También pueden ser obligatorios algunos requisitos específicos de experiencia o nivel educativo. Se debe proporcionar una justificación y la evidencia de respaldo para cada uno de estos requisitos. En las pruebas y evaluaciones, la evidencia incluye, entre otros, la confiabilidad/precisión de los puntajes y la correlación entre las pruebas y evaluaciones.

En las pruebas de empleo, durante la toma de decisiones, la persona responsable puede integrar los puntajes de las pruebas con los datos de la entrevista, verificaciones de referencias y muchas otras fuentes de información. Las inferencias extraídas de los puntajes de las pruebas se deben limitar a las que cuentan con evidencia de validación disponible. Por ejemplo, en una prueba que mide un solo dominio pertinente muy concreto (como el conocimiento de trabajo) sería una inferencia incorrecta considerar un puntaje alto como indicador de idoneidad global para el trabajo (excluyendo, por tanto, la necesidad de verificar las referencias). En otras circunstancias, los responsables de tomar decisiones pueden integrar los puntajes de varias pruebas, o de varias escalas de una prueba.

Unidad 2. Estándares para las pruebas de empleo

Estándar 11.5

Cuando una prueba se usa para predecir un criterio, la decisión de llevar a cabo estudios empíricos locales de las relaciones predictor-criterio y la interpretación de los resultados se deben basar en el conocimiento de las investigaciones pertinentes.

Comentario: La literatura acumulada sobre la relación entre un tipo específico de predictor y un tipo de criterio puede ser suficientemente grande y sistemática como para respaldar la relación predictor-criterio sin investigación adicional. En algunos contextos, la literatura de investigación acumulada puede ser tan importante y sistemática que una conclusión dispar de un estudio acotado deberá tomarse con precaución, a menos que el estudio acotado sea excepcionalmente sólido. Los

estudios acotados tienen el máximo valor en contextos donde la literatura de investigación acumulada es escasa (p. ej., debido a la novedad del predictor o del criterio usado), donde el registro acumulado no es sistemático, o donde la literatura acumulada no incluye estudios similares al estudio del contexto local (p. ej., el estudio de una prueba con una literatura acumulada considerable que trata exclusivamente de trabajos de producción y un contexto local que abarca trabajos administrativos).

Estándar 11.6

La dependencia en la evidencia local de las relaciones predictor-criterio determinadas empíricamente como una estrategia de validez está supeditada a la determinación de la viabilidad técnica.

Comentario: La evidencia pertinente de las relaciones predictor-criterio está supeditada a un número de características, incluyendo (a) los trabajos que son relativamente estables y no de un periodo de rápida evolución; (b) la disponibilidad de una medida de criterio pertinente y confiable; (c) la disponibilidad de una muestra razonablemente representativa de la población de interés; y (d) un tamaño de muestra adecuado para estimar la solidez de la relación predictor-criterio. Si alguna de estas condiciones no se cumple, se deberá usar una estrategia de validez alternativa. Por ejemplo, como se observó en el comentario del Estándar 11.5, la literatura de investigación acumulada podría proporcionar una evidencia de validación sólida.

Estándar 11.7

Cuando la evidencia empírica de las relaciones predictor-criterio es parte de un patrón de evidencia usado para respaldar el uso de una prueba, las medidas de criterio usadas deben reflejar el dominio de constructo de criterio de interés para la organización. Todos los criterios deben representar comportamientos o resultados de trabajo importantes, ya sea en el trabajo o en la capacitación relacionada con el trabajo, según lo indique una revisión apropiada de la información sobre el trabajo.

Comentario: Cuando se construyen criterios para representar actividades o comportamientos de trabajo (p. ej., calificaciones de supervisión de subordinados en dimensiones importantes del trabajo), la recopilación sistemática de información sobre el trabajo deberá informar el desarrollo de las medidas de criterio. Sin embargo, no hay una opción clara entre los numerosos métodos disponibles de análisis del trabajo. Observe que el análisis del trabajo no se limita a la observación directa del trabajo o al muestreo directo de expertos en la materia; a menudo, las bases de datos analíticas a gran escala ofrecen información útil. Cuando los criterios de interés son medidas como el absentismo, la rotación o los accidentes, no hay una clara necesidad de análisis del trabajo para respaldar el uso de criterios.

Estándar 11.8

Los individuos que realizan o interpretan estudios empíricos de las relaciones predictor-criterio deben identificar artefactos que pueden haber influido en las conclusiones del estudio, como errores de medida, restricción de rango, deficiencias de criterios, contaminación de criterios o datos omitidos. Se debe documentar la evidencia de presencia o ausencia de talas características (y de las acciones realizadas para eliminar o controlar su influencia) y ponerse a disposición según sea necesario.

Comentario: Los errores de medida en los criterios y las restricciones en la variabilidad de los puntajes del predictor o el criterio reducen sistemáticamente las estimaciones de la relación entre las medidas del predictor y el dominio de constructo de criterio, pero están disponibles procedimientos de corrección de los efectos de estos artefactos. Cuando se aplican estos procedimientos, se deben presentar tanto los valores corregidos como los no corregidos, junto con la justificación de los procedimientos de corrección elegidos. No se deben usar pruebas de relevancia estadísticas para correlaciones no

corregidas en correlaciones corregidas. Otras características a considerar incluyen cuestiones como, por ejemplo, los datos omitidos para variables de algunos individuos, las decisiones sobre la retención o eliminación de valores extremos de los datos, los efectos de la capitalización aleatoria en la selección de predictores a partir de un conjunto mayor basándose en la solidez de las relaciones predictor-criterio, como en el caso de la recopilación de calificaciones de criterios de supervisores que conocen los puntajes de las pruebas de selección. El capítulo 3, sobre imparcialidad, describe temas adicionales que se deben tener en cuenta.

Estándar 11.9

La evidencia de las relaciones predictor-criterio en una situación local actual no se debe inferir a partir de un solo estudio previo de validez, a menos que el estudio previo de las relaciones predictor-criterio haya sido hecho bajo condiciones favorables (es decir, con un tamaño de muestra grande y un criterio pertinente) y la situación actual se corresponda estrechamente con la situación anterior.

Comentario: Una estrecha correspondencia significa que los criterios (p. ej., los requisitos del trabajo o los constructos psicológicos subyacentes) son básicamente los mismos (p. ej., tal como se han determinado por un análisis del trabajo) y que el predictor es básicamente el mismo. Los juicios sobre el grado de correspondencia se deben basar en factores que bien pueden afectar a la relación predictor-criterio. Por ejemplo, una prueba de juicio situacional que prediga el desempeño de los gerentes en un país puede o no predecir el desempeño gerencial en otro país de cultura muy diferente.

Estándar 11.10

Si las pruebas se van a utilizar para tomar decisiones de clasificación de trabajos (p. ej., si el patrón de puntajes de predictor se va a usar para hacer asignaciones diferenciales de trabajo), se necesita evidencia de que los puntajes están vinculados a diferentes niveles o probabilidades de éxito en los trabajos, grupos de trabajos o niveles de trabajos.

Comentario: Como se señaló en el capítulo 1, es posible que las pruebas sean altamente predictivas del desempeño en diferentes trabajos, pero no proporcionen evidencia del diferente grado de éxito entre los trabajos. Por ejemplo, podría predecirse que las mismas personas van a tener éxito en todos los trabajos.

Estándar 11.11

Si la evidencia basada en el contenido de la prueba es una fuente primaria de evidencia de validación que respalda el uso de una prueba para la selección en un trabajo específico, solo se debe formular una inferencia similar sobre la prueba en una nueva situación si el trabajo y la situación son básicamente los mismos que el trabajo y la situación donde se recopiló la evidencia de validación original.

Comentario: El uso apropiado de la prueba en este contexto requiere que los factores críticos de contenido del trabajo sean básicamente los mismos (p. ej., tal como se han determinado por un análisis del trabajo) y que el nivel de lectura del material de la prueba no exceda el apropiado para el nuevo trabajo. Además, el significado original de los materiales de la prueba no debe cambiar sustancialmente en la nueva situación. Por ejemplo, "*la sal es a la pimienta*" puede ser la respuesta correcta al ítem por analogía "*el blanco es al negro*" en una cultura donde las personas usan pimienta negra de forma cotidiana, pero el ítem tendría un significado diferente en una cultura donde la norma es la pimienta blanca.

Estándar 11.12

Cuando el uso de una determinada prueba para la selección de personal depende de las relaciones entre un dominio de constructo de predictor que la prueba representa y un dominio de constructo de criterio, es necesario establecer dos vinculaciones. En primer lugar, debe existir evidencia de

que los puntajes de la prueba son confiables y de que el contenido de la prueba presenta adecuadamente el dominio de constructo de predictor. En segundo lugar, debe existir evidencia de la relación entre el dominio de constructo de predictor y los principales factores del dominio de constructo de criterio.

Comentario: Debe existir una justificación conceptual clara para estas relaciones. Se deberá definir con claridad tanto el dominio de constructo de predictor como el dominio de constructo de criterio con el que se le vincula. No existe una sola ruta de preferencia para establecer estas relaciones. La evidencia que respalda las vinculaciones entre los dos dominios de constructo incluye patrones de conclusiones de la literatura de investigación y una evaluación sistemática del contenido del trabajo para identificar constructos de predictor vinculados al dominio de criterio. Se deben documentar las bases de los juicios que vinculan los dominios de constructo de predictor y criterio.

Por ejemplo, se podría usar una prueba de capacidad cognitiva para predecir el desempeño en un trabajo complejo que requiere un análisis sofisticado de muchos factores. Aquí, el dominio de constructo de predictor sería la capacidad cognitiva, y verificar el primer vínculo requeriría demostrar que la prueba es una medida adecuada del dominio de capacidad cognitiva. El segundo vínculo podría respaldarse con varias líneas de evidencia, incluyendo un conjunto de hallazgos de investigaciones que demuestren una relación sistemática entre la capacidad cognitiva y el desempeño en tareas complejas, y por los juicios de expertos en la materia relacionados con la importancia de la capacidad cognitiva para el desempeño en el dominio de desempeño.

Unidad 3. Estándares para la acreditación

Estándar 11.13

Se debe definir claramente el dominio de contenido que abarcará una prueba de acreditación y se debe justificar en términos de la importancia del contenido para el desempeño acreditado de una profesión o actividad. Se debe proporcionar una justificación o evidencia que respalde el argumento de que los conocimientos o habilidades evaluadas son necesarios para el desempeño acreditado de esa actividad y que son coherentes con el propósito para el cual se estableció el programa de acreditación.

Comentario: Habitualmente, alguna forma de análisis del trabajo o práctica proporcionan la base principal para definir el dominio de contenido. Si se usa el mismo examen en la acreditación de personas empleadas en una variedad de contextos y especialidades, es posible que sea necesario analizar un número de distintos contextos de trabajo. Aunque las técnicas de análisis del trabajo pueden ser similares a las usadas en las pruebas de empleo, el enfoque de la acreditación se limita apropiadamente a los conocimientos y habilidades necesarias para un ejercicio eficaz. Los conocimientos y habilidades incluidas en un plan de estudios básico diseñado para capacitar a personas para el trabajo o actividad pueden ser pertinentes, especialmente si el plan de estudios se ha diseñado para ser coherente con análisis empíricos del trabajo o la práctica.

En las pruebas que se usan para la obtención de licencias, no se deben incluir los conocimientos y habilidades que pueden ser importantes para el éxito, pero no se relacionan directamente con el propósito de la obtención de una licencia (p. ej., la protección del público). Por ejemplo, en contabilidad, las habilidades de marketing pueden ser importantes para el éxito y la evaluación de esas habilidades podría resultar útil para las organizaciones que seleccionan contadores. Sin embargo, es posible que la carencia de esas habilidades no sea un riesgo para el público y, por lo tanto, estas habilidades podrían excluirse correctamente del examen para la obtención de licencias. El hecho de que los profesionales con éxito presenten algunos conocimientos o habilidades es pertinente, pero no convincente. Esa información se debe cotejar con un análisis del propósito del programa de acreditación y las razones por las que se

requieren los conocimientos o habilidades en una profesión o actividad.

Estándar 11.14

Se deben proporcionar valoraciones sobre la coherencia de las decisiones de acreditación basadas en pruebas, además de otras fuentes de evidencia de confiabilidad.

Comentario: Los estándares para la coherencia de la decisión descritos en el capítulo 2 se aplican a las pruebas que se usan en la obtención de licencias y certificación. También pueden ser útiles otros tipos de valoraciones de confiabilidad y errores estándar de medida asociados, especialmente el error estándar condicional en el puntaje de corte. No obstante, es de importancia fundamental la coherencia de las decisiones en relación con la certificación.

Estándar 11.15

Las reglas y procedimientos que se usan para combinar puntajes de diferentes partes de una evaluación o puntajes de varias evaluaciones para determinar el resultado general de una prueba de acreditación se deben reportar a los examinandos, preferentemente antes de la administración de la prueba.

Comentario: En algunos casos de acreditación, es posible que se exija a los candidatos que obtengan puntajes iguales o superiores a un mínimo especificado en todas las pruebas. En otros casos, la decisión de aprobar-reprobar se puede basar exclusivamente en un puntaje agregado total. Si las pruebas se van a combinar en un puntaje agregado, se deberá proporcionar información a los candidatos sobre el peso relativo de las pruebas. No siempre es posible informar a los candidatos de la ponderación exacta antes de la administración de la prueba ya que los pesos pueden depender de propiedades empíricas de las distribuciones de los puntajes (p. ej., sus varianzas). No obstante, se deberá informar a los candidatos de la intención de ponderación (p. ej., la prueba A contribuye con un 25 % y la prueba B contribuye con un 75 % al puntaje total).

Estándar 11.16

El nivel de desempeño requerido para aprobar una prueba de acreditación depende de los conocimientos y habilidades necesarios para el desempeño acreditado en la actividad o profesión y no se debe ajustar para controlar el número o proporción de personas que superan la prueba.

Comentario: El puntaje de corte se debe determinar mediante un cuidadoso análisis y juicio del desempeño acreditado (véase el capítulo 5). Cuando existan formularios alternativos de una prueba, el puntaje de corte debe hacer referencia al mismo nivel de desempeño en todos los formularios.

12. PRUEBAS Y EVALUACIÓN EDUCATIVAS

ANTECEDENTES

El uso de pruebas educativas para informar decisiones sobre el aprendizaje, la instrucción y la política educativa tiene una larga historia. Los resultados de las pruebas se usan para establecer juicios sobre el estado, los avances o los logros de estudiantes individuales, así como de entidades como escuelas, distritos escolares, estados o países. Las pruebas usadas en contextos educativos representan una variedad de enfoques, desde formatos tradicionales de ítems abiertos y de selección múltiple hasta evaluaciones de desempeño, incluyendo porfolios puntuables. Como se señaló en el capítulo de introducción, en ocasiones se hace una distinción entre los términos *prueba* y *evaluación*, donde este último abarca fuentes de información más amplias que el puntaje mediante un solo instrumento. En este capítulo usamos ambos términos, a veces de forma intercambiable, porque los estándares examinados se aplican, en general, a ambos.

Este capítulo no trata explícitamente las cuestiones relacionadas con las pruebas desarrolladas o seleccionadas exclusivamente para informar sobre el aprendizaje y la instrucción en el nivel del aula. Con frecuencia, esas pruebas tienen consecuencias para los estudiantes e influyen en las acciones didácticas, en la ubicación de estudiantes en programas educativos, y en categorías que pueden afectar a la admisión universitaria. Los *Estándares* proporcionan criterios deseables de calidad que se pueden aplicar a estas pruebas. Sin embargo, como en las ediciones anteriores, hay consideraciones prácticas que limitan la aplicabilidad de los *Estándares* en el nivel del aula. A menudo, las prácticas formales de validez no son factibles en las pruebas de aula porque las escuelas y profesores no tienen los recursos para documentar las características de sus pruebas y estas no se publican para uso general. No obstante, se deben considerar las expectativas básicas de validez, confiabilidad/precisión e imparcialidad en el desarrollo de tales pruebas.

Los *Estándares* se aplican claramente a pruebas formales cuyos puntajes u otros resultados se usan para propósitos que van más allá del aula, como parámetros de referencia o pruebas provisionales que escuelas y distritos usan para supervisar los progresos de los estudiantes. Los *Estándares* también se aplican a evaluaciones que se adoptan para su uso en diversas aulas y presentan afirmaciones de validez de las interpretaciones de los puntajes para los usos previstos por parte de sus desarrolladores. Sin duda, esta distinción no siempre resulta clara. Distritos, escuelas y profesores usan cada vez más una gama de sistemas didácticos y de evaluación coordinados, muchos de los cuales se basan en tecnología. Estos sistemas pueden incluir, por ejemplo, bancos de ítems de prueba que los profesores individuales pueden usar en el desarrollo de pruebas para sus propios fines, ejercicios de evaluación focalizados que se adjuntan a las lecciones, o simulaciones y juegos diseñados para fines didácticos o de evaluación. Incluso si no siempre es posible separar en estos sistemas las cuestiones de medida de las cuestiones didácticas y de aprendizaje correspondientes, las evaluaciones que forman parte de esos sistemas y que sirven a propósitos que exceden la enseñanza individual de un profesor, se inscriben en el ámbito de los *Estándares*. Los desarrolladores de estos sistemas tienen la responsabilidad de adherirse a los *Estándares* para respaldar sus argumentos.

Tanto el tema introductorio como los estándares proporcionados en este capítulo se organizan en tres grandes unidades: (1) diseño y desarrollo de evaluaciones educativas; (2) uso e interpretación de evaluaciones educativas; y (3) administración, calificación y presentación de reportes de evaluaciones educativas. Aunque las unidades están relacionadas con los capítulos que examinan áreas operativas de los estándares, el análisis se basa en los principios y conceptos presentados en los capítulos principales sobre

validez, confiabilidad/precisión e imparcialidad, y los aplica a los contextos educativos. Se debe señalar que este capítulo no trata específicamente sobre el uso de los resultados de las pruebas en sistemas obligatorios de rendición de cuentas que pueden imponer recompensas o sanciones basadas en el desempeño a instituciones como, por ejemplo, escuelas o distritos escolares, o a individuos como profesores o directores. Las aplicaciones de rendición de cuentas que comportan agregados de puntajes se tratan en el capítulo 13 ("Uso de pruebas para la evaluación de programas, estudios de políticas y rendición de cuentas").

Diseño y desarrollo de evaluaciones educativas

Las pruebas educativas se diseñan y desarrollan para proporcionar puntajes que respalden las interpretaciones para los propósitos y usos previstos. Por lo tanto, el diseño y desarrollo de pruebas educativas empieza considerando el propósito de la prueba. Una vez que se establecen los propósitos de las pruebas, se pueden examinar las consideraciones relacionadas con aspectos específicos del diseño y el desarrollo.

Propósitos principales de las pruebas educativas

Aunque las pruebas educativas se usan de muchas maneras, la mayoría aborda al menos uno de tres propósitos principales: (a) formular inferencias que informen la enseñanza y el aprendizaje a nivel individual o curricular; (b) formular inferencias sobre los resultados de estudiantes individuales y grupos de estudiantes; y (c) informar las decisiones sobre los estudiantes, como la certificación de adquisición de conocimientos o habilidades específicos para la promoción, participación en programas especiales de instrucción o la graduación.

Información de la enseñanza y el aprendizaje. Las evaluaciones que informan la enseñanza y el aprendizaje empiezan con objetivos claros para el aprendizaje de los estudiantes y pueden implicar una variedad de estrategias para la evaluación de la condición y el progreso de los estudiantes. Por lo general, los objetivos son de naturaleza cognitiva, como la comprensión por parte del estudiante de los números racionales equivalentes, pero también puede abordar estados afectivos o habilidades psicomotoras. Por ejemplo, los objetivos de enseñanza y aprendizaje podrían incluir el interés creciente del estudiante por la ciencia o enseñar a los estudiantes a formar letras con lápices o plumas.

Muchas evaluaciones que informan la enseñanza y el aprendizaje se usan para fines formativos. Los profesores las usan en contextos cotidianos de aula para guiar la instrucción continua. Por ejemplo, los profesores pueden evaluar a los estudiantes antes de empezar una nueva unidad para comprobar si han adquirido los conocimientos y capacidades indispensables previos. A continuación, los profesores pueden recabar evidencias a lo largo de la unidad para ver si los estudiantes están consiguiendo los progresos anticipados e identificar cualquier laguna o concepto erróneo que necesite resolverse.

Muchas evaluaciones formales usadas para propósitos de enseñanza y aprendizaje no solo informan la instrucción en clase, sino que también proporcionan datos de evaluación individuales y agregados que otros pueden usar para respaldar mejoras en el aprendizaje. Por ejemplo, los profesores de un distrito pueden administrar periódicamente evaluaciones construidas comercial o localmente que respondan a los estándares estatales de contenido o a los planes de estudio del distrito. Estas pruebas se podrían usar para evaluar el aprendizaje de los estudiantes en una o más unidades de instrucción. Los resultados se pueden reportar de inmediato a los estudiantes, profesores y/o responsables de la escuela o el distrito. Asimismo, los resultados se pueden desglosar por estándar o subdominio de contenido para ayudar a los profesores y responsables didácticos a identificar los puntos fuertes y débiles del aprendizaje de los estudiantes, o para identificar a los estudiantes, profesores o escuelas que pueden necesitar asistencia especial. Por ejemplo, se podrían diseñar programas especiales para dar tutorías a los estudiantes en las áreas específicas que, según los resultados de las pruebas, presentan carencias.

Debido a que los resultados de las pruebas pueden influir en las decisiones sobre la instrucción posterior, es importante que los puntajes de dominios o subdominios de contenido se basen en un número suficiente de ítems o tareas que respalde de forma confiable los usos previstos.

En algunos casos, las evaluaciones administradas durante el año escolar se pueden usar para predecir el desempeño del estudiante en una evaluación sumativa a final de año. Si el desempeño pronosticado en la evaluación de final de año es bajo, podrían estar justificadas intervenciones formativas adicionales. Se pueden usar técnicas estadísticas, como la regresión lineal, para establecer las relaciones predictivas. Una variable confusa en tales predicciones puede ser el grado en que las intervenciones formativas que se basan en resultados provisionales mejoran, a lo largo del año escolar, el desempeño de estudiantes con puntajes inicialmente bajos; las relaciones predictivas se reducirán en la medida que el aprendizaje del estudiante mejora.

Evaluación de los resultados de los estudiantes. Normalmente, la evaluación de los resultados de los estudiantes presenta funciones sumativas, esto es, ayuda a evaluar el aprendizaje de los alumnos a la finalización de una secuencia formativa específica (p. ej., al final del año escolar). Los resultados de pruebas educativas de los estudiantes pueden ser considerados con varios tipos de interpretaciones de puntajes, incluyendo interpretaciones basadas en estándares, interpretaciones basadas en el crecimiento e interpretaciones normativas. Estos resultados se pueden relacionar con el estudiante individual o agregarse por grupos de estudiantes, por ejemplo, clases, subgrupos, escuelas, distritos, estados o países.

Por lo general, las interpretaciones basadas en estándares de los resultados de los estudiantes empiezan con *estándares de contenido*, que especifican qué se espera que los estudiantes conozcan y sean capaces de hacer. Normalmente, estos estándares los establecen comités de expertos en el área que se someterá a prueba. Los estándares de contenido deben ser claros y específicos, y dar a los profesores, estudiantes y padres instrucciones suficientes para guiar la enseñanza y el aprendizaje. Los *estándares de rendimiento* académico, que a veces se denominan *estándares de desempeño*, conectan los estándares de contenido con la información que describe en qué medida los estudiantes están adquiriendo los conocimientos y capacidades incluidos en los estándares de contenido académico. Los estándares de desempeño pueden incluir etiquetas de desempeño (p. ej., "básico", "competente", "avanzado"), descripciones de lo que saben y pueden hacer estudiantes de diferentes niveles de desempeño, ejemplos de trabajos de estudiantes que ilustren el rango de rendimiento en cada nivel de desempeño, y puntajes de corte que especifiquen los niveles de desempeño en una evaluación que separa niveles adyacentes de consecución. El proceso de establecer los puntajes de corte para estándares de rendimiento académico se conoce normalmente como *fijación de estándar*.

Aunque a partir de la consideración de las pruebas basadas en estándares se desprende que las evaluaciones deben alinearse estrechamente con los estándares de contenido, en general no es posible medir exhaustivamente todos los estándares de contenido usando una sola prueba sumativa. Por ejemplo, los estándares de contenido que se centran en la colaboración del estudiante, la argumentación oral o las actividades en el laboratorio de ciencias no se prestan fácilmente a la medición mediante pruebas tradicionales. Como resultado, se ha restado importancia a algunos estándares de contenido en la instrucción a expensas de estándares que se pueden medir con pruebas sumativas de final de año. Estas limitaciones se pueden solventar mediante el desarrollo de componentes de evaluación que se centren en diversos aspectos de un conjunto de estándares de contenido comunes. Por ejemplo, las evaluaciones de desempeño que están más estrechamente conectadas con las unidades formativas podrían medir determinados estándares de contenido que no se evalúan fácilmente mediante una evaluación sumativa de final de año más tradicional.

La evaluación de los resultados de los estudiantes también puede comportar interpretaciones relacionadas con los progresos de los

estudiantes o el crecimiento a lo largo del tiempo, y no con solo el desempeño en un momento específico. En pruebas basadas en estándares, una consideración importante es medir el crecimiento de los estudiantes de un año al siguiente, tanto al nivel de estudiante individual como en un nivel agregado de varios estudiantes, por ejemplo, en el nivel del profesor, el subgrupo o la escuela. Se usan varias evaluaciones educativas para supervisar el progreso o crecimiento de estudiantes individuales en uno o varios años escolares. En ocasiones, las pruebas usadas con esta finalidad están respaldadas por escalas verticales que abarcan un amplio rango de niveles educativos o de desarrollo, e incluyen (entre otros) baterías de pruebas multinivel convencionales y evaluaciones adaptables computarizadas. En la construcción de escalas verticales para pruebas educativas, es importante alinear los estándares u objetivos de aprendizaje verticalmente en los distintos niveles y diseñar pruebas en niveles adyacentes (o grados) que tengan una superposición sustancial en el contenido medido.

Sin embargo, existe una variedad de modelos estadísticos alternativos para la medición del crecimiento de los estudiantes y no todos requieren el uso de escalas verticales. Al usar y evaluar varios modelos de crecimiento, es importante entender claramente las preguntas que cada modelo puede (y no puede) responder, en qué supuestos se basa cada modelo de crecimiento y qué inferencias apropiadas se pueden derivar de los resultados de cada modelo. Los datos incompletos pueden crear problemas en algunos modelos de crecimiento. Se debe prestar atención a la posibilidad de que algunas poblaciones queden excluidas del modelo debido a datos incompletos (por ejemplo, estudiantes móviles o con baja asistencia). Otros factores que considerar en el uso de modelos de crecimiento son la confiabilidad/precisión relativa de los puntajes estimados para grupos con diferentes volúmenes de datos incompletos, y la posibilidad de que el modelo trate de forma similar a los estudiantes independientemente de su ubicación en un continuo de desempeño.

En ocasiones, los resultados de los estudiantes en pruebas educativas se evalúan a través de interpretaciones referenciadas a la norma. Una interpretación referenciada a la norma compara el desempeño de un estudiante con el desempeño de otros estudiantes. Estas interpretaciones se pueden realizar cuando se evalúa tanto el estado como el crecimiento. Las comparaciones se pueden hacer para todos los estudiantes, un subgrupo específico (p. ej., otros examinandos que se han especializado en el campo de estudio de interés para el examinando) o para subgrupos basados en muchas otras condiciones (p. ej., estudiantes con desempeño académico similar, estudiantes de escuelas similares). Se pueden desarrollar normas para una variedad de poblaciones de interés que van desde muestras de estudiantes nacionales o internacionales hasta estudiantes de un distrito escolar específico (es decir, normas locales). Las interpretaciones referenciadas a normas deben considerar las diferencias entre las poblaciones objetivo en diferentes momentos de un año escolar y en diferentes años. Cuando se administra una prueba de forma rutinaria a una población objetivo completa, como en el caso de una evaluación estatal, resulta relativamente fácil producir interpretaciones referenciadas a normas y, por lo general, solo se aplican a un único punto del año escolar. Sin embargo, las normas nacionales para una prueba de rendimiento estandarizada se suelen facilitar en varios intervalos dentro del año escolar. En ese caso, los desarrolladores deben indicar si las normas que abarcan un intervalo de tiempo específico se basaron en datos o se interpolaron de datos recopilados en otros momentos del año. Por ejemplo, las normas de invierno se basan a menudo en una interpolación de las normas empíricas recopiladas en otoño y primavera. La base para calcular las normas interpoladas se debe documentar, de manera que los usuarios puedan tener conocimiento de los supuestos subyacentes sobre el crecimiento de los estudiantes a lo largo del año escolar.

Debido al tiempo y a los gastos asociados con el desarrollo de normas nacionales, muchos desarrolladores de pruebas reportan *normas de usuario* alternativas que se componen de estadísticas descriptivas, basadas en todos aquellos que han realizado esa prueba o en un subconjunto

demográficamente representativo de los examinandos a lo largo de un periodo de tiempo. Aunque tales estadísticas (que se basan en personas que han hecho la prueba) suelen ser útiles, las normas basadas en ellas cambiarán a medida que cambie la composición del grupo de referencia. En consecuencia, las normas de usuario no se deben confundir con normas representativas de grupos muestreados más sistemáticamente.

Información de las decisiones sobre los estudiantes. A menudo, los resultados de las pruebas se usan en el proceso de toma de decisiones sobre individuos específicos, por ejemplo, sobre la graduación en escuelas secundarias, la asignación a determinados programas educativos o la promoción de un grado al siguiente. En niveles de educación superiores, los resultados de las pruebas informan las decisiones de admisión y la asignación del nivel de los estudiantes en diferentes cursos (p. ej., normales o de apoyo) o programas formativos.

La imparcialidad es una cuestión fundamental en todas las pruebas, pero debido a que las decisiones respecto de la participación, promoción o graduación educativas pueden tener un profundo efecto individual, la imparcialidad resulta esencial cuando las pruebas se usan para informar tales decisiones. En este contexto, la imparcialidad se puede mejorar a través de una atenta consideración de las condiciones que afectan a las oportunidades de los estudiantes para demostrar sus capacidades. Por ejemplo, cuando las pruebas se usan para la promoción y graduación, la imparcialidad de las interpretaciones de puntajes individuales se puede mejorar (a) proporcionando a los estudiantes varias oportunidades para demostrar sus capacidades a través de la repetición de pruebas con formularios alternativos u otros medios equivalentes de constructo; (b) proporcionando a los estudiantes un aviso adecuado de las habilidades y el contenido sometidos a prueba, junto con los materiales de preparación apropiados; (c) proporcionando a los estudiantes el plan de estudios y la instrucción para darles la oportunidad de aprender el contenido y las habilidades sometidos a prueba; (d) proporcionando a los estudiantes un acceso equitativo al contenido y las respuestas de la prueba, así como a cualquier instrucción específica para la ejecución de la prueba (p. ej., estrategias de realización de pruebas); (e) proporcionando a los estudiantes las adecuaciones apropiadas para la prueba a fin de solventar necesidades de acceso específicas; y (f) en los casos pertinentes, teniendo en cuenta varios criterios y no solo un único puntaje de prueba.

Las pruebas que informan las decisiones de admisión universitaria se usan junto con otra información sobre las capacidades de los estudiantes. Los criterios de selección pueden variar dentro de una institución por especialización académica, expedientes y promedio de calificaciones o clasificación en clase. Los puntajes de las pruebas usadas para certificar estudiantes para la graduación de enseñanza secundaria o las pruebas administradas al final de cursos específicos de secundaria se pueden usar en las decisiones de admisión universitaria. Las interpretaciones inherentes de los usos de las pruebas de enseñanza secundaria deberán tener el respaldo de varias líneas de evidencia de validación pertinente (p. ej., evidencia concurrente y predictiva). Otras medidas que usan algunas instituciones para la toma de decisiones de admisión son muestras de trabajos anteriores de los estudiantes, listas de logros académicos y de servicio, cartas de recomendación y declaraciones compuestas por los estudiantes evaluados para informarse sobre la idoneidad de los objetivos y la experiencia del estudiante y/o sus competencias en la redacción.

Las pruebas usadas para situar a los estudiantes en el nivel universitario apropiado o en cursos de apoyo juegan un papel importante en las facultades universitarias y en instituciones con programas de cuatro años. La mayoría de las instituciones usan pruebas de nivel comerciales o desarrollan sus propias pruebas para estos fines. Por lo general, los ítems de las pruebas de nivel se seleccionan para servir únicamente a este propósito de forma eficaz y en general no miden exhaustivamente el contenido previo indispensable. Por ejemplo, una prueba de nivel de álgebra solo abarcará un subconjunto del contenido de álgebra que se enseña en secundaria. Los resultados de las

pruebas de nivel se usan para exonerar a los estudiantes de asignaturas que normalmente deberían cursar. Los asesores usan otras pruebas de nivel para situar a los estudiantes en cursos de apoyo o en el curso más apropiado de una secuencia de introducción de nivel universitario. En algunos casos, las decisiones de nivel se mecanizan a través de la aplicación de puntajes de corte localmente determinados en el examen de nivel. Estos puntajes de corte se deben establecer a través de un proceso documentado que involucre a los agentes apropiados y que se valide a través de la investigación empírica.

Los resultados de las pruebas educativas también pueden informar las decisiones relacionadas con la asignación de nivel de los estudiantes en programas formativos especiales, incluyendo a estudiantes con discapacidades, estudiantes de lengua inglesa y estudiantes dotados y talentosos. Los puntajes de las pruebas nunca se deben usar como único fundamento para la inclusión de un estudiante en un programa de educación especial o para la exclusión de un estudiante de tales programas. Los puntajes de las pruebas se deben interpretar en el contexto del historial, el funcionamiento y las necesidades del estudiante. No obstante, los resultados de las pruebas pueden proporcionar una base importante para determinar si un estudiante tiene una discapacidad y cuáles son sus necesidades educativas.

Desarrollo de pruebas educativas

Al igual que en todas las pruebas, una vez que se han delineado el constructo y los propósitos de una prueba educativa, se debe tener en cuenta la población prevista de examinandos, así como los problemas prácticos como, por ejemplo, el tiempo y los recursos de evaluación disponibles que respaldan las tareas de desarrollo. En el desarrollo de pruebas educativas, la atención se centra en la medición de los conocimientos, competencias y habilidades de todos los examinandos de la población prevista, sin introducir ventajas o desventajas que se deban a características individuales (p. ej., cultura, discapacidad, género, idioma, raza/origen étnico) que sean irrelevantes para el constructo que la prueba trata de medir. Los principios de diseño universal (un método para el desarrollo de evaluaciones que intenta maximizar la accesibilidad de una prueba para todos los examinandos previstos) proporcionan una base para desarrollar evaluaciones educativas de este modo. Un factor esencial en el proceso es la documentación explícita de los pasos que se toman durante el proceso de desarrollo a fin de proporcionar evidencia de imparcialidad, confiabilidad/precisión y validez para los usos previstos de la prueba. Cuantos mayores son los riesgos asociados con la evaluación, más atención se deberá prestar a esta documentación. En el capítulo sobre imparcialidad en las pruebas (cap. 3) y en el capítulo sobre el diseño y desarrollo de pruebas (cap. 4) se detallan consideraciones relacionadas con el desarrollo de pruebas educativas.

En el desarrollo de pruebas educativas se usan una variedad de formatos, desde formatos tradicionales de ítems abiertos y de selección múltiple hasta evaluaciones de desempeño, incluyendo porfolios puntuables, simulaciones y juegos. Ejemplos de estas evaluaciones de desempeño podrían incluir la resolución de problemas usando materiales manipulables, hacer inferencias complejas después de recopilar información, o explicar oralmente o por escrito la justificación de un curso de acción gubernamental concreto bajo determinadas condiciones económicas. Se podría usar un porfolio individual como otro tipo de evaluación de desempeño. Los porfolios puntuables son colecciones sistemáticas de productos educativos normalmente recopilados, y posiblemente revisados, a lo largo del tiempo.

En contextos educativos, se suele usar la tecnología para presentar material de evaluación y para registrar y puntuar las respuestas de los examinandos. Ejemplos incluyen mejoras del texto mediante instrucciones por audio para facilitar la comprensión del estudiante, pruebas adaptables y basadas en computadora, y ejercicios de simulación donde los atributos de las evaluaciones de desempeño se refuerzan mediante tecnología. Algunos formatos de administración de pruebas también pueden tener la capacidad de capturar aspectos de los procesos de los estudiantes a medida que resuelven los ítems de la prueba. Por ejemplo,

se podría monitorizar el tiempo empleado en los ítems, las soluciones probadas y rechazadas, o la edición de secuencias de texto creadas por los examinandos. Las tecnologías también permiten proporcionar condiciones de administración de pruebas diseñadas para adecuarse a estudiantes con necesidades especiales como, por ejemplo, distintos orígenes lingüísticos, trastornos de déficit de atención o discapacidades físicas.

Las interpretaciones de los puntajes en pruebas basadas en tecnología se evalúan con los mismos estándares de validez, confiabilidad/precisión e imparcialidad que las pruebas administradas a través de medios más tradicionales. Es especialmente importante que los examinandos se familiaricen con las tecnologías de evaluación, de manera que cualquier desconocimiento de un dispositivo de entrada o interfaz de evaluación no suponga inferencias basadas en varianza irrelevante de constructo. Además, la consideración explícita de las fuentes de varianza irrelevante de constructo deberá ser parte del marco de validez a medida que nuevas tecnologías e interfaces se incorporan a los programas de evaluación. Finalmente, es importante describir los algoritmos de calificación usados en las pruebas basadas en tecnología y los modelos expertos en los que se puedan basar, y proporcionar datos técnicos que respalden su uso en la documentación del sistema de pruebas. Sin embargo, esta documentación no debe comprometer la seguridad de la evaluación de forma que la validez de las interpretaciones de los puntajes pueda quedar afectada de manera adversa.

Evaluación que sirve para distintos propósitos

Mediante la evaluación de los conocimientos y habilidades de los estudiantes relacionados con un conjunto específico de objetivos académicos, los resultados de las pruebas pueden servir para una variedad de propósitos, incluyendo la mejora de la instrucción para satisfacer mejor las necesidades de los estudiantes; la evaluación de planes de estudios y planes didácticos en el ámbito distrital; la identificación de estudiantes, escuelas o profesores que requieren ayuda; o la predicción de las probabilidades de éxito de cada estudiante en una evaluación sumativa. En tales evaluaciones, es importante validar las interpretaciones hechas a partir de los puntajes de las pruebas para cada uno de los usos previstos.

Con frecuencia, se producen tensiones asociadas con el uso de evaluaciones educativas para distintos propósitos. Por ejemplo, no es probable que una prueba desarrollada para controlar el progreso o crecimiento de estudiantes individuales en distintos años escolares también proporcione eficazmente información de diagnóstico detallada y factible sobre los puntos fuertes y débiles de los estudiantes. De forma similar, es improbable que una evaluación diseñada para ser administrada varias veces a lo largo del curso anual escolar para predecir el desempeño de un estudiante en una evaluación sumativa de final de año proporcione información útil sobre el aprendizaje del estudiante con respecto a unidades didácticas específicas. La mayoría de las pruebas educativas servirán mejor para un propósito que para otros, y cuanto más propósitos se supone atiende una prueba educativa, menos probable será que sirva eficazmente a cualquiera de esos propósitos. Por esta razón, los desarrolladores y usuarios de la prueba deben diseñar y/o seleccionar evaluaciones educativas para conseguir los propósitos que consideran más importantes, y deben considerar si se pueden lograr propósitos adicionales y supervisar la idoneidad de cualquier uso adicional identificado.

Uso e interpretación de evaluaciones educativas

Riesgos y consecuencias de la evaluación

Con frecuencia, la importancia de los resultados de los programas de evaluación para individuos, instituciones o grupos hace referencia a los *riesgos* del programa de evaluación. Cuando los riesgos para un individuo son altos y decisiones importantes dependen sensiblemente del desempeño en la prueba, la responsabilidad de proporcionar evidencia que respalde el propósito previsto de una prueba es mayor de la que cabría esperar para pruebas usadas en contextos de bajo riesgo. Aunque no es posible lograr la exactitud perfecta

en la descripción del desempeño de un individuo, es necesario hacer esfuerzos para minimizar los errores de medida o los errores de clasificación de los individuos en categorías como "aprobado", "reprobado", "admitido" o "rechazado". Además, respaldar la validez de interpretaciones para propósitos de alto riesgo (ya sean individuales o institucionales), requiere generalmente la recopilación de información colateral fidedigna que se pueda usar para ayudar a la comprensión de los factores que contribuyen a los resultados de la prueba y para corroborar la evidencia que respalda las inferencias basadas en los resultados. Por ejemplo, los resultados de las pruebas pueden verse influidos por distintos factores, tanto institucionales como individuales, como la calidad de la educación proporcionada, la exposición de los estudiantes a la educación (p. ej., a través de la asistencia regular a la escuela) y la motivación de los estudiantes para realizar bien la prueba. Recopilar este tipo de información puede contribuir a interpretaciones apropiadas de los resultados de las pruebas.

La naturaleza de alto riesgo de algunos programas de prueba puede crear dificultades especiales cuando se introducen nuevas versiones. Por ejemplo, un estado puede introducir una serie de pruebas de final de curso para secundaria que se basen en nuevos estándares de contenido y estén parcialmente vinculadas a los requisitos de graduación. El uso operativo de estas nuevas pruebas debe ir acompañado de documentación que haya sido impartida a los estudiantes sobre contenido que responda a los nuevos estándares. Debido a las limitaciones de viabilidad, esto puede requerir un periodo escalonado cuidadosamente planificado que incluya encuestas especiales o estudios de investigación cualitativos que proporcionen la documentación necesaria para la oportunidad de aprendizaje. Hasta que no esté disponible tal documentación, no se deben usar las pruebas para los propósitos de alto riesgo previstos.

Muchos tipos de pruebas educativas se ven como herramientas de política educativa. Por encima de los objetivos de la política fijada, es importante considerar los efectos potenciales imprevistos de los programas de evaluación a gran escala. Estos efectos potenciales imprevistos incluyen (a) la contracción de los planes de estudios de algunas escuelas para centrarse exclusivamente en el contenido anticipado de la prueba, (b) la restricción de la gama de métodos didácticos para corresponderse al formato de la prueba, (c) índices de abandono más altos entre los estudiantes que no aprueban la prueba, y (d) el fomento de prácticas institucionales o administrativas que pueden elevar el puntaje de las pruebas sin mejorar la calidad de la educación. Resulta esencial que quienes encargan y usan pruebas educativas conozcan esas consecuencias negativas potenciales (incluyendo las oportunidades perdidas para mejorar la enseñanza y el aprendizaje) para recabar información relacionada con estos problemas y tomar decisiones sobre el uso de las evaluaciones que tengan en cuenta esta información.

Evaluaciones para estudiantes con discapacidades y estudiantes que están aprendiendo la lengua inglesa

En la edición de 1999 de los *Estándares*, el material sobre pruebas educativas para poblaciones especiales se centraba en la evaluación diagnóstica individualizada y en la asignación educativa de los estudiantes con necesidades especiales. Desde entonces, los requisitos emanados de la legislación federal han incrementado notablemente la participación de las poblaciones especiales en los programas de evaluación educativa a gran escala. Las poblaciones especiales también se han hecho más diversas y ahora representan un porcentaje más alto de los examinandos que participan en programas educativos generales. Se diagnostica a más estudiantes con discapacidades y se incluye más de estos estudiantes en los programas de educación general y en las evaluaciones basadas en estándares de un estado. Además, el número de estudiantes que son estudiantes de lengua inglesa ha aumentado considerablemente y el número incluido en las evaluaciones educativas ha crecido en consonancia.

Como se examinó en el capítulo 3 ("Imparcialidad en las pruebas"), las evaluaciones para poblaciones especiales requieren un continuo de adaptaciones potenciales, que van desde evaluaciones alternativas especialmente desarrolladas

hasta modificaciones y adecuaciones de evaluaciones normales. La finalidad de las evaluaciones y adaptaciones alternativas es incrementar la accesibilidad de pruebas que, de otro modo, no permitirían a estudiantes con determinadas características exponer sus conocimientos y habilidades. Las evaluaciones para poblaciones especiales también podrían incluir evaluaciones desarrolladas para estudiantes de lengua inglesa y evaluaciones administradas individualmente que se usen para el diagnóstico y la ubicación.

Evaluaciones alternativas. El término *evaluaciones alternativas* que aquí se usa, en el contexto de las pruebas educativas, se refiere a las evaluaciones desarrolladas para estudiantes con discapacidades cognitivas importantes. Basadas en otros estándares de desempeño que los utilizados en las evaluaciones habituales, las evaluaciones alternativas proporcionan a los estudiantes la oportunidad de demostrar su situación y progreso en el aprendizaje. Una evaluación alternativa puede consistir en una lista de comprobación de observaciones, una evaluación multinivel con tareas de desempeño o un porfolio que incluya respuestas a tareas abiertas o de selección de respuestas. Las tareas de evaluación se desarrollan teniendo en mente las características especiales de esa población. Por ejemplo, una evaluación multinivel con tareas de desempeño podría incluir procedimientos de andamiaje donde el examinador elimine los distractores de las preguntas cuando los estudiantes responden de forma incorrecta, a fin de reducir la complejidad de la pregunta. O bien, en una evaluación de porfolio, el profesor podría incluir muestras y otra información de evaluación adaptada específicamente al estudiante. El profesor podría evaluar el mismo estándar de lengua inglesa pidiendo a un estudiante que escriba una historia y a otro que secuencie una historia usando tarjetas con gráficos, en función de la actividad que proporcione acceso a los estudiantes para que demuestren lo que saben y pueden hacer.

El desarrollo y uso de las pruebas alternativas en educación se ha visto enormemente influido por la legislación federal. Las regulaciones federales pueden exigir que las evaluaciones alternativas usada en un determinado estado tengan conexiones explícitas con los estándares de contenido medidos por la evaluación habitual estatal, aunque admita un contenido con menor profundidad, amplitud y complejidad. Estos requisitos influyen claramente en el diseño y desarrollo de evaluaciones alternativas en los programas basados en estándares de los estados.

Las evaluaciones alternativas en educación se deben llevar a cabo con los mismos requisitos técnicos que se aplican a las evaluaciones habituales a gran escala. Esto incluye documentación y datos empíricos que respalden al desarrollo de la prueba, la fijación de estándares, la validez, la confiabilidad/precisión y las características técnicas de la prueba. Cuando el número de estudiantes atendidos por las pruebas alternativas es demasiado pequeño para generar datos estadísticos estables, el desarrollador y los usuarios de la prueba deben describir dictámenes alternativos u otros procedimientos usados para documentar la evidencia de validación de las interpretaciones de la prueba.

Cuando las evaluaciones alternativas se usan para programas de pruebas a nivel estatal puede surgir una variedad de problemas de compatibilidad, por ejemplo, en la agregación de resultados de las evaluaciones alternativas y habituales o en la comparación de datos de tendencias de subgrupos cuando se han usado evaluaciones alternativas en unos años y habituales en otros.

Adecuaciones y modificaciones. Para permitir que los sistemas de evaluación incluyan a todos los estudiantes, se facilitan adecuaciones y modificaciones para aquellos estudiantes que las requieren, incluyendo a quienes participan en evaluaciones alternativas debido a discapacidades cognitivas significativas. Las adaptaciones, que incluyen tanto las adecuaciones como las modificaciones, proporcionan acceso a las evaluaciones educativas.

Las *adecuaciones* son adaptaciones del formato o administración de la prueba (por ejemplo, cambios en la forma en que se presenta la prueba, el entorno de la prueba o el modo en que los estudiantes responden) que mantienen el

mismo constructo y producen resultados que son comparables a los obtenidos por estudiantes que no usan adecuaciones. Las adecuaciones se pueden facilitar a estudiantes que estudian la lengua inglesa para solventar sus necesidades lingüísticas, así como a estudiantes con discapacidades para gestionar características individuales específicas que, de otro modo, interferirían con la accesibilidad. Por ejemplo, se puede facilitar a un estudiante con dislexia extrema un lector de pantalla que lea en voz alta escenarios y preguntas de una prueba que mida las capacidades de investigación en ciencias. El lector de pantalla se consideraría una adecuación porque la lectura no es parte del constructo definido (la investigación en ciencias) y se asume que los puntajes obtenidos por el estudiante de la prueba serían comparables a los obtenidos por estudiantes que hicieran la prueba bajo condiciones habituales.

El uso de adecuaciones se debe respaldar por la evidencia de que su aplicación no cambia el constructo que mide la evaluación. Tal evidencia puede estar disponible de estudios de aplicaciones similares, pero también podría requerir una investigación especialmente diseñada.

Las *modificaciones* son adaptaciones del formato o administración de la prueba que cambian el constructo que se mide a fin de hacerla accesible para los estudiantes designados, manteniendo tanto como sea posible el constructo original. Las modificaciones pueden dar como resultado puntajes que difieren en significado de aquellos obtenidos mediante evaluaciones habituales. Por ejemplo, se puede facilitar a un estudiante con dislexia extrema un lector de pantalla que lea en voz alta los pasajes y preguntas de una prueba de comprensión lectora que incluya la decodificación como parte del constructo. En este caso, el lector de pantalla se consideraría una modificación porque cambia el constructo que se mide y los puntajes obtenidos por el estudiante de la prueba no serían comparables a los obtenidos por estudiantes que hicieran la prueba bajo condiciones habituales. En muchos casos, las adecuaciones pueden atender las necesidades de acceso del estudiante sin el uso de modificaciones, pero en otros casos, las modificaciones son la única opción para proporcionar a algunos estudiantes acceso a la evaluación educativa. Como con las evaluaciones alternativas, el uso de modificaciones en programas de pruebas educativas presenta problemas de compatibilidad.

Las pruebas modificadas se deben diseñar y desarrollar con las mismas consideraciones de validez, confiabilidad/precisión e imparcialidad que las pruebas habituales. No es suficiente suponer que la evidencia de validación asociada con una evaluación habitual se puede generalizar para una versión modificada.

En el capítulo 3 (“Imparcialidad en las pruebas”) se examinan en detalle las modificaciones y adecuaciones para poblaciones especiales.

Evaluaciones de competencia en el idioma inglés. La presencia cada vez mayor de estudiantes de lengua inglesa en las aulas de EE. UU. se ha reflejado en una atención creciente en la medición de su competencia en el idioma inglés (ELP, por sus siglas en inglés). Como con las pruebas de contenido basadas en estándares, las pruebas ELP se basan en estándares ELP y se llevan a cabo con los mismos estándares de precisión de validez e imparcialidad de las interpretaciones de puntajes para los usos previstos, como otras pruebas a gran escala.

Las pruebas ELP pueden servir para una diversidad de propósitos. Se usan para identificar estudiantes como educandos de inglés y clasificarlos para programas y servicios especiales para estudiantes del idioma inglés, para redesignar estudiantes como competentes en inglés y para fines de diagnóstico e instrucción. Asimismo, estados, distritos y escuelas usan las pruebas ELP para monitorizar el progreso de estos estudiantes y para la rendición de cuentas de escuelas y educadores respecto del aprendizaje y progreso de los educandos de inglés hacia un nivel de competencia.

Como en cualquier prueba educativa, se puede proporcionar evidencia de validación de las medidas de ELP mediante el examen del proyecto de la prueba, la concordancia del contenido con los estándares ELP, la comparabilidad del constructo entre los estudiantes, la coherencia de la clasificación y otras afirmaciones del argumento

de validez. La justificación y la evidencia que respaldan la definición del dominio ELP y las funciones/relaciones de las modalidades del lenguaje (p. ej., lectura, escritura, competencia oral, auditiva) con respecto a la competencia en el idioma inglés, son consideraciones importantes en la articulación del argumento de validez para una prueba ELP y pueden informar la interpretación de los resultados de la prueba. Dado que una sola evaluación no tiene el mismo grado de eficacia para atender a todos los propósitos deseados, los usuarios deben considerar los usos de las pruebas ELP que tengan mayor prioridad y elegir o desarrollar los instrumentos en consonancia.

Las adecuaciones asociadas con las pruebas ELP se deben considerar cuidadosamente, ya que las adaptaciones que son apropiadas para evaluaciones de contenido habituales pueden poner en riesgo los estándares ELP que se evalúan. Además, los usuarios deben establecer directrices comunes para el uso de los resultados de ELP en la toma de decisiones sobre educandos del idioma inglés. Estas directrices deben incluir políticas y procedimientos explícitos para el uso de los resultados en la identificación y redesignación de los educandos de inglés como competentes en el idioma inglés, un proceso importante debido a la importancia legal y educativa de estas designaciones. Los organismos y escuelas de educación locales deben disponer de un fácil acceso a estas directrices.

Evaluaciones individuales. Psicólogos y otros profesionales de escuelas y contextos relacionados usan las pruebas administradas individualmente para informar decisiones sobre una variedad de servicios que se pueden administrar a los estudiantes. Los servicios se facilitan a estudiantes dotados, así como a aquellos que tienen dificultades académicas (p. ej., estudiantes que requieren clases de apoyo para la lectura). Hay otros servicios que se proporcionan a estudiantes que presentan dificultades conductuales, emocionales, físicas o de aprendizaje más severas. Los servicios pueden prestarse a estudiantes que reciben clases en aulas normales, así como a aquellos que reciben instrucción más especializada (p. ej., estudiantes de educación especial).

Si procede, cuando se usen los resultados de la prueba como ayuda para decisiones de asignación, los profesionales de evaluación cualificados deben tener en cuenta aspectos de la prueba que pueden generar varianza irrelevante de constructo en estudiantes con determinadas características pertinentes. Por ejemplo, la competencia en el idioma inglés de los estudiantes o la experiencia educativa previa podría interferir con su desempeño en una prueba de capacidad académica y, si no se tiene en cuenta, podría conducir a una clasificación errónea en educación especial. Una vez que se ha ubicado a un estudiante, se pueden administrar las pruebas para supervisar el progreso del estudiante con respecto a las metas y objetivos de aprendizaje prescritos. Los resultados de las pruebas también se pueden usar para informar evaluaciones de la eficacia didáctica y determinaciones sobre la necesidad de continuar, modificar o interrumpir los servicios especiales.

Se usan muchos tipos de pruebas en la evaluación de necesidades individualizadas y especiales. Esto incluye pruebas de capacidades cognitivas, rendimiento académico, procesos de aprendizaje, memoria visual y auditiva, habla y lenguaje, vista y audición, y comportamiento y personalidad. Por lo general, estas pruebas se usan junto con otros métodos de evaluación (por ejemplo, entrevistas, observaciones conductuales y revisión de registros) para fines de identificación y ubicación de estudiantes con discapacidades. Independientemente de las cualidades en evaluación y de los métodos de recopilación de datos empleados, los datos de evaluación que se usan en la toma de decisiones de educación especial se evalúan en términos de la evidencia que respalda las interpretaciones previstas en relación con las necesidades específicas de los estudiantes. Los datos también se deben juzgar en términos de su utilidad para la designación de programas educativos apropiados para estudiantes que tengan necesidades especiales. Para obtener más información, vea el capítulo 10 (“Pruebas y evaluación psicológicas”).

Capacidad de evaluar y desarrollo profesional

La *capacidad de evaluar* se puede definir, en sentido amplio, como el conocimiento de los

principios básicos de la práctica de evaluación correcta, incluyendo la terminología, el desarrollo y uso de metodologías y técnicas de evaluación, y la familiaridad con los estándares por los cuales se juzga la calidad de las prácticas de evaluación. Los resultados de las evaluaciones educativas se usan para la toma de decisiones en una variedad de contextos de aulas, escuelas, distritos y estados. Dado la amplitud y la complejidad de propósitos de las pruebas, es importante que los desarrolladores de pruebas y los responsables de los programas de pruebas educativas animen a que los educadores se conviertan en consumidores informados de las pruebas, y entiendan a cabalidad y usen de forma apropiada los reportes de resultados que les llegan. De forma similar, como usuarios de la prueba, es responsabilidad de los educadores buscar y conseguir la capacidad de evaluar en lo tocante a sus funciones en el sistema educativo.

Los promotores y desarrolladores de pruebas pueden promover la capacidad de evaluar de los educadores de muchas formas, incluyendo talleres, el desarrollo de materiales escritos y audiovisuales, y la colaboración con los educadores en el proceso de desarrollo de las pruebas (p. ej., desarrollo de los estándares de contenido, redacción y revisión de los ítems, y fijación de estándares). En particular, los responsables de programas de pruebas educativas deben incorporar la capacidad de evaluar en el desarrollo profesional continuo de los educadores. Además, se deben hacer intentos continuos para educar a otros agentes del proceso educativo, incluyendo a los padres, estudiantes y responsables de políticas.

Administración, calificación y presentación de reportes de evaluaciones educativas

Administración de pruebas educativas

La mayoría de las pruebas educativas conllevan procedimientos estandarizados de administración. Estos procedimientos incluyen instrucciones para los administradores y examinandos de la prueba, especificaciones para las condiciones de la evaluación y procedimientos de calificación. Debido a que, por lo general, el personal de la escuela administra las pruebas educativas, es importante que el organismo promotor proporcione la supervisión apropiada sobre el proceso y que las escuelas asignen funciones y responsabilidades locales (p. ej., la coordinación de la prueba) para capacitar a las personas que administrarán la prueba. De forma similar, los desarrolladores de la prueba tienen la obligación de respaldar el proceso de administración de la prueba y proporcionar recursos que ayuden a resolver los problemas que puedan surgir. Por ejemplo, en pruebas de alto riesgo administradas por computadora, un soporte técnico eficaz resulta crítico para la administración local y debe incluir a personas que conozcan el contexto del programa de pruebas, así como los aspectos técnicos del sistema de suministro.

Los responsables de los programas de pruebas educativas deben tener procedimientos formales para admitir adecuaciones de la prueba e implicar a personal cualificado en el proceso de toma de decisiones. Para los estudiantes con discapacidades, los cambios didácticos y de evaluación se suelen especificar en un programa de educación individualizado (IEP, por sus siglas en inglés). Para los estudiantes de lengua inglesa, las escuelas pueden usar las directrices del estado o distrito para compaginar la competencia idiomática de los estudiantes y la experiencia didáctica con las adecuaciones apropiadas del idioma. Personal cualificado debe seleccionar las adecuaciones de la prueba basándose en las necesidades individuales de los estudiantes. En programas de evaluación a gran escala, resulta especialmente importante establecer políticas y procedimientos claros para la asignación y uso de las adecuaciones. Estos pasos ayudan a mantener la comparabilidad de los puntajes de las pruebas con adecuaciones en evaluaciones académicas de distintos distritos y escuelas. Una vez seleccionadas, las adecuaciones se deben usar de forma sistemática en la instrucción y la evaluación, y los administradores de la prueba deben estar familiarizados con los procedimientos para una evaluación con adecuaciones. En el capítulo 3 (“Imparcialidad en las pruebas”) se proporciona

información relacionada con las adecuaciones de administración de pruebas.

Calificación ponderada y compuesta

La calificación de pruebas y evaluaciones educativas requiere el desarrollo de reglas para la combinación de puntajes de ítems y/o tareas para obtener un puntaje total y, en algunos casos, para la combinación de varios puntajes en un puntaje agregado. A veces, los puntajes de varias pruebas se combinan en agregados lineales usando pesos nominales, que se asignan a cada puntaje componente de acuerdo con un criterio lógico de su importancia relativa. En ocasiones, los pesos nominales pueden se equívocos debido a que la varianza del agregado también está determinada por las varianzas y covarianzas de los puntajes individuales componentes. Como resultado, es posible que el "peso efectivo" de cada componente no refleje el peso nominal. Cuando se usan puntajes agregados, se deben conocer y documentar las diferencias entre los pesos nominal y efectivo.

Para una sola prueba, a menudo los puntajes totales se basan en una simple suma de los puntajes de ítems y tareas. Sin embargo, se pueden aplicar sistemas de ponderación diferencial para reflejar el énfasis diferencial sobre contenidos o constructos específicos. Por ejemplo, en una prueba de idioma inglés, se podría asignar un mayor peso a un extenso ensayo debido a la importancia de la tarea y porque no es factible incluir en la prueba más de una tarea escrita extensa. Además, la calificación basada en modelos de la teoría de respuesta al ítem (IRT) puede dar como resultado pesos de ítems que difieren de los pesos nominales o deseados. Estas aplicaciones de la IRT deben incluir la consideración y explicación de los pesos de los ítems en la calificación. En general, las reglas de calificación que se usan en pruebas educativas se deben documentar y deben incluir una justificación basada en la validez.

Además, los desarrolladores de la prueba deben tratar con los responsables de políticas sobre los diversos métodos de combinación de los resultados de diferentes pruebas educativas usadas para tomar decisiones sobre los estudiantes, y deben documentar y comunicar claramente estos métodos, denominados también *reglas de decisión*. Por ejemplo, como parte de los requisitos de graduación, un estado puede requerir que un estudiante obtenga niveles establecidos de desempeño en varias pruebas que miden diferentes áreas de contenido usando una regla de decisión compensatoria o una no compensatoria. Bajo la regla de decisión no compensatoria, el estudiante tiene que conseguir un nivel determinado de desempeño en cada prueba; bajo la regla de decisión compensatoria, es posible que el estudiante solo tenga que conseguir un determinado puntaje agregado total basado en una combinación de puntajes de las distintas pruebas. Para una decisión de alto riesgo, como la relacionada con la graduación, las reglas usadas para combinar puntajes de distintas pruebas se deben establecer con un conocimiento claro de las implicaciones asociadas. En estas situaciones, las consecuencias importantes (como calificaciones de aprobado o índices de errores de clasificación) serán diferentes en función de las reglas para combinar los resultados de las pruebas. Los desarrolladores de pruebas deben documentar y comunicar estas implicaciones a los responsables de las políticas para propiciar decisiones plenamente informadas.

Reportes de puntajes

Los reportes de puntajes para evaluaciones educativas deben respaldar las interpretaciones y decisiones de sus audiencias previstas, que incluyen estudiantes, profesores, padres, directores, responsables de políticas y otros educadores. Se pueden desarrollar y producir diferentes reportes para diferentes audiencias, y los diseños de los reportes de puntajes pueden diferir en consonancia. Por ejemplo, los reportes preparados para estudiantes individuales y padres pueden incluir información sobre el propósito de la evaluación, definiciones de categorías de desempeño y representaciones de error de medida más accesibles para el usuario (p. ej., márgenes de error sobre gráficas de puntajes). Quienes desarrollan estos reportes se deben esforzar en proporcionar información que ayude a los estudiantes a tomar decisiones productivas sobre su propio aprendizaje. En contraste, los reportes preparados para directores y personal del

distrito pueden incluir resúmenes más detallados, pero menos información básica, ya que estas personas suelen tener un conocimiento mucho mayor de estas evaluaciones.

Como se examinó en el capítulo 3, cuando se han hecho modificaciones a una prueba para algunos examinandos y estas afectan al constructo que se mide, se puede considerar el reporte de esa modificación ya que afecta a la confiabilidad/precisión de los puntajes de la prueba o a la validez de las interpretaciones de los puntajes. Por el contrario, cuando se hacen adecuaciones que no afectan a la comparabilidad de los puntajes de la prueba, no resulta apropiado indicarlas.

En general, los reportes de puntajes de pruebas educativas se deben diseñar para proporcionar información que sea comprensible y útil para los interesados, y no lleven a interpretaciones injustificadas de los puntajes. Los desarrolladores de pruebas pueden mejorar significativamente el diseño de los reportes de puntajes llevando a cabo investigaciones de respaldo. Por ejemplo, el estudio de los reportes disponibles de otras pruebas educativas puede aportar ideas para una presentación eficaz de los resultados de las pruebas. Además, los estudios de usabilidad con consumidores de reportes de puntajes proporcionan indicaciones sobre el diseño del reporte. Se pueden usar diversas técnicas en este tipo de investigaciones, incluyendo grupos de enfoque, encuestas y análisis de protocolos verbales. Por ejemplo, las ventajas y desventajas de diseños de prototipos alternativos se pueden comparar mediante la recopilación de datos sobre las interpretaciones e inferencias formuladas por los usuarios basadas en los datos presentados en cada reporte.

La capacidad de presentación de reportes online da a los usuarios acceso flexible a los resultados de las pruebas. Por ejemplo, el usuario puede seleccionar opciones online para desglosar los resultados por contenido o subgrupo. Las opciones proporcionadas a los usuarios de la prueba para realizar consultas de resultados deben respaldar los usos e interpretaciones previstos de la prueba. Por ejemplo, los sistemas online pueden disuadir o anular la presentación de resultados, en algunos casos exigida por ley, si los tamaños de muestra de subgrupos específicos están por debajo de un número aceptable. Además, se deben tomar las medidas necesarias para permitir el acceso únicamente a los individuos apropiados. Al igual que con los reportes de puntajes, la validez de las interpretaciones a partir de sistemas de apoyo online se puede mejorar a través de estudios de usabilidad donde participen los usuarios previstos.

La tecnología facilita la estrecha concordancia de los materiales didácticos y los resultados de las pruebas educativas. Por ejemplo, los resultados reportados para un estudiante individual podrían incluir no solo sus puntos fuertes y débiles sino también vínculos directos con materiales didácticos específicos que un profesor podría usar con el estudiante en el futuro. Se debe proporcionar la justificación y documentación que respalda la eficacia de las intervenciones recomendadas, y se debe recomendar a los usuarios que consideren esta información junto con otras evidencias y criterios sobre las necesidades formativas de los estudiantes.

Cuando se reportan resultados para evaluaciones a gran escala, los promotores o usuarios de la prueba deben preparar directrices complementarias para fomentar el uso correcto y las interpretaciones válidas de los datos por los medios de comunicación y otros interesados en el proceso de evaluación. Estas comunicaciones deben abordar, probablemente, las consecuencias de la evaluación (tanto positivas como negativas), así como los usos indebidos anticipados de los resultados.

ESTÁNDARES PARA PRUEBAS Y EVALUACIÓN EDUCATIVAS

Los estándares de este capítulo se han separado en tres unidades temáticas denominadas de la siguiente manera:

1. Diseño y desarrollo de evaluaciones educativas
2. Uso e interpretación de evaluaciones educativas
3. Administración, calificación y presentación de reportes de evaluaciones educativas

Los usuarios de pruebas educativas para evaluación, políticas o rendición de cuentas deben consultar los estándares del capítulo 13 ("Uso de pruebas para la evaluación de programas, estudios de políticas y rendición de cuentas").

Unidad 1. Diseño y desarrollo de evaluaciones educativas

Estándar 12.1

Cuando escuelas, distritos, estados u otras autoridades encargan programas de pruebas educativas, se deben describir claramente los usos previstos de los resultados de las pruebas por parte de quien las ha encargado. También es responsabilidad de quienes encargan el uso de pruebas supervisar el impacto e identificar y minimizar las consecuencias negativas potenciales cuando sea factible. El desarrollador y/o usuario de la prueba deberá examinar las consecuencias resultantes de los usos de la prueba, tanto previstas como imprevistas.

Comentario: Los programas de pruebas obligatorios se suelen justificar en términos de sus potenciales beneficios para la enseñanza y el aprendizaje. Se han planteado interrogantes sobre el impacto negativo potencial de los programas de pruebas obligatorios, sobre todo cuando se traducen directamente en decisiones importantes para los individuos e instituciones. Existe la preocupación de que algunas escuelas estén restringiendo sus planes de estudio para centrarse exclusivamente en los objetivos de las pruebas, fomenten prácticas didácticas o administrativas diseñadas simplemente para subir los puntajes y no para mejorar la calidad de la educación, y pierdan un mayor número de estudiantes debido al posible abandono después de pruebas fallidas. La necesidad de supervisar el impacto de los programas de pruebas educativas se relaciona directamente con la imparcialidad en las pruebas, lo que requiere garantizar que los puntajes de una determinada prueba reflejan el mismo constructo y tienen básicamente el mismo significado para todos los individuos de la población de examinandos de destino. En consonancia con los objetivos de evaluación apropiados, se deben supervisar las consecuencias negativas potenciales y, cuando se identifiquen, se deben solventar en el máximo grado posible. En función del uso previsto, la persona responsable de examinar las consecuencias podría ser la autoridad que encomienda, el desarrollador o el usuario de la prueba.

Estándar 12.2

En contextos educativos, cuando una prueba se diseña o se usa para servir varios propósitos, se debe proporcionar evidencia de validación, confiabilidad/precisión e imparcialidad para cada uno de los usos previstos.

Comentario: En evaluaciones educativas, se ha convertido en una práctica común usar la misma prueba para varios propósitos. Por ejemplo, las pruebas provisionales/de referencia se pueden usar para una diversidad de propósitos, incluyendo el diagnóstico de los puntos fuertes y débiles del estudiante, el seguimiento del crecimiento individual del estudiante, el suministro de información para apoyar la planificación didáctica para individuos y grupos de estudiantes, y la evaluación de escuelas o distritos. Ninguna prueba servirá a todos los propósitos con la misma eficacia. Elecciones de diseño y desarrollo de la prueba que mejoran la validez para un propósito podrían

reducir la validez para otros propósitos. Diferentes propósitos pueden requerir diferentes tipos de evidencia técnica, y el desarrollador de la prueba debe proporcionar la evidencia apropiada de validez, confiabilidad/precisión e imparcialidad para cada propósito. Si el usuario de la prueba desea usarla para un propósito no respaldado por la evidencia disponible, corresponderá al usuario proporcionar la evidencia adicional necesaria. Vea el capítulo 1 ("Validez").

Estándar 12.3

Los responsables del desarrollo y uso de evaluaciones educativas deben diseñar todos los pasos pertinentes del proceso de pruebas para promover el acceso al constructo de todos los individuos y subgrupos a quienes se destina la prueba.

Comentario: En contextos educativos, es importante facilitar a todos los estudiantes (independientemente de sus características individuales) la oportunidad de demostrar su competencia en el constructo sometido a medición. Las especificaciones de la prueba deben especificar claramente todos los subgrupos pertinentes de la población objetivo, incluyendo aquellos para quienes la prueba no permitiría la demostración de conocimientos o habilidades. Los ítems y las tareas se deben diseñar para maximizar el acceso al contenido de la prueba a todos los individuos de la población de examinandos prevista. Se deben implementar herramientas y estrategias para familiarizar a todos los examinandos con la tecnología y el formato de evaluación utilizados, y se debe evitar que el método de administración y calificación introduzca alguna varianza irrelevante de constructo en el proceso de la prueba. En situaciones en que se cree que características individuales (como la competencia en el inglés, los orígenes culturales o lingüísticos, la discapacidad o la edad) pueden interferir con el acceso a los constructos que la prueba intenta medir, se deben proporcionar adaptaciones apropiadas que permitan el acceso al contenido, contexto y formatos de respuesta de los ítems de la prueba. Esto podría incluir tanto adecuaciones (cambios que, se supone, mantienen el constructo sometido a medición) como modificaciones (cambios que, se supone, crean una versión alterada del constructo accesible). El capítulo 3 ("Imparcialidad en las pruebas") incluye consideraciones adicionales relacionadas con la imparcialidad y la accesibilidad en pruebas y evaluaciones educativas.

Estándar 12.4

Cuando una prueba se usa como indicador de rendimiento en un dominio didáctico o con respecto a estándares específicos de contenido, se debe proporcionar evidencia del grado en que la prueba abarca el rango de conocimientos y revela los procesos reflejados en el dominio objetivo. Tanto el dominio probado como el objetivo se deben describir con suficiente detalle para que pueda evaluarse esta relación. El análisis debe explicitar los aspectos del dominio objetivo que la prueba representa y también los que no representa.

Comentario: Normalmente, las pruebas se desarrollan para controlar el estado o progreso de individuos o grupos con respecto a estándares de contenido locales, estatales, nacionales o profesionales. Es muy raro que una sola prueba abarque la gama completa de desempeños reflejada en los estándares de contenido. En el desarrollo de una nueva prueba o en la selección de una prueba existente, la interpretación apropiada de los puntajes como indicadores de desempeño en estos estándares requiere documentar y evaluar la relevancia de la prueba respecto de los estándares y el grado de alineación de la prueba con estos estándares. Estos estudios de alineación deben abordar varios criterios, incluyendo no solo la alineación de la prueba con las áreas de contenido incluidas en los estándares, sino también la alineación con los estándares en términos de variedad y complejidad de los conocimientos y habilidades que se espera demuestren los estudiantes. Además, realizar estudios de las estrategias y capacidades cognitivas de los examinandos, o estudios de las relaciones entre los puntajes de la prueba y otros indicadores de desempeño pertinentes al dominio objetivo más general, permite la evaluación del grado de

respaldo de las generalizaciones en ese dominio. Esta información se debe poner a disposición de todos quienes usen la prueba o interpreten los puntajes de la prueba.

Estándar 12.5

Cuando corresponda, se deben desarrollar normas locales para respaldar las interpretaciones previstas de los usuarios de la prueba.

Comentario: La comparación de los puntajes de los examinandos con grupos de normas representativas locales o más generales puede ser informativa. De este modo, si el tamaño de muestra lo admite, las normas locales suelen ser útiles en combinación con las normas publicadas, especialmente si las poblaciones locales difieren marcadamente de la población en que se basan las normas publicadas. En algunos casos, las normas locales pueden usarse de manera exclusiva.

Estándar 12.6

Se debe proporcionar la documentación del diseño, los modelos y los algoritmos de calificación para las pruebas que se administran y califican usando computadoras o recursos multimedia.

Comentario: Las pruebas por computadora y multimedia se deben llevar a cabo con los mismos requisitos de calidad técnica que otras pruebas. Por ejemplo, el uso de formatos de ítems mejorados mediante tecnología debe estar respaldado con evidencia de que los formatos son un método viable de recopilar información sobre el constructo, que no introducen varianza irrelevante de constructo y que se han tomado medidas para promover la accesibilidad para todos los estudiantes.

Unidad 2. Uso e interpretación de evaluaciones educativas

Estándar 12.7

En contextos educativos, los usuarios de la prueba deben tomar las medidas necesarias para evitar actividades de preparación de la prueba y distribución de materiales a los estudiantes que puedan afectar negativamente a la validez de las inferencias obtenidas de los puntajes.

Comentario: En la mayoría de los contextos de evaluaciones educativas, el objetivo es usar una muestra de ítems de prueba para formular inferencias respecto de un dominio más general. Cuando se producen actividades inadecuadas de preparación de la prueba (por ejemplo, la enseñanza excesiva de ítems que son equivalentes a aquellos que se usarán en la prueba), la validez de las inferencias de los puntajes de la prueba se ve afectada de forma negativa. La idoneidad de las actividades de preparación de la prueba se puede evaluar, por ejemplo, determinando el grado en que las actividades se reflejan en ítems específicos de la prueba y considerando el grado en que los puntajes de la prueba podrían mejorarse artificialmente en consecuencia, sin aumentar el verdadero nivel de rendimiento de los estudiantes.

Estándar 12.8

Cuando los resultados de la prueba contribuyen sustancialmente a decisiones sobre la promoción o graduación de estudiantes, se debe proporcionar evidencia de que los estudiantes han tenido la oportunidad de aprender el contenido y las habilidades medidas por la prueba.

Comentario: Se debe informar a estudiantes, padres y personal educativo sobre los dominios que incluirá la prueba, la naturaleza de los tipos de ítems y los criterios para determinar la destreza. Se deben hacer esfuerzos razonables para documentar la enseñanza impartida sobre el contenido y las habilidades sometidas a prueba, incluso si no resulta posible o viable determinar el contenido específico de la instrucción para cada estudiante. Además, y cuando sea apropiado, se debe proporcionar evidencia de que los estudiantes han tenido la oportunidad de familiarizarse con el modo de administración y los formatos de ítems usados en la evaluación.

Estándar 12.9

Los estudiantes que deben demostrar destreza en determinados conocimientos o habilidades para obtener una promoción o un título deben disponer de un número razonable de oportunidades para tener éxito en formularios alternativos de la prueba, o se les debe facilitar alternativas técnicamente adecuadas para demostrar su destreza en los mismos conocimientos o habilidades. En la mayoría de las circunstancias, cuando se proporciona a los estudiantes varias oportunidades para demostrar su destreza, el intervalo de tiempo entre las oportunidades debe permitirles obtener experiencias didácticas pertinentes.

Comentario: El número de oportunidades de evaluación y el tiempo entre las oportunidades variará con las circunstancias específicas del contexto. Además, la política puede dictar que algunos estudiantes dispongan de oportunidades para demostrar su rendimiento usando un método diferente. Por ejemplo, algunos estados que administran pruebas de graduación en secundaria permiten que los estudiantes que hayan participado en el plan de estudios regular, pero que no han podido demostrar el nivel de desempeño requerido en una o más pruebas, muestren, a través de un porfolio estructurado de los trabajos del curso y otros indicadores (p. ej., participación en programas de apoyo aprobados, satisfacción de otros requisitos de graduación), que tienen los conocimientos y capacidades necesarios para obtener un título de secundaria. Si se usa otro método de evaluación, deberá llevarse a cabo con los mismos estándares de calidad técnica que la evaluación principal. En particular, se debe proporcionar evidencia de que el método alternativo mide las mismas habilidades y tiene las mismas expectativas de calificación de aprobación que la evaluación principal.

Estándar 12.10

En contextos educativos, una decisión o caracterización que vaya a tener un impacto significativo en un estudiante debe tener en cuenta no solo los puntajes de una sola prueba sino otra información pertinente.

Comentario: En general, distintas medidas o fuentes de datos suelen mejorar la idoneidad de las decisiones sobre los estudiantes en contextos educativos y, por lo tanto, los promotores y usuarios de la prueba deben tenerlos en cuenta a la hora de establecer reglas y políticas de decisión. Es importante que, además de los puntajes de una sola prueba, se tome en consideración otra información pertinente (p. ej., trabajos de la escuela, observación en el aula, reportes parentales, otros puntajes de pruebas) cuando esté justificado. Estas fuentes de datos adicionales deben demostrar información pertinente para el constructo previsto. Por ejemplo, tal vez no sea recomendable o legal admitir automáticamente estudiantes en un programa de talento si su CI medido está por encima de 130, sin considerar información pertinente adicional sobre su desempeño. De forma similar, algunos estudiantes con CI medidos por debajo de 130 podrían ser admitidos basándose en otras medidas o fuentes de datos como, por ejemplo, una prueba de creatividad, un porfolio de trabajos o recomendaciones de los profesores. En estos casos, otro tipo de evidencia de desempeño talentoso sirve para compensar un puntaje de CI más bajo.

Estándar 12.11

Cuando se usan puntajes de diferencia o de crecimiento para estudiantes individuales, estos puntajes se deben definir claramente y se debe reportar evidencia de validación, confiabilidad/precisión e imparcialidad.

Comentario: Se debe reportar el error estándar de la diferencia entre puntajes de pretest y postest, la regresión de puntajes de postest en puntajes de pretest, o datos pertinentes de otros métodos apropiados para examinar el cambio.

En los casos donde se predicen puntajes de crecimiento para estudiantes individuales, se pueden usar resultados basados en diferentes versiones de pruebas realizadas a lo largo del tiempo. Por ejemplo, los puntajes de matemáticas en los grados 3,

4 y 5 se podrían usar para predecir el puntaje de matemáticas esperado en el grado 6. En tales casos, si se usan modelos estadísticos complejos para estudiantes individuales, el método para la construcción de modelos deberá ser explícito y estar justificado, y se deberá proporcionar información técnica e interpretativa de respaldo a los usuarios de los puntajes. El capítulo 13 ("Uso de pruebas para la evaluación de programas, estudios de políticas y rendición de cuentas") aborda la aplicación de modelos más complejos a grupos o sistemas en contextos de rendición de cuentas.

Estándar 12.12

Cuando se comparan los puntajes de distintas pruebas de un estudiante individual, cualquier decisión educativa basada en la comparación debe tener en cuenta el grado de superposición entre los dos constructos y la confiabilidad o error estándar del puntaje de diferencia.

Comentario: Cuando se usan puntajes de diferencia entre dos pruebas como ayuda para la toma de decisiones educativas, es importante que las dos pruebas se coloquen sobre una escala común, ya sea mediante estandarización o mediante otros medios, y, si resulta apropiado, se normalicen con respecto a la misma población en aproximadamente el mismo momento. Además, la confiabilidad y el error estándar de los puntajes de diferencia entre las dos pruebas se ven afectados por la relación entre los constructos medidos por las pruebas, así como por los errores estándar de medida de los puntajes de las dos pruebas. Por ejemplo, cuando puntajes de una capacidad no verbal se comparan con puntajes de pruebas de rendimiento, el carácter superpuesto de los dos constructos puede generar una confiabilidad más baja de los puntajes de diferencia de lo que esperarían normalmente los usuarios de la prueba. Si las pruebas de habilidad y/o rendimiento incluyen una cantidad significativa de error de medida, esto también reducirá la confianza que se pueda poner en los puntajes de diferencia. Todos estos factores afectan a la confiabilidad de los puntajes de diferencia entre las pruebas y se deben considerar cuando estos puntajes se usen como fundamento para tomar decisiones importantes sobre un estudiante. Este estándar también es pertinente en comparaciones de subpuntajes o puntajes de diferentes componentes de la misma prueba, como pueden ser los reportados por varias baterías de pruebas de aptitudes, pruebas educativas o pruebas de selección.

Estándar 12.13

Cuando se prevé que los puntajes de las pruebas se usen como parte del proceso de toma de decisiones sobre ubicación o promoción educativas, implementación de programas educativos individualizados o suministro de servicios para estudiantes de lengua inglesa, se debe proporcionar evidencia empírica que documente la relación entre los puntajes de pruebas específicas, los programas didácticos y los resultados deseados de los estudiantes. Cuando no esté disponible la evidencia empírica, debe advertirse a los usuarios que ponderen los resultados de la prueba en función de otra información pertinente sobre los estudiantes.

Comentario: El uso de los puntajes de una prueba para decisiones de asignación o promoción debe estar respaldado por evidencia sobre la relación entre los puntajes de la prueba y los beneficios previstos de los programas educativos resultantes. De este modo, se debe recopilar evidencia empírica para respaldar el uso de una prueba por una escuela universitaria para ubicar a los estudiantes que ingresan en diferentes cursos de matemáticas. De forma similar, en educación especial, cuando los puntajes de las pruebas se usen en el desarrollo de objetivos educativos y estrategias didácticas específicos, se necesitará la evidencia que demuestre que la instrucción prescrita (a) está directamente vinculada con los puntajes de la prueba, y (b) probablemente mejore el aprendizaje del estudiante. Cuando haya evidencia limitada sobre la relación entre los resultados de la prueba, los planes didácticos y los resultados de rendimiento de los estudiantes, los desarrolladores y usuarios de la prueba deberán enfatizar la naturaleza preliminar

de las recomendaciones basadas en la prueba y recomendar a los profesores y a otros responsables de tomar decisiones a ponderar la utilidad de los puntajes a la luz de otra información pertinente sobre los estudiantes.

Estándar 12.14

En contextos educativos, quienes supervisan a otros en la selección, administración e interpretación de puntajes de pruebas, deben estar familiarizados con la evidencia de confiabilidad/precisión, la validez de las interpretaciones previstas y la imparcialidad de los puntajes. Deben tener la capacidad de articular y preparar eficazmente a otros para que articulen una explicación lógica de las relaciones entre las pruebas usadas, los propósitos de las pruebas y las interpretaciones de los puntajes de las pruebas para los usos previstos.

Comentario: Las interpretaciones apropiadas de los puntajes en pruebas educativas dependen de la preparación efectiva de los individuos que llevan a cabo la administración de la prueba y de la capacitación apropiada de aquellos que hacen uso de los resultados de la prueba. Establecer programas de desarrollo profesional continuo que hagan hincapié en la mejora de la capacidad de evaluar de los profesores e interesados es un mecanismo que permite a los responsables del uso de pruebas en contextos educativos facilitar la validez de las interpretaciones de los puntajes. La fijación de requisitos educativos (p. ej., un grado avanzado, trabajos académicos pertinentes o asistencia a talleres proporcionados por el desarrollador o promotor de la prueba) es otra estrategia que se puede usar para suministrar documentación de cualificaciones y especialización.

Estándar 12.15

Los responsables de programas de pruebas educativas deben tomar las medidas necesarias para verificar que los individuos que interpretan los resultados de la prueba para la toma de decisiones en el contexto escolar estén cualificados para hacerlo o tengan la asistencia o asesoría de personas que disponen de esa cualificación.

Comentario: Cuando los programas educativos se usan como estrategia para orientar la instrucción, el personal de la escuela que se prevé deberá formular inferencias sobre la planificación didáctica, puede necesitar asistencia en la interpretación de los resultados de la prueba para esa finalidad. Esta asistencia puede consistir en desarrollo profesional continuo, guías de interpretación, capacitación, sesiones informativas y la disponibilidad de expertos para responder a las preguntas que surjan a medida que se diseminan los resultados de la prueba.

La interpretación de algunos puntajes de pruebas es suficientemente compleja para requerir que el usuario tenga capacitación y experiencia pertinentes o cuente con la ayuda o asesoría de personas con esa capacitación y experiencia. Los ejemplos incluyen las pruebas de inteligencia administradas individualmente, inventarios de interés, puntajes de crecimiento en evaluaciones estatales, pruebas proyectivas y pruebas neuropsicológicas.

Unidad 3. Administración, calificación y presentación de reportes de evaluaciones educativas

Estándar 12.16

Los responsables de los programas de pruebas educativas deben proporcionar la capacitación, documentación y supervisión apropiadas, de manera que los individuos que administren o califiquen las pruebas sean competentes en los procedimientos apropiados de administración o calificación de las pruebas y entiendan la importancia de adherirse a las instrucciones facilitadas por el desarrollador.

Comentario: Además de estar familiarizados con la documentación y los procedimientos estandarizados de administración de pruebas (incluyendo los protocolos de seguridad de pruebas), es

importante que los coordinadores y administradores de pruebas se familiaricen con los materiales y procedimientos de las adecuaciones y modificaciones en la evaluación. Por lo tanto, los desarrolladores de pruebas deben proporcionar los manuales y el material de capacitación apropiados que aborden específicamente la administración de pruebas con adecuaciones. Los coordinadores y administradores de pruebas también deben recibir información sobre las características de las poblaciones de estudiantes incluidas en el programa de evaluación.

Estándar 12.17

En contextos educativos, cuando sea posible, los reportes de las diferencias entre grupos en los puntajes de las pruebas deben ir acompañados de información contextual pertinente para facilitar la interpretación significativa de las diferencias. Cuando la información contextual apropiada no esté disponible, los usuarios deben ser cautos respecto de las interpretaciones indebidas.

Comentario: Las diferencias entre los puntajes de las pruebas entre subgrupos pertinentes (p. ej., clasificados por género, raza/origen étnico, escuela/distrito o región geográfica) pueden verse influidas, por ejemplo, por las diferencias en las características de los estudiantes, los patrones de elección de cursos, el plan de estudios, las cualificaciones de los profesores o los niveles educativos parentales. Las diferencias en el desempeño en cohortes de estudiantes a lo largo del tiempo pueden verse influidas por los cambios en la población de estudiantes bajo prueba o los cambios en las oportunidades de aprendizaje para los estudiantes. Se debe recomendar a los usuarios que consideren la información contextual apropiada y sean cautos respecto de las interpretaciones indebidas.

Estándar 12.18

En contextos educativos, los reportes de puntajes deben ir acompañados de una presentación clara de información sobre cómo interpretarlos, incluyendo el grado de error de medida asociado con cada puntaje o nivel de clasificación, y de información complementaria relacionada con los puntajes de resumen de grupo. Además, los reportes de puntajes deben incluir las fechas de administración de las pruebas y los estudios de normalización pertinentes.

Comentario: La información de puntajes se debe comunicar de forma que sea accesible para las personas que reciben el reporte. La investigación empírica relacionada con los usuarios de reportes de puntajes puede ser útil para mejorar la claridad de los reportes. Por ejemplo, el grado de incertidumbre de los puntajes se podría representar mediante errores estándar de medida presentados gráficamente; o se podría proporcionar la probabilidad de clasificación incorrecta asociada con los niveles de desempeño. De forma similar, cuando se reporten los promedios o puntajes de resumen de grupos de estudiantes, deben complementarse con información adicional sobre los tamaños de muestra y los perfiles o dispersiones de la distribución de puntajes. En los reportes de puntajes, se debe tener especial cuidado al representar la información de subpuntajes de manera que facilite una interpretación apropiada. Los reportes de puntajes deben incluir la fecha de administración, de modo que los usuarios de los puntajes puedan considerar la validez de las inferencias con el paso del tiempo. Los reportes de puntajes también deben incluir las fechas de los estudios de normalización pertinentes, de manera que los usuarios puedan tener en cuenta la antigüedad de las normas cuando formulen inferencias sobre el desempeño de los estudiantes.

Estándar 12.19

En contextos educativos, cuando los reportes de puntajes incluyan recomendaciones de intervención formativa o estén vinculados a planes recomendados o materiales didácticos, se debe proporcionar la justificación y evidencia que respalde estas recomendaciones.

Comentario: La tecnología permite asignar, cada vez en mayor medida, intervenciones formativas específicas a los estudiantes basándose

en los resultados de las evaluaciones. Se puede poner a disposición de los estudiantes contenido digital específico (p. ej., fichas de trabajo o lecciones) usando una interpretación basada en reglas de su desempeño en una prueba basadas en estándares. En estos casos, se debe proporcionar documentación que respalde la idoneidad de las asignaciones formativas. De forma similar, cuando el patrón de subpuntajes de una prueba se use para asignar estudiantes a intervenciones formativas concretas, es importante proporcionar una justificación y evidencia empírica que respalde la alegación de idoneidad de estas asignaciones. Además, se debe recomendar a los usuarios que consideren estas recomendaciones pedagógicas junto con otra información pertinente sobre los puntos fuertes y débiles de los estudiantes.

13. USO DE PRUEBAS PARA LA EVALUACIÓN DE PROGRAMAS, ESTUDIOS DE POLÍTICAS Y RENDICIÓN DE CUENTAS

ANTECEDENTES

Las pruebas se utilizan extensamente para informar decisiones como parte de políticas públicas. Un ejemplo es el uso de pruebas en el contexto del diseño y evaluación de programas o iniciativas de políticas. La *evaluación de programas* es el conjunto de procedimientos usados para emitir juicios sobre el diseño, la implementación y los resultados de un programa. Los *estudios de políticas* son más amplios que las evaluaciones de programas; contribuyen a la evaluación de los planes, principios o procedimientos dictados para conseguir objetivos públicos generales. Con frecuencia, las pruebas proporcionan datos que son analizados para calcular el efecto de una política, programa o iniciativa en resultados como, por ejemplo, el rendimiento o la motivación de los estudiantes. Una segunda categoría general del uso de pruebas en contextos de políticas son los *sistemas de* rendición de cuentas, que establecen consecuencias (p. ej., recompensas y sanciones) al desempeño de instituciones (por ejemplo, escuelas o distritos escolares) o individuos (por ejemplo, profesores o proveedores de servicios de salud mental). Las evaluaciones de programas, estudios de políticas y sistemas de rendición de cuentas no se deben ver necesariamente como categorías discretas. Con frecuencia, se adoptan combinando unas y otras, como en el caso de sistemas de rendición de cuentas que imponen requisitos o recomendaciones para usar los resultados de pruebas en la evaluación de los programas adoptados por escuelas o distritos.

El uso de pruebas para evaluaciones de programas, estudios de políticas y rendición de cuentas comparte algunas características, incluyendo la medición del desempeño de un grupo de personas y el uso de puntajes de pruebas como evidencia del éxito o las carencias de una institución o iniciativa. Este capítulo examina estos usos. El análisis de la rendición de cuentas se centra en sistemas que involucran agregados de puntajes (como las medias de toda una escuela o institución), porcentajes de estudiantes o pacientes con calificaciones por encima de determinado nivel, o el crecimiento o resultados de modelos de valor añadido agregados a nivel de aula, escuela o institución. Los sistemas o programas que se centran en la rendición de cuentas para estudiantes individuales (por ejemplo, a través de políticas de promoción o exámenes de graduación basados en pruebas) se tratan en el capítulo 12. Sin embargo, muchas de las cuestiones tratadas en ese capítulo son pertinentes para el uso de pruebas educativas para fines de evaluación de programas o rendición de cuentas en la escuela. Si los programas o sistemas de rendición de cuentas incluyen pruebas administradas a profesores, directores u otros proveedores para fines de evaluación de sus prácticas o desempeño (p. ej., programas de "pago por desempeño" para profesores que incluyan una prueba de conocimientos o una medida basada en la observación de sus prácticas), esas prácticas se deben evaluar según los estándares relacionados con las pruebas y acreditación en el centro de trabajo del capítulo 11.

Los contextos en que tiene lugar las pruebas de evaluación y de rendición de cuentas varían en cuanto a los riesgos para los examinandos y para quienes son responsables de promover resultados específicos (p. ej., profesores o proveedores de servicios de salud). Los programas de pruebas para instituciones pueden tener riesgos altos cuando el desempeño agregado de una muestra o de toda la población de examinandos se usa para

formular inferencias sobre la calidad de los servicios suministrados y, como resultado, se toman decisiones sobre estados, recompensas o sanciones institucionales. Por ejemplo, la calidad del plan de estudio y la enseñanza de la lectura se podría juzgar en parte sobre la base de los resultados de las pruebas del nivel alcanzado por grupos de estudiantes. De forma similar, a veces los puntajes agregados de pruebas psicológicas se usan para evaluar la eficacia del tratamiento que proporcionan programas u organismos de salud mental, y se pueden incluir en los sistemas de rendición de cuentas. Incluso cuando se reportan los resultados de pruebas de forma agregada y se destinan a fines de bajo riesgo, la comunicación pública de los datos se podría usar para informar juicios sobre la calidad del programa, el personal o sobre los programas educativos y podría tener influencia sobre las decisiones normativas.

Evaluación de programas e iniciativas de políticas

Como se indicó anteriormente, un programa de evaluación implica habitualmente la formulación de juicios sobre un solo programa, mientras que los estudios de políticas abordan planes, principios o procedimientos dictados para conseguir objetivos públicos generales. Los estudios de políticas pueden tratar políticas en varios niveles de gobierno, incluyendo el local, estatal, federal e internacional, y se pueden llevar a cabo en contextos organizacionales o institucionales tanto públicos como privados. No hay una distinción nítida entre estudios de políticas y evaluaciones de programas, y en muchos casos hay una superposición sustancial entre los dos tipos de investigaciones. Los resultados de las pruebas suelen ser una fuente importante de evidencia para el inicio, continuación, modificación, terminación o expansión de diversos programas y políticas.

Las pruebas se pueden usar en evaluaciones de programas o estudios de políticas para proporcionar información sobre el estado de clientes, estudiantes u otros grupos antes, durante y después de una intervención o adopción de política, así como para proporcionar información de puntajes para grupos de comparación apropiados. Si bien muchas actividades de evaluación se dirigen a documentar el desempeño de examinandos individuales, la evaluación de programas y estudios de políticas tiene como objetivo el desempeño de grupos o el impacto de los resultados de las pruebas en estos grupos. Se puede usar una variedad de pruebas para la evaluación de programas y políticas; los ejemplos incluyen las pruebas de rendimiento estandarizadas administradas por estados y distritos, pruebas psicológicas publicadas que miden resultados de interés, y medidas desarrolladas específicamente para los propósitos de la evaluación. Además, las evaluaciones de programas y políticas resumen a veces los resultados de distintos estudios y pruebas.

Es importante evaluar cualquier prueba propuesta en términos de su relevancia para los objetivos del programa o política y/o las preguntas específicas que se pueden solventar con su uso. Es relativamente raro que una prueba esté específicamente diseñada para propósitos de evaluación de programas o estudios de políticas; por lo tanto, a menudo es necesario que aquellos que realizan pruebas se basen en medidas desarrolladas para otros propósitos. Además, por razones de coste o evidencia, algunas pruebas se pueden adoptar para usarlas en una evaluación de programa o estudio de políticas, incluso si se han desarrollado para una población de examinandos un tanto diferente. Algunas pruebas se pueden seleccionar porque son bastante conocidas y se las considera especialmente confiables desde la perspectiva de los clientes o consumidores públicos, o porque ya existen datos útiles de administraciones anteriores. Se debe proporcionar evidencia de validación de los puntajes de pruebas para los usos previstos siempre que se usen pruebas para la evaluación de programas o políticas o para fines de rendición de cuentas.

Debido a realidades administrativas, como las limitaciones de costo y la carga de respuestas, se pueden adoptar ajustes metodológicos para aumentar la eficiencia de las pruebas. Una estrategia es obtener una muestra de participantes a evaluar a partir de un conjunto más grande de participantes expuestos a un programa o política. Cuando

un número suficiente de clientes se ve afectado por el programa o la política que se va a evaluar, y cuando exista el deseo de limitar el tiempo que se dedica a la evaluación, los evaluadores pueden crear diversos formularios de pruebas cortas a partir de un conjunto más grande de ítems. Mediante la construcción de un número de formularios de pruebas compuestos cada uno por un número relativamente bajo de ítems y la asignación de los formularios a diferentes submuestras de examinandos (un procedimiento conocido como *muestreo de matriz*), se puede incluir en el estudio un mayor número de ítems del que podría administrarse razonablemente a un solo examinando. Este método se suele usar cuando es deseable representar un dominio con un gran número de ítems de prueba. No obstante, en las pruebas con muestreo de matriz, normalmente los puntajes individuales no se crean ni interpretan. Debido a que los procedimientos para el muestreo de individuos o ítems de prueba pueden variar en distintas formas, el análisis e interpretación adecuados de los resultados de las pruebas dependen de una clara descripción del modo cómo se forman las muestras y de cómo se diseñan, califican y reportan las pruebas. Los reportes de resultados de las pruebas usados para la evaluación o la rendición de cuentas, deben describir la estrategia de muestreo y el grado de representatividad de la muestra respecto de la población pertinente para las inferencias previstas.

En ocasiones, las evaluaciones y estudios de política se basan en *análisis de datos secundarios*: el análisis de los datos recopilados anteriormente para otros propósitos. En algunos casos, puede ser difícil garantizar una concordancia correcta entre la prueba existente y la intervención o política bajo examen, o reconstruir en detalle las condiciones bajo las cuales se recopilaron originalmente los datos. El análisis de datos secundarios también requiere la consideración de los derechos de privacidad de los examinandos y de otros afectados por el análisis. A veces esto requiere determinar si el consentimiento informado obtenido de los participantes en la recopilación original de datos resulta adecuado para que se realice un análisis secundario sin necesidad de un consentimiento adicional. También puede ser necesario conocer el grado en que la información de identificación personal ha sido suprimida del conjunto de datos de acuerdo con la normativa vigente. Al seleccionar (o desarrollar) una prueba o al decidir el uso de datos existentes en evaluaciones o estudios de políticas, los investigadores prudentes intentan equilibrar el propósito de la prueba, la probabilidad de que sea sensible a la intervención en estudio, su credibilidad con respecto a las partes interesadas y los costos de administración. De lo contrario, los resultados de las pruebas pueden llevar a conclusiones inapropiadas sobre el progreso, el impacto y el valor general de los programas y las políticas bajo revisión

La interpretación de puntajes de las pruebas en evaluación de programas y estudios de políticas requiere habitualmente el análisis complejo de un número de variables. Por ejemplo, algunos programas son obligatorios para un grupo de población; otros se dirigen solo a determinados subgrupos. Algunos están diseñados para afectar a las actitudes, creencias o valores; en tanto que otros tienen como meta tener un impacto directo en el comportamiento, los conocimientos o las habilidades. Es importante que los participantes incluidos en cualquier estudio cumplan los criterios especificados para participar en el programa o política bajo revisión, de manera que sea posible una interpretación apropiada de los resultados de la prueba. Los resultados de la prueba reflejarán no solo los efectos de las reglas para la selección de participantes y el impacto en los participantes de programas o tratamientos, sino también las características de los participantes. Se puede obtener información contextual pertinente sobre clientes o estudiantes para reforzar las inferencias derivadas de los resultados de la prueba. Las interpretaciones válidas pueden depender de consideraciones adicionales que no tengan nada que ver con la idoneidad de la prueba o su calidad técnica, incluyendo el diseño del estudio, la viabilidad administrativa y la calidad de otros datos disponibles. Este capítulo se centra en las pruebas y no examina esas otras consideraciones de manera sustancial. Sin embargo, para el desarrollo de conclusiones defendibles, los investigadores que

llevan a cabo evaluaciones de programas y estudios de políticas deben complementar los resultados de las pruebas con datos de otras fuentes. Estos datos podrían incluir información sobre características del programa, prestaciones, costos, antecedentes de clientes, grado de participación y evidencias de efectos secundarios. Debido a que los resultados de las pruebas tienen un peso importante para la evaluación y los estudios de políticas, resulta crucial que cualquier prueba usada en estas investigaciones sea sensible con respecto a las preguntas del estudio y apropiada para los examinandos.

Sistemas de rendición de cuentas basada en pruebas

La inclusión de puntajes de pruebas en sistemas de rendición de cuentas del ámbito educativo se ha hecho común en los Estados Unidos y otros países. En la mayoría de los casos, la rendición de cuentas basada en pruebas tiene lugar en el nivel K-12, pero muchos de los problemas que surgen en el contexto K-12 son pertinentes para los esfuerzos de adoptar una rendición de cuentas basada en resultados en la educación postsecundaria. Además, los sistemas de rendición de cuentas pueden incorporar información de sistemas de datos longitudinales que relacionan el desempeño de los estudiantes en las pruebas y otros indicadores, incluyendo sistemas que capturan el desempeño de una cohorte desde el nivel pre-escolar hasta educación superior y en la fuerza laboral. En ocasiones, la rendición de cuentas basada en pruebas se produce en sectores distintos a la educación; un ejemplo es el uso de pruebas psicológicas con el fin crear medidas de eficacia para proveedores de servicios de salud mental. Estos usos de las pruebas plantean cuestiones similares a las que surgen en contextos educativos.

Los sistemas de rendición de cuentas basada en pruebas adoptan una variedad de métodos para medir el desempeño y exigir a individuos o grupos responsabilidad por ese desempeño. Estos sistemas varían en un número de dimensiones, incluyendo la unidad de la rendición de cuentas (p. ej., distrito, escuela, profesor), los riesgos aparejados con los resultados, la frecuencia de la medición y la inclusión o no de indicadores externos a la prueba en el sistema de rendición de cuentas. Una cuestión de medición importante en la rendición de cuentas se deriva de la construcción de un índice de rendición de cuentas: un número o etiqueta que refleja un conjunto de reglas para la combinación de puntajes y otra información para llegar a conclusiones e informar la toma de decisiones. Un índice de rendición de cuentas podría ser tan sencillo como un puntaje promedio de pruebas para los estudiantes de un grado específico de una escuela concreta, pero la mayoría de los sistemas dependen de índices más complejos. Estos pueden incluir un conjunto de reglas (a menudo, denominadas *reglas de decisión*) para sintetizar distintas fuentes de información como, por ejemplo, puntajes de pruebas, calificaciones de graduación, calificaciones de elección de curso y cualificaciones del profesor. Un índice de rendición de cuentas también podría crearse a partir de aplicaciones de modelos estadísticos complejos como, por ejemplo, los utilizados en métodos de modelos de valor añadido. Como se expuso en el capítulo 12, para decisiones de alto riesgo, como la clasificación de escuelas o profesores en categorías de desempeño vinculadas a recompensas o sanciones, el establecimiento de reglas usadas para crear índices de rendición de cuentas deberá estar informado por la consideración de la naturaleza de la información que se prevé proporcionará el sistema y por el conocimiento del efecto de estas reglas en las consecuencias. Las implicaciones de estas reglas se deben comunicar a los responsables de las decisiones, de manera que conozcan las consecuencias de cualquier decisión sobre las políticas que se basan en el índice de rendición de cuentas.

Los sistemas de rendición de cuentas basada en pruebas incluyen interpretaciones y supuestos que van más allá de la interpretación de los puntajes de las pruebas en las que se basan; por lo tanto, requieren evidencia adicional que respalde su validez. Por lo general, los sistemas de rendición de cuentas en educación agregan los puntajes de los estudiantes de una clase o

escuela, y se pueden usar complejos modelos matemáticos para generar un resumen estadístico, o índice, para cada profesor o escuela. Estos índices se suelen interpretar como estimaciones de la eficacia del profesor o escuela. Los usuarios de la información de los sistemas de rendición de cuentas podrían asumir que los índices de rendición de cuentas proporcionan indicadores válidos de los resultados educativos previstos (p. ej., competencia en las habilidades y conocimientos descritos en los estándares de contenido de un estado), que las diferencias entre índices se pueden atribuir a diferencias en la eficacia del profesor o escuela, y que esas diferencias son razonablemente estables a lo largo del tiempo y para distintos estudiantes e ítems. Estos supuestos deben estar respaldados por evidencias. Además, los responsables del desarrollo e implementación de sistemas de rendición de cuentas basada en pruebas sostienen, a menudo, que estos sistemas conducen a resultados específicos, como una mayor motivación del educador o mejoras de rendimiento; estas afirmaciones también se deben respaldar con evidencias. En particular, se deben adoptar medidas para investigar cualquier consecuencia positiva o negativa potencial del sistema de rendición de cuentas seleccionado.

De modo similar, la elección de reglas y datos específicos que se usan para crear un índice de rendición de cuentas debe reflejar los objetivos y valores de quienes están desarrollando el sistema de rendición de cuentas, así como las inferencias que el diseño del sistema respalda. Por ejemplo, si el objetivo principal de un sistema de rendición de cuentas es identificar a profesores que sean eficaces en la mejora del rendimiento de los estudiantes, el índice de rendición de cuentas se debe basar en evaluaciones que estén estrechamente alineadas con el contenido que se prevé cubrirá el profesor y deberá tener en cuenta factores fuera del control del profesor. Normalmente, el proceso conlleva decisiones como, por ejemplo, determinar si se miden los porcentajes sobre un puntaje de corte o sobre una media de los puntajes de escala, si se mide el estado o el crecimiento, cómo combinar la información de varios sujetos y niveles de grado, y determinar si se mide el desempeño con respecto a un objetivo fijo o se usa un método basado en clasificaciones. El desarrollo de un índice de rendición de cuentas también implica consideraciones políticas, por ejemplo, cómo equilibrar las cuestiones técnicas y la transparencia.

Problemas en la evaluación de programas y políticas y en la rendición de cuentas

En ocasiones, los resultados de las pruebas se usan como una forma de motivar a los administradores de programas u otros proveedores de servicios, así como para inferir la eficacia institucional. Se cree que el uso de estas pruebas, incluyendo el reporte público de los resultados, recomienda a que una institución mejore los servicios que ofrece a sus clientes. Por ejemplo, en algunos sistemas de rendición de cuentas basada en pruebas, resultados sistemáticamente deficientes en las pruebas de rendimiento en el nivel escolar pueden dar como resultado intervenciones que afectan al personal o a las operaciones de la escuela. La interpretación de los resultados de las pruebas es particularmente compleja cuando las pruebas se usan como mecanismo de políticas institucionales y también como una medida de eficacia. Por ejemplo, una política o programa se puede basar en el supuesto de que proporcionar objetivos claros y especificaciones generales del contenido de una prueba (p. ej., tipos de temas, constructos, dominios cognitivos y formatos de respuestas incluidos en la prueba) puede ser una estrategia razonable para comunicar nuevas expectativas a los educadores. Sin embargo, el deseo de influir en los resultados de una prueba o evaluación para demostrar un desempeño institucional aceptable podría llevar a prácticas de evaluación inapropiadas como, por ejemplo, enseñar los ítems de la prueba con antelación, modificar los procedimientos de administración, desanimar a que determinados estudiantes o clientes participen en las sesiones de evaluación, o centrar la enseñanza exclusivamente en las capacidades

que se someten a prueba. Estas respuestas ilustran que cuanto más se usa un indicador para la toma de decisiones, más probabilidades hay que se corrompa y distorsione el proceso que debe medir. Prácticas no deseables (por ejemplo, un énfasis excesivo en las habilidades sometidas a prueba) podrían sustituir a las prácticas que tienen como objetivo que los examinandos aprendan los dominios más generales medidos por la prueba. Debido a que los resultados que se derivan de tales prácticas pueden conducir a estimaciones artificialmente altas del desempeño, el investigador diligente debe estimar el impacto de los cambios en las prácticas de enseñanza que puedan deducirse de la evaluación a fin de interpretar correctamente los resultados de la prueba. Examinar las consecuencias potenciales inapropiadas de las pruebas, así como sus beneficios, dará como resultado una evaluación más precisa de los argumentos políticos sobre los tipos específicos de programas de pruebas que inducen a mejores desempeños.

Es posible que los investigadores que llevan a cabo estudios de políticas y evaluaciones de programas no den razones claras a los examinandos sobre la participación en el procedimiento de evaluación y, a menudo, oculten los resultados a los examinandos. Cuando se usa el muestreo de matriz para la evaluación de programas, es posible que no sea viable suministrar tales reportes. Si se hacen escasos esfuerzos para motivar a los examinandos para que se tomen la prueba con seriedad (p. ej., si no se explica el propósito de la prueba), los examinandos tendrían pocas razones para maximizar su esfuerzo en la prueba. De este modo, los resultados de la prueba podrían tergiversar el impacto de un programa, institución o política. Cuando existan sospechas de que una prueba no se ha realizado seriamente, se puede explorar la motivación de los examinandos mediante la recogida de información adicional donde sea factible, usando métodos de observación o entrevista. Los problemas de preparación inapropiada y desempeño desmotivado plantean preguntas sobre la validez de las interpretaciones de los resultados de las pruebas. En todo caso, es importante considerar el impacto potencial en el examinando del propio proceso de evaluación, incluyendo las prácticas de administración y presentación de reportes.

Raras veces las decisiones de políticas públicas se basan exclusivamente en los resultados de estudios empíricos, ni siquiera cuando los estudios son de alta calidad. Cuanto más expansiva e indirecta es la política, más probable es que entren en juego otras consideraciones como, por ejemplo, el impacto político y económico de abandonar, cambiar o mantener la política, o las reacciones de diversos agentes cuando las instituciones se convierten en objetivo de recompensas o sanciones. Las pruebas usadas en contextos de políticas pueden estar sujetas a un intenso y detallado escrutinio por motivos políticos. Cuando los resultados de las pruebas contradicen una posición favorecida, es posible que se hagan intentos de desacreditar el procedimiento, contenido o interpretación de la evaluación. Los usuarios de la prueba deben tener la capacidad de defender el uso de la prueba y la interpretación de los resultados, pero también deben reconocer que no pueden controlar las reacciones de los grupos interesados.

Es esencial que todas las pruebas usadas en contextos de rendición de cuentas, evaluación de programas o políticas cumplan los estándares de validez, confiabilidad e imparcialidad apropiados para las interpretaciones y usos previstos de los puntajes de las pruebas. Además, como se describe en el capítulo 6, las pruebas deben administrarse por personal con la capacitación apropiada para implementar los procedimientos de administración. También es esencial que se asista a los responsables de interpretar los resultados del estudio para profesionales, medios de comunicación y público general. Una cuidadosa comunicación sobre los objetivos, procedimientos, conclusiones y limitaciones aumenta la probabilidad de que las interpretaciones de los resultados sean precisas y útiles.

Consideraciones adicionales

Este capítulo y los estándares asociados se dirigen a los usuarios de pruebas para la evaluación de

programas, estudios de políticas y sistemas de rendición de cuentas. Los usuarios incluyen a aquellos que encargan, diseñan o implementan estas evaluaciones, estudios o sistemas, y aquellos que toman decisiones basándose en la información que proporcionan. Los usuarios incluyen, entre otros, a los psicólogos que desarrollan, evalúan o aplican políticas, así como a los educadores, administradores y responsables de políticas que trabajan en la medición del desempeño de las escuelas y la evaluación de la eficacia de programas y políticas de educación. Además de los estándares siguientes, los usuarios deben considerar otros documentos disponibles que contienen estándares pertinentes.

ESTÁNDARES PARA EL USO DE PRUEBAS PARA LA EVALUACIÓN DE PROGRAMAS, ESTUDIOS DE POLÍTICAS Y RENDICIÓN DE CUENTAS

Los estándares de este capítulo se han separado en dos unidades temáticas denominadas de la siguiente manera:

1. Diseño y desarrollo de programas de pruebas e índices para la evaluación de programas, estudios de políticas y sistemas de rendición de cuentas
2. Interpretaciones y usos de la información de pruebas usadas en evaluación de programas, estudios de políticas y sistemas de rendición de cuentas

Los usuarios de pruebas educativas para la evaluación, políticas o rendición de cuentas también deben consultar los estándares del capítulo 12 ("Pruebas y evaluación educativas") y el resto de los estándares de este volumen.

Unidad 1. Diseño y desarrollo de programas de pruebas e índices para la evaluación de programas, estudios de políticas y sistemas de rendición de cuentas

Estándar 13.1

Los usuarios de pruebas que llevan a cabo evaluaciones de programas o estudios de políticas deben describir claramente la población que ese programa o política tiene por objetivo servir y deben documentar el grado de representatividad de la muestra de examinandos respecto de esa población. Además, cuando se usen procedimientos de muestreo de matriz, se deben proporcionar las reglas para el muestreo de ítems y examinandos, y los cálculos de error deben tener en cuenta el método de muestreo. Cuando se combinan varios estudios como parte de la evaluación de un programa o estudio de política, se debe proporcionar información sobre las muestras incluidas en cada estudio individual.

Comentario: Es importante proporcionar información sobre las ponderaciones de muestreo que podría ser necesario aplicar para obtener inferencias precisas sobre el desempeño. Cuando se use el muestreo de matriz, la documentación debe abordar las limitaciones que se derivan de este método de muestreo, por ejemplo, la dificultad para crear puntajes a nivel individual. Si no se ha usado un muestreo aleatorio simple, los desarrolladores de la prueba también deben reportar estimaciones apropiadas de la varianza de error de muestreo.

Estándar 13.2

Cuando se usan puntajes de cambio o ganancia, se deben reportar los procedimientos para la construcción de puntajes, así como sus cualidades y limitaciones técnicas. Además, se deben reportar los periodos de tiempo entre las administraciones de pruebas y se debe prestar atención para evitar efectos prácticos.

Comentario: El uso de puntajes de cambio o ganancia asume que se utiliza la misma prueba, formularios equivalente de la prueba o formularios de una prueba escalada verticalmente, y que la prueba (o formulario o escala vertical) no ha sido alterada materialmente entre las administraciones Se debe reportar el error estándar de la diferencia entre puntajes de pretest y el postest, el error asociado con la regresión de puntajes de postest en puntajes de pretest, o los datos pertinentes de otros métodos para examinar el cambio, por ejemplo, aquellos basados en modelos de ecuaciones estructurales. Además de las consideraciones técnicas o metodológicas, los detalles relacionados con la administración de la prueba también pueden ser pertinentes para la interpretación de los puntajes de cambio o ganancia. Por ejemplo, es importante considerar que el error asociado con

los puntajes de cambio es más alto que el error asociado con los puntajes originales en los cuales aquellos se basan. Si se usan puntajes de cambio, se debe reportar la información sobre la confiabilidad/precisión de estos puntajes. También es importante reportar el periodo de tiempo entre las administraciones de las pruebas y, si se usa la misma prueba en varias ocasiones, se debe examinar la posibilidad de efectos prácticos (es decir, la mejora del desempeño debido a la familiaridad con los ítems de la prueba).

Estándar 13.3

Cuando se usen índices de rendición de cuentas, indicadores de eficacia en evaluaciones de programas o estudios de políticas u otros modelos estadísticos (por ejemplo, modelos de valor añadido), se debe describir y justificar el método para construir tales índices, indicadores o modelos, y se deben reportar sus cualidades técnicas.

Comentario: Un índice que se construye mediante la manipulación y combinación de puntajes de pruebas deberá estar sujeto a las mismas investigaciones de validez, confiabilidad e imparcialidad que se esperan para los puntajes de las pruebas que fundamentan el índice. Los métodos y reglas para construir estos índices deberán estar disponibles para los usuarios, junto con la documentación de sus propiedades técnicas. Se deberá evaluar las cualidades y limitaciones de diversos métodos para la combinación de puntajes, y deberá estar disponible la información que permitiría una replicación independiente de la construcción de los índices, indicadores o modelos para uso de las partes pertinentes.

Al igual que con los puntajes de pruebas habituales, deberá presentarse un argumento de validez para justificar las inferencias sobre los índices como medidas de un resultado deseado. Es importante ayudar a que los usuarios entiendan el grado en que estos modelos respaldan las inferencias causales. Por ejemplo, cuando se usan estimaciones de valor añadido como medidas de la eficacia de los profesores en la mejora del rendimiento de los estudiantes, será necesario proporcionar evidencia de la idoneidad de esta inferencia. De forma similar, si las calificaciones publicadas de proveedores de servicios de salud se basan en índices construidos a partir de puntajes de pruebas psicológicas de sus pacientes, la información pública deberá incluir información que ayude a los usuarios a entender qué inferencias sobre el desempeño del proveedor están justificadas. Los desarrolladores y usuarios de índices deben tener en cuenta las formas en que el proceso de combinación de puntajes individuales en un índice puede introducir problemas técnicos que no repercuten en los puntajes originales. Errores de vinculación, efectos suelo o techo, diferencias de variabilidad en distintas medidas y carencia de una escala de intervalos son algunos ejemplos que podrían no ser problemáticos para el propósito de interpretar puntajes individuales, pero pueden representar un problema cuando los puntajes se combinan en una medida agregada. Finalmente, cuando las evaluaciones o sistemas de rendición de cuentas se basan en medidas que combinan varias fuentes de información (por ejemplo, cuando se combinan puntajes de varios formularios de una prueba o cuando se incluye información externa a la prueba en un índice de rendición de cuentas), será necesario formular explícitamente y justificar las reglas para la combinación de la información. Es importante reconocer que cuando varias fuentes de datos se reducen a un solo puntaje o calificación agregados, los pesos y características de distribución de las fuentes afectarán a la distribución de los puntajes agregados. Se deben investigar los efectos de la ponderación y las características de distribución en el puntaje agregado.

Cuando los índices combinan puntajes de pruebas administradas bajo condiciones estándar con aquellos que incluyen modificaciones u otros cambios en las condiciones de administración, deberá existir una justificación clara de la combinación de la información en un solo índice, y se deberán examinar las implicaciones para la validez y la confiabilidad.

Unidad 2. Interpretaciones y usos de la información de pruebas usadas en evaluación de programas, estudios de políticas y sistemas de rendición de cuentas

Estándar 13.4

Se debe recopilar y poner a disposición la evidencia de validación, confiabilidad e imparcialidad del propósito del uso de una prueba en la evaluación de un programa, estudio de política o sistema de rendición de cuentas.

Comentario: Se debe proporcionar evidencia de la idoneidad del uso de una prueba en la evaluación de programas, estudios de políticas o sistemas de rendición de cuentas, incluyendo la relevancia de la prueba respecto de los objetivos del programa, política o sistema en estudio y la idoneidad de la prueba para las poblaciones interesadas. Los responsables de la publicación o presentación de reportes de resultados de pruebas deben proporcionar y explicar cualquier información complementaria que minimice posibles interpretaciones o usos indebidos de los datos. En particular, si una evaluación o sistema de rendición de cuentas se diseña para respaldar interpretaciones relacionadas con la eficacia de un programa, institución o proveedor, se deberá investigar y documentar la validez de esas interpretaciones para los usos previstos. Los reportes deben incluir precauciones contra inferencias no justificadas, por ejemplo, la exigencia de responsabilidades a proveedores de servicios de salud por cambios en los puntajes de pruebas que posiblemente no están bajo su control. Si el uso implica una clasificación de personas, instituciones o programas en distintas categorías, se debe reportar la coherencia, precisión e imparcialidad de las clasificaciones. Si la misma prueba se usa para varios propósitos (p. ej., supervisión del rendimiento de estudiantes individuales; proporcionar información para ayudar a la planificación didáctica para individuos o grupos de estudiantes; la evaluación de distritos, escuelas o profesores), se debe recopilar y proporcionar a los usuarios la evidencia relacionada con la validez de las interpretaciones para cada uno de esos usos, y será necesario considerar y mitigar los efectos negativos potenciales de algunos usos (p. ej., mejora de la enseñanza) que podrían dar como resultado uso no previstos (p. ej., responsabilidad de alto riesgo). Cuando las pruebas se usan para evaluar el desempeño del personal, se deberá examinar la idoneidad de las pruebas para diferentes grupos de personal (p. ej., profesores habituales, profesores de educación especial, directores),

Estándar 13.5

Los responsables del desarrollo y uso de pruebas para fines de evaluación y rendición de cuentas deben tomar medidas para promover interpretaciones precisas y usos apropiados para todos los grupos a los que se apliquen los resultados.

Comentario: Los responsables de la medición de resultados deben, en la medida de lo posible, diseñar el proceso de evaluación para promover el acceso y maximizar la validez de las interpretaciones (p. ej., proporcionando las adecuaciones apropiadas) para todos los subgrupos pertinentes de examinandos que participen en la evaluación del programa o política. Los usuarios de datos secundarios deben describir claramente el grado en que la población incluida en la base de datos de puntajes incluye a todos los subgrupos pertinentes. Los usuarios también deben documentar cualquier regla de exclusión que se aplique y cualquier otro cambio en el proceso de evaluación que pueda afectar a las interpretaciones de los resultados. De forma similar, los usuarios de pruebas para fines de rendición de cuentas deben hacer lo posible para incluir a todos los subgrupos pertinentes en el programa de evaluación; proporcionar documentación sobre cualquier regla de exclusión, modificaciones de las pruebas u otros cambios en la prueba o en las condiciones de administración; y facilitar la evidencia relacionada con la validez de las interpretaciones de los puntajes para los subgrupos. Cuando se reporten de forma separada resúmenes de los puntajes por subgrupo (p. ej., por grupo racial/origen étnico),

los usuarios de la prueba deben llevar a cabo análisis para evaluar la confiabilidad/precisión de los puntajes para tales grupos y la validez de las interpretaciones de los puntajes, y se debe reportar esta información cuando se publiquen los resúmenes. Los análisis de índices complejos usados para la rendición de cuentas o para la medición de la eficacia de un programa deben considerar la posibilidad de sesgo hacia subgrupos específicos o hacia programas o instituciones que prestan servicios a esos grupos. Si se detecta sesgo (p. ej., si se demuestra que los puntajes del índice están sujetos a un error sistemático relacionado con las características del examinando como la raza u origen étnico), estos índices no se deben usar a menos que se modifiquen de forma que se elimine el sesgo. El capítulo 3 incluye consideraciones adicionales relacionadas con la imparcialidad y la accesibilidad en pruebas y evaluaciones educativas.

Cuando los resultados de la prueba se usan para respaldar acciones relacionadas con la adopción o cambios de programas o políticas, es posible que los profesionales que hagan las interpretaciones que conduzcan a tales acciones necesiten asistencia en la interpretación de los resultados para este propósito. Los avances tecnológicos han permitido una creciente disponibilidad de los datos y reportes para profesores, administradores y otros agentes que pueden no haber recibido capacitación en el uso e interpretación apropiados de la prueba o en el análisis de los datos de puntajes. Quienes proporcionan los datos o herramientas tienen la responsabilidad de ofrecer soporte y asistencia a los usuarios, y los usuarios tienen la responsabilidad de buscar orientación sobre el análisis e interpretación apropiados. Los responsables de la publicación o presentación de reportes de resultados de pruebas deben proporcionar y explicar cualquier información complementaria que minimice posibles interpretaciones indebidas de los datos.

A menudo, los resultados de las pruebas para la evaluación de programas o el análisis de políticas se examinan bastante después de que se hayan realizado las pruebas. Cuando este sea el caso, el usuario deberá investigar y describir el contexto en el cual se llevaron a cabo las pruebas. Factores como las reglas de inclusión/exclusión, el propósito de la prueba, el muestreo de contenido, la alineación didáctica y la vinculación con altos riesgos pueden afectar a los resultados agregados y se deben poner en conocimiento de las audiencias para su análisis o evaluación.

Estándar 13.6

Cuando sea posible, los reportes de las diferencias entre grupos en el desempeño de las pruebas deben ir acompañados de la información contextual pertinente para facilitar la interpretación significativa de las diferencias. Cuando la información contextual apropiada no esté disponible, los usuarios deben ser cautos respecto de las interpretaciones indebidas.

Comentario: Las diferencias observadas en los puntajes promedio de pruebas entre grupos (p. ej., clasificados por género, raza/origen étnico, discapacidad, competencia en el idioma, condición socioeconómica o región geográfica) pueden verse influidas por las diferencias en factores como, por ejemplo, oportunidad de aprendizaje, experiencia en capacitación, esfuerzo, calidad del instructor, y el nivel y tipo de apoyo parental. En educación, las diferencias en el desempeño de grupos a lo largo del tiempo pueden verse influidas por los cambios en la población que se somete a la prueba (incluyendo cambios en el tamaño de muestra) o cambios en sus experiencias. Se debe recomendar a los usuarios que tenga en cuenta la información contextual apropiada cuando interpreten estas diferencias entre grupos y cuando se diseñen políticas o prácticas para solventar esas diferencias. Además, si las evaluaciones conllevan comparaciones de puntajes de pruebas a nivel internacional, se debe proporcionar evidencia de la comparabilidad de los puntajes.

Estándar 13.7

Cuando se seleccionan pruebas para usarlas en contextos de evaluación o rendición de cuentas, se deben describir claramente los usos previstos de los resultados y las consecuencias que

se espera promover, junto con las precauciones contra usos inapropiados.

Comentario: En algunos contextos, como la evaluación de un programa curricular específico, es posible que una prueba tenga un propósito limitado y no se destine a promover otros resultados específicos distintos a informar la evaluación. En otros contextos, especialmente con sistemas de rendición de cuentas basada en pruebas, el uso de pruebas se suele justificar con el argumento de que mejorará la calidad de la educación al proporcionar información útil a los responsables de tomar decisiones y crear incentivos para promover un mejor desempeño por parte de educadores y estudiantes. Este tipo de afirmaciones se deberán formular explícitamente cuando el sistema sea obligatorio o haya sido adoptado y, cuando esté disponible, se deberá proporcionar evidencia que respalde su validez. El diseño del programa deberá incorporar la recopilación y el reporte de la evidencia del argumento de validez específico. Un argumento determinado respecto de los beneficios del uso de la prueba, como la mejora del rendimiento de los estudiantes, podría estar respaldado por razonamientos lógicos o teóricos, así como por datos empíricos. Se deberá asignar el peso debido a los hallazgos de la literatura científica que pueden ser incompatibles con el argumento expuesto.

Estándar 13.8

Quienes encargan el uso de pruebas en contextos de políticas, evaluación o rendición de cuentas, y aquellos que usan pruebas en tales contextos, deben supervisar su impacto y deben identificar y minimizar las consecuencias negativas.

Comentario: El uso de pruebas en contextos de políticas, evaluación y rendición de cuentas puede, en algunos casos, acarrear consecuencias imprevistas. Especialmente cuando hay una vinculación con altos riesgos, quienes encargan las pruebas (así como quienes usan los resultados) deben adoptar medidas para identificar las consecuencias potenciales imprevistas. Las consecuencias negativas imprevistas pueden incluir la enseñanza de ítems de la prueba con antelación, la modificación de los procedimientos de administración de la prueba, y la disuasión o exclusión de algunos examinandos con respecto a la prueba. Estas prácticas pueden llevar a la obtención de puntajes artificialmente altos y que no reflejen el desempeño en el constructo subyacente o el dominio de interés. Además, estas prácticas podrían estar prohibidas por ley. Los procedimientos de evaluación deben estar diseñados para minimizar la probabilidad de tales consecuencias, y los usuarios deben recibir orientación y estímulo para abstenerse de prácticas inapropiadas en la preparación para las pruebas.

Se pueden anticipar algunas consecuencias sobre la base de investigaciones anteriores y entender cómo responden las personas a los incentivos. Por ejemplo, las investigaciones demuestran que las pruebas de rendición de cuentas en educación influyen en el plan de estudios y la instrucción al señalar lo que los estudiantes consideran importante conocer y ser capaces de hacer. Esta influencia puede ser positiva si una prueba potencia la atención en resultados útiles de aprendizaje, pero es negativa si restringe el plan de estudios en formas no previstas. Se deben estudiar y tener en cuenta los resultados de estas y otras consecuencias negativas comunes, como el posible impacto emocional en profesores y estudiantes (incluso cuando los resultados de las pruebas se usan como se tiene previsto) y el aumento de las tasas de abandono. Se debe mantener la integridad de los resultados de las pruebas esforzándose en eliminar las prácticas diseñadas para elevar los puntajes sin mejorar el desempeño en el constructo o dominio medido por la prueba. Además, la administración de una medida de auditoría (es decir, otra medida del constructo sometido a prueba) podría detectar una posible corrupción de los puntajes.

Estándar 13.9

En contextos de evaluación o rendición de cuentas, los resultados de las pruebas se deben usar junto con información de otras fuentes cuando el uso de la información adicional contribuya a la validez de la interpretación general.

Comentario: El desempeño en otros indicadores distintos a las pruebas resulta casi siempre útil y, en muchos casos, es esencial. Suele ser necesaria la descripción o el análisis de variables como los criterios de selección de clientes, las características del cliente, el contexto y los recursos, a fin de proporcionar una imagen completa del programa o política sometida a revisión y como ayuda para la interpretación de los resultados de la prueba. En contexto de rendición de cuentas, una decisión que tenga un gran impacto sobre un individuo (como un profesor o proveedor de servicios de salud) u organización (como una escuela o centro de tratamiento) deberá tener en consideración otra información pertinente además de los puntajes de las pruebas. Ejemplos de esta información adicional que se podría incorporar en las evaluaciones o sistemas de rendición de cuentas son las medidas de las prácticas de educadores o proveedores de servicios de salud (p. ej., observaciones en el aula, listas de comprobación) y medidas externas a la prueba de logros de los estudiantes (elección de cursos, asistencia al centro educativo).

En el caso de modelos de valor añadido, algunos investigadores defienden la inclusión de características demográficas del estudiante (p. ej., raza/origen étnico, condición socioeconómica) como controles, mientras que otros trabajos sugieren que la inclusión de esas variables no mejora el desempeño de las medidas y pueden promover consecuencias no deseadas como, por ejemplo, la percepción de que se establecer estándares más bajos para unos estudiantes que para otros. Las decisiones respecto a qué variables incluir en tales modelos deberá estar informada por evidencia empírica relacionada con los efectos de su inclusión o exclusión.

En contextos de políticas, un tipo adicional de información pertinente para la interpretación de resultados es el grado de motivación de los examinandos. Es importante determinar si los examinandos consideran seriamente las experiencias de evaluación, sobre todo cuando los puntajes individuales no se reportan a los examinandos o cuando los puntajes no se asocian con consecuencias para los examinandos. Se deben documentar claramente los criterios de decisión respecto a la inclusión o no de puntajes de individuos con motivaciones cuestionables.

GLOSARIO

Este glosario incluye definiciones de los términos tal como se emplean en el texto y los estándares. Muchos de estos términos presentan diversas definiciones en la literatura relacionada; asimismo, el uso técnico puede diferir del uso común.

accesibilidad: Grado en que los ítems o tareas de una prueba permiten al máximo número posible de examinandos demostrar su situación respecto del constructo de destino sin que lo impidan las características del ítem irrelevantes para la medición del constructo. Una prueba con una alta clasificación en este criterio se considera una prueba *accesible*.

aceleración: Grado de dependencia de los puntajes de los examinandos respecto de la velocidad a la que se ejecuta una tarea, así como de la exactitud de las respuestas. El término no se usa para describir pruebas de velocidad.

acreditación: Otorgar una credencial autorizada a una persona (por ejemplo, un certificado, una licencia o diploma) que denota un nivel aceptable de desempeño en un determinado dominio de conocimiento o actividad.

aculturación: Proceso relacionado con la adquisición de conocimientos y artefactos culturales, evolutivo por naturaleza y dependiente del tiempo de exposición y la oportunidad de aprendizaje.

adaptación/adaptación de prueba: 1. Cualquier cambio que se realice en el contenido, el formato (incluyendo el formato de las respuestas) o las condiciones de administración con la finalidad de aumentar la accesibilidad de la prueba para personas que, de otro modo, se enfrentarían a obstáculos irrelevantes de constructo en la prueba original. Una adaptación puede cambiar o no el significado del constructo que se mide o alterar las interpretaciones del puntaje. Una adaptación que cambia el significado del puntaje se denomina *modificación*; una adaptación que no cambia el significado del puntaje se denomina *adecuación* (consulte las definiciones en este glosario). 2. Cambio en una prueba que ha sido traducido al idioma del grupo de destino y que tiene en cuenta los matices del idioma y la cultura de ese grupo.

adecuación/adecuaciones de la prueba: Ajustes que no alteran el constructo evaluado y que se aplican a la presentación, el entorno, el contenido, el formato (incluyendo el formato de las respuestas) o las condiciones de administración de la prueba para examinandos específicos, y que se incorporan en las evaluaciones o se aplican después de diseñar la evaluación. Las pruebas o evaluaciones con este tipo de adecuaciones (así como sus puntajes) se consideran pruebas o evaluaciones *adaptadas*. Los puntajes adaptados deben ser suficientemente comparables a los puntajes no adaptados de manera que puedan agregarse.

algoritmos patentados: Procedimientos (a menudo, código informático) usados por editores comerciales o desarrolladores de pruebas que no se divulgan al público por motivos comerciales.

alineación: Grado en que el contenido o las demandas cognitivas de las preguntas de la prueba se corresponden con el contenido o las demandas cognitivas objetivo descritas en las especificaciones de la prueba.

análisis de empleo: Investigación de los puestos o clases de trabajo para obtener información sobre los deberes y tareas, las responsabilidades, las características requeridas (p. ej., conocimientos, capacidades y competencias), las condiciones laborales u otros aspectos del trabajo. Véase *análisis práctico*.

análisis de factores: Cualquiera de los métodos estadísticos para describir las interrelaciones de un conjunto de variables mediante la derivación estadística de nuevas variables, denominadas *factores*, menos numerosas que el conjunto original de variables.

análisis laboral: Investigación de una determinada ocupación o profesión para obtener

información descriptiva sobre las actividades y responsabilidades de la ocupación o profesión, y sobre los conocimientos, habilidades y capacidades necesarias para desempeñar con éxito esa ocupación o profesión. Véase *análisis de empleo*.

argumento de validez: Justificación explícita del grado en que la evidencia acumulada y la teoría respaldan una interpretación propuesta de los puntajes para el uso previsto.

batería: Conjunto de pruebas que normalmente se administran como una unidad. Por lo general, los puntajes de las pruebas se escalan de manera que se puedan comparar o usar fácilmente en combinación para la toma de decisiones.

bilingüe/multilingüe: Tener un nivel de competencia en dos o más idiomas.

calibración: 1. En vinculación de puntajes de pruebas, el proceso de relacionar los puntajes de una prueba con los puntajes de otra prueba que difieren en confiabilidad/precisión respecto de la primera prueba, de manera que tengan el mismo significado relativo para un grupo de examinandos. 2. En teoría de respuesta al ítem, el proceso de estimación de los parámetros de la función de respuesta al ítem. 3. En calificación de las tareas de respuestas construidas, los procedimientos usados durante la capacitación y la calificación para conseguir un nivel deseado de conformidad de la persona que otorga el puntaje.

capacidad de evaluación: Conocimientos sobre las evaluaciones que respaldan las interpretaciones válidas de los puntajes de prueba para los fines previstos, por ejemplo, conocimientos sobre prácticas de desarrollo de una prueba, interpretaciones de los puntajes de una prueba, riesgos para las interpretaciones válidas de los puntajes, confiabilidad y precisión de los puntajes, administración de la prueba, etc.

certificación: Proceso mediante el cual se reconoce (o certifica) que las personas han demostrado un determinado nivel de conocimientos y capacidades en un dominio específico. Véase *licencia, acreditación*.

ciencia cognitiva: Estudio interdisciplinario del aprendizaje y el procesamiento de la información.

ciencia del comportamiento: Disciplina científica, como la sociología, la antropología o la psicología, que estudia las acciones y reacciones de seres humanos y animales a través de métodos observacionales y experimentales.

coeficiente alfa: Coeficiente de confiabilidad de coherencia interna basada en el número de partes en que se divide una prueba (p. ej., ítems, subpruebas o calificadores), las interrelaciones de las partes y la varianza del puntaje total de la prueba. También denominado *Alfa de Cronbach* y, para ítems dicotómicos, *KR-20*. Véase *coeficiente de coherencia interna, coeficiente de confiabilidad*.

coeficiente de coherencia interna: Índice de confiabilidad de los puntajes de las pruebas derivado de las interrelaciones estadísticas entre las respuestas a los ítems o los puntajes de diferentes partes de una prueba. Véase *coeficiente alfa, coeficiente de confiabilidad dividido*.

coeficiente de confiabilidad dividido: Coeficiente de coherencia interna que se obtiene del uso de la mitad de los ítems de una prueba para generar un puntaje y de la otra mitad para generar un segundo puntaje independiente. Véase *coeficiente de coherencia interna, coeficiente alfa*.

coeficiente de confiabilidad test-retest: Coeficiente de confiabilidad obtenido mediante la administración de la misma prueba por segunda vez al mismo grupo después de un intervalo de tiempo y correlacionando los dos conjuntos de puntajes; por lo general, se usa como medida de estabilidad de los puntajes. Véase *estabilidad*.

coeficiente de confiabilidad: Indicador sin unidades que refleja el grado en que los puntajes están libres del error de medida aleatorio. Véase *teoría de generabilidad*.

coeficiente de generabilidad: Índice de confiabilidad/precisión basado en la teoría de generabilidad (teoría G). Un coeficiente de generabilidad es la relación de la varianza del puntaje del universo

con respecto a la varianza del puntaje observado, donde la varianza del puntaje observado es igual a la varianza del puntaje del universo más la varianza de error total. Véase *teoría de generabilidad*.

comparabilidad/comparabilidad de puntaje: En vinculación de pruebas, grado de comparabilidad del puntaje que se deriva de la aplicación de un procedimiento de vinculación. La comparabilidad del puntaje varía a lo largo de un continuum que depende del tipo de vinculación efectuado. Véase *formularios alternativos, equiparación, vinculación, moderación, proyección, escalamiento vertical*.

componentes de varianza: Acumulación de varianzas de fuentes constituyentes independientes que, en teoría, contribuyen a la varianza global de los puntajes observados. Tales varianzas, estimadas mediante métodos de análisis de varianza, suelen reflejar la situación, la ubicación, el tiempo, el formulario de la prueba, el evaluador y otros efectos relacionados. Véase *teoría de generabilidad*.

concordancia/coherencia entre los evaluadores: Nivel de coherencia con el que dos o más evaluadores califican el trabajo o desempeño de los examinandos. Véase *confiabilidad entre los evaluadores*.

concordancia: En vinculación de puntajes de pruebas para las pruebas que miden constructos similares, proceso de relacionar el puntaje de una prueba con el puntaje de otra, de manera que los puntajes tengan el mismo significado relativo para un grupo de examinandos.

confiabilidad de los evaluadores: Nivel de coherencia entre las repeticiones de un solo evaluador en la calificación de las respuestas de los examinandos. Las incoherencias en el proceso de calificación que se derivan de influencias que son internas respecto del evaluador y no de diferencias verdadera en el desempeño de los examinandos, dan como resultado una baja confiabilidad de los evaluadores.

confiabilidad entre los evaluadores: Nivel de coherencia en el orden de clasificación de las calificaciones entre los evaluadores. Véase *concordancia/coherencia entre los evaluadores*.

confiabilidad/precisión: Grado de coherencia de los puntajes de una prueba para un grupo de examinandos a través de aplicaciones repetidas de un procedimiento de medida y que, por consiguiente, permite deducir la confiabilidad y coherencia para un examinando individual; grado en que los puntajes están libres de errores aleatorios de medida para un grupo determinado. Véase *teoría de generabilidad, teoría clásica de los tests, precisión de medida*.

conjunto de ítems/banco de ítems: Colección o grupo de ítems a partir del cual se seleccionan los ítems de una prueba o escala de prueba durante el desarrollo de una evaluación, o el conjunto total de ítems a partir del cual se selecciona un subconjunto concreto para un examinando durante una prueba adaptable.

consecuencias: Resultados, previstos o imprevistos, del uso de las pruebas de manera concreta, en determinados contextos y con ciertas poblaciones.

consentimiento informado: Autorización de una persona, o del representante legal de una persona, para la ejecución de un procedimiento en o por esa persona, por ejemplo, la realización de una prueba o la cumplimentación de un cuestionario.

constructo: Concepto o característica para cuya medición se diseña una prueba.

contenido estándar: En evaluación educativa, una declaración del contenido y las competencias que se espera que adquieran los estudiantes en una asignatura; con frecuencia, en un grado concreto o al término de un nivel determinado de escolarización,

curva característica de ítem (ICC, por sus siglas en inglés): Función matemática que relaciona la probabilidad de una determinada respuesta de ítem (por lo general, una respuesta correcta) con el nivel del atributo medido por el ítem. También denominada curva de respuesta al ítem o función de respuesta al ítem.

desarrollador de la prueba: Personas u organizaciones responsables del diseño y la construcción de una prueba y de la documentación

respecto de la calidad técnica para una finalidad prevista.

desarrollo de la prueba: Proceso a través del cual se planifica, construye, evalúa y modifica una prueba, incluyendo la consideración del contenido, formato, administración, puntaje, propiedades de los ítems, escalamiento y calidad técnica para la finalidad prevista.

descriptor de nivel de desempeño: Descripciones de lo que los examinandos saben y pueden hacer en niveles específicos de desempeño.

diseño de la prueba: Proceso de desarrollo de especificaciones detalladas sobre el objeto de medición de una prueba y sobre el contenido, nivel cognitivo, formato y tipos de ítems que se van a utilizar.

diseño universal: Método de evaluación del desarrollo que intenta maximizar la accesibilidad de una prueba para todos los examinandos a los que se dirige.

distrito escolar: Organismo educativo local administrado por un consejo público de autoridades educativas o de otro tipo que supervisa las escuelas públicas de educación primaria y secundaria en una subdivisión política estatal.

documentación: Conjunto de publicaciones (p. ej., manuales de la prueba, manuales complementarios, reportes de investigación, guías de usuario) desarrolladas por el autor, desarrollador, usuario o editor de la prueba como ayuda para las interpretaciones de los puntajes para el uso previsto.

documentos de la prueba: Documentos como manuales de la prueba, manuales técnicos, guías de usuario, conjuntos de muestras e instrucciones para los administradores y evaluadores de la prueba, que proporcionan información para evaluar la idoneidad y pertinencia técnica de una prueba para la finalidad prevista.

dominio de constructo: Conjunto de atributos interrelacionados (por ej., comportamiento, actitudes, valores) que se incluyen bajo una etiqueta de constructo.

dominio de contenido: Conjunto de comportamientos, conocimientos, capacidades, competencias, actitudes u otras características que medirá una prueba, descrito en las especificaciones detalladas de la prueba y que se suele organizar en categorías clasificatorias de ítems.

dominio de criterios: Dominio de constructo de una variable que se usa como criterio. Véase *dominio de criterios*.

editor de la prueba: Entidad, persona, organización u organismo que produce o distribuye una prueba.

efecto de contexto de ítem: Influencia de la posición del ítem, otros ítems administrados, los límites de tiempo, las condiciones de administración, etc., en la dificultad de un ítem y en otras características estadísticas de un ítem.

equiparación: Proceso de relacionar los puntajes de formularios alternativos de una prueba de manera que tengan básicamente el mismo significado. Por lo general, los puntajes equiparados se reportan sobre una escala de puntaje común.

equivalencia de constructo: 1. Grado en que un constructo medido por una prueba es básicamente el mismo que el constructo medido por otra prueba. 2. Grado en el que un constructo medido por una prueba en un grupo cultural o lingüístico es comparable al constructo medido por la misma prueba en otro grupo cultural o lingüístico.

error aleatorio: Error no sistemático; un componente de los puntajes de pruebas que parece no tener relación con otras variables.

error de medida: Diferencia entre un puntaje observado y el puntaje verdadero correspondiente. Véase *error estándar de medida, error sistemático, error aleatorio, error verdadero*.

error estándar de medida condicional: Desviación estándar de los errores de medida que afecta a los puntajes de los examinandos en un nivel específico de puntaje de prueba.

error estándar de medida: Desviación estándar de los puntajes observados de un individuo en administraciones repetidas de una prueba (o de

formularios paralelos de una prueba) bajo condiciones idénticas. Debido a que, en general, tales datos no se pueden recopilar, el error estándar de medida se suele estimar a partir de datos de grupo. Véase *error de medida*.

error sistemático: Error que incrementa o reduce de manera sistemática los puntajes de todos los examinandos o de algunos subconjuntos de examinandos, pero que no está relacionado con el constructo que la prueba intenta medir. Véase *sesgo*.

escala: 1. Sistema numérico, y sus unidades, mediante el cual se reporta un valor en una determinada dimensión de medida. 2. En pruebas, conjunto de ítems o subpruebas usadas para medir una característica específica (p. ej., una prueba de habilidad verbal o una escala de extroversión-introversión).

escalamiento vertical: En vinculación de pruebas, proceso de relacionar puntajes de pruebas que miden el mismo constructo pero difieren en dificultad. En general, se usa con las pruebas de rendimiento y capacidad con contenido o dificultad que abarca una variedad de grados y niveles de edad.

escalamiento: Proceso de creación de una escala o un puntaje de escala para mejorar la interpretación de los puntajes de una prueba a través de la colocación de los puntajes de diferentes pruebas o formularios en una escala común, o mediante la generación de puntajes de escala diseñados para respaldar las interpretaciones. Véase *escala*.

especificaciones de la prueba: Documentación de la finalidad y los usos previstos de una prueba, así como del contenido, formato, duración, características psicométricas (de los ítems o de la prueba en general), modo de ejecución, administración, puntaje y reportes de puntajes de una prueba.

especificidad: En clasificación, diagnóstico y selección, proporción de casos que se evalúan como no satisfactorios o que se prevé no satisfagan los criterios y los que, en realidad, no satisfacen los criterios.

estabilidad: Grado de invariabilidad a lo largo del tiempo de los puntajes de una prueba, evaluado mediante la correlación de los puntajes de un grupo de individuos con los puntajes de la misma prueba o de una prueba equiparada realizada por el mismo grupo en un momento posterior. Véase *coeficiente de confiabilidad test-retest*.

estándares alternos o alternativos: Estándares de contenido y desempeño en evaluaciones educativas para estudiantes con discapacidades cognitivas.

estándares de desempeño: Descripciones de niveles de adquisición de conocimientos y capacidades incluidos en los estándares de contenido, tal como se articulan a través de las etiquetas de nivel de desempeño (p. ej., "básico," "competente", "avanzado"); enunciados de lo que los examinandos saben y pueden hacer en diferentes niveles de desempeño; y puntajes de corte o rangos de puntajes en la escala de una evaluación que diferencia niveles de desempeño. Véase *puntaje de corte, nivel de desempeño, descriptor de nivel de desempeño*.

estándares de rendimiento: Véase *estándares de desempeño*.

estandarización: 1. En administración de pruebas, mantener un entorno de evaluación coherente y llevar a cabo las pruebas de acuerdo con reglas y especificaciones detalladas, de manera que las condiciones de evaluación sean las mismas para todos los examinandos en una o varias ocasiones. 2. En desarrollo de pruebas, establecer una escala de presentación de reportes usando normas basadas en el desempeño en las pruebas de una muestra representativa de individuos de la población sobre la que se prevé se aplicará la prueba.

estrategias de ejecución de una prueba: Estrategias que los examinandos pueden usar cuando realizan una prueba con el fin de mejorar su desempeño (p. ej., la gestión del tiempo o la eliminación de las opciones claramente erróneas en una pregunta de respuestas múltiples) antes de responder a la pregunta.

estructura interna: En análisis de pruebas, la estructura factorial de las respuestas a los ítems o subescalas de una prueba.

estudiante de lengua inglesa (ELL, por sus siglas en inglés): Persona que aún no ha alcanzado un nivel de competencia en inglés. Un ELL puede ser una persona cuya lengua materna no es el inglés, alguien perteneciente a una minoría lingüística que empieza el aprendizaje del inglés, o una persona que ha desarrollado una competencia considerable en este idioma. Los términos relacionados incluyen estudiante de inglés (EL), competencia limitada en inglés (LEP), inglés como segunda lengua (ESL) y cultural y lingüísticamente diverso.

estudio de políticas: Estudio que contribuye a la evaluación de los planes, principios o procedimientos dictados para conseguir objetivos públicos generales.

evaluación basada en estándares: Evaluación de la situación de un individuo con respecto a un contenido descrito sistemáticamente y a estándares de desempeño.

evaluación cognitiva: Proceso de recolección sistemática de puntajes de pruebas y datos relacionados con la finalidad de formular un juicio sobre la competencia de una persona para realizar diversas actividades mentales requeridas para el procesamiento, adquisición, retención, conceptualización y organización de información sensorial, perceptual, verbal, espacial y psicomotora.

evaluación de capacidad: Uso de pruebas para evaluar el desempeño actual de una persona en dominios definidos del funcionamiento cognitivo, psicomotor o físico.

evaluación del programa: Recolección y síntesis de evidencias sobre el uso, el funcionamiento y los efectos de un programa; conjunto de procedimientos usados para formular juicios sobre el diseño, la implementación y los resultados de un programa.

evaluación formativa: Un proceso de evaluación usado por los profesores y estudiantes durante la instrucción y que proporciona información para adaptar la enseñanza y el aprendizaje en curso con el objetivo de mejorar el rendimiento de los estudiantes en los resultados educativos previstos.

evaluación neuropsicológica: Tipo especializado de evaluación psicológica de procesos normales o patológicos que afectan al sistema nervioso central y a las funciones o disfunciones psicológicas y conductuales resultantes.

evaluación psicológica: Examen del funcionamiento psicológico que comporta la recopilación, evaluación e integración de resultados de pruebas e información colateral, y la presentación de reportes sobre un individuo.

evaluación sumativa: Evaluación de los conocimientos y capacidades de un examinando que, por lo general, se realiza al finalizar un programa de aprendizaje, por ejemplo, al terminar una unidad educativa.

evaluación vocacional: Tipo especializado de evaluación psicológica diseñada para generar hipótesis e inferencias sobre los intereses, las necesidades laborales, el desarrollo profesional, la madurez vocacional y la indecisión.

evaluación: Método sistemático de obtención de información, usado para formular deducciones sobre las características de personas, objetos o programas; proceso sistemático para medir o evaluar las características o el desempeño de individuos, programas u otras entidades con la finalidad de hacer inferencias; en ocasiones se usa como sinónimo de *prueba*.

evaluaciones alternativas/pruebas alternativas: Evaluaciones o pruebas usadas para evaluar el desempeño de estudiantes en contextos educativos que no les permiten participar en evaluaciones estandarizadas de rendición de cuentas, ni siquiera con adecuaciones. Por lo general, las evaluaciones o pruebas alternativas miden el rendimiento respecto de estándares de contenido alternativos.

evaluaciones de desempeño: Evaluaciones en las cuales el examinando demuestra realmente las capacidades que la prueba pretende medir

mediante la ejecución de las tareas que requieren esas capacidades.

evaluaciones de referencia: Evaluaciones administradas en contextos educativos a horas especificadas durante una secuencia curricular, a fin de evaluar los conocimientos y habilidades de los estudiantes relacionados con un conjunto explícito de objetivos de aprendizaje a largo plazo. Véase *evaluaciones o pruebas provisionales*.

evaluaciones o pruebas provisionales: Evaluaciones administradas durante la instrucción para evaluar los conocimientos y capacidades de los estudiantes relacionados con un conjunto específico de objetivos académicos, con la finalidad de informar las decisiones del responsable de las políticas o del educador en el nivel de aula, escuela o distrito. Véase *evaluaciones de referencia*.

evidencia de convergencia: Evidencia basada en la relación entre los puntajes de la prueba y otras medidas del mismo constructo o de un constructo relacionado.

evidencia de validación predictiva: Evidencia que indica la precisión con que los datos de prueba recolectados en un determinado momento pueden predecir los puntajes de criterios que se obtienen en un momento posterior.

evidencia de validación relacionada con el contenido: Evidencia basada en el contenido de la prueba y que respalda la interpretación prevista de los puntajes de la prueba para un propósito determinado. Esta evidencia puede abordar ámbitos como la fidelidad del contenido de la prueba para actuar en el dominio en cuestión y el grado en el cual el contenido de una prueba muestra de forma representativa un dominio, por ejemplo, un plan de estudios o un trabajo.

evidencia discriminante: Evidencia que indica si dos pruebas interpretadas como medidas de diferentes constructos son suficientemente independientes (no correlacionadas) y que miden realmente dos constructos distintos.

evidencia empírica: Evidencia basada en datos, en contraposición a las evidencias basadas en la lógico o la teoría.

evidencia local: Evidencia (por lo general, relacionada con la confiabilidad/precisión o validez) recogida en una prueba específica y un conjunto específico de examinandos, en una sola institución o en una ubicación específica.

factor: Cualquier variable, real o hipotética, que sea un aspecto de un concepto o constructo.

falso negativo: Error de clasificación, diagnóstico o selección que conduce a determinar que un individuo no cumple el estándar basándose en una evaluación para la inclusión en un grupo concreto, cuando en realidad sí cumple ese estándar (o lo cumpliría en ausencia del error de medida). Véase *sensibilidad, especificidad*.

falso positivo: Error de clasificación, diagnóstico o selección que conduce a determinar que un individuo cumple el estándar basándose en una evaluación para la inclusión en un grupo concreto, cuando en realidad no cumple ese estándar (o no lo cumpliría en ausencia del error de medida). Véase *sensibilidad, especificidad*.

fijación de estándar: Proceso (a menudo basado en juicios) de fijación de puntajes de corte usando un procedimiento estructurado que intenta asignar puntajes de pruebas a niveles discretos de desempeño que, por lo general, se especifican mediante descriptores de nivel de desempeño.

formato de respuesta: Mecanismo que usa un examinando para responder a un ítem, por ejemplo, selección en una lista de opciones (pregunta de opciones múltiples) o la presentación de una respuesta escrita (respuesta de rellenado o escrita a una pregunta de respuesta abierta o construida); respuesta oral o desempeño físico.

formato/modo de prueba: Forma de presentación del contenido de la prueba al examinando: con papel y lápiz, por computadora, por Internet u oralmente con un examinador.

formulario de la prueba: Conjunto de ítems o ejercicios de una prueba que cumple los requisitos de las especificaciones de un programa de evaluación. Muchos programas de evaluación usan formularios alternativos, generados de acuerdo

con las mismas especificaciones, pero con parte o la totalidad de los ítems adaptados de manera exclusiva para cada formulario. Véase *formularios alternativos*.

formularios alternativos: Dos o más versiones de una prueba que se consideran intercambiables, en el sentido de que miden los mismos constructos de la misma forma, tienen el mismo contenido y las mismas especificaciones estadísticas, y se administran bajo las mismas condiciones, usando las mismas instrucciones. Véase *formularios equivalentes, formularios paralelos*.

formularios equiparados: Formularios alternativos de una prueba cuyas puntuaciones se han relacionado a través de un proceso estadístico, conocido como equiparación, que permite escalar los puntajes de formularios equiparados para que se puedan usar indistintamente.

formularios equivalentes: Véase *formularios alternativos, formularios paralelos*.

formularios paralelos: En teoría clásica de los tests, formularios de prueba estrictamente paralelos que, en teoría, miden el mismo constructo y tiene los mismos significados y las mismas desviaciones estándar en la población de interés. Véase *formularios alternativos*.

fraude negativo: Exagerar o falsificar las respuestas a ítems de la prueba en un intento de aparentar deficiencias.

fraude positivo: Exagerar o falsificar las respuestas a ítems de la prueba en un intento de presentarse a sí mismo de manera excesivamente positiva.

función de información de prueba: Función matemática que relaciona cada uno de los niveles de una capacidad o rasgo latente, tal como se define en la teoría de respuesta al ítem (IRT), con el recíproco de la varianza de error de medida condicional correspondiente.

funcionamiento diferencial de la prueba (DTF, por sus siglas en inglés): Desempeño individual en la prueba o nivel de dimensión que indica que individuos de diferentes grupos que tienen la misma situación respecto de la característica evaluada por una prueba, no tienen el mismo puntaje de prueba esperado.

funcionamiento diferencial de los ítems (DIF, por sus siglas en inglés): Para un ítem específico de una prueba, un indicador estadístico del grado en que diferentes grupos de examinandos que están en el mismo nivel de capacidad tienen diferentes frecuencias de respuestas correctas o, en algunos casos, diferentes índices de elección de distintas opciones de ítems.

generalización de validez: Aplicación de las evidencias de validez obtenidas en una o más situaciones a otras situaciones similares sobre la base de métodos como el meta análisis.

guía de usuario: Publicación preparada por los desarrolladores o editores de la prueba para proporcionar información sobre la finalidad, los usos apropiados, la correcta administración, los procedimientos de puntaje, los datos normativos, la interpretación de resultados y los estudios de caso de una prueba. Véase *manual de la prueba*.

imparcialidad: Validez de las interpretaciones del puntaje de una prueba para el uso previsto y para individuos de todos los subgrupos pertinentes. Una prueba equitativa minimiza la varianza irrelevante de constructo asociada con las características individuales y los contextos de la prueba que, de otro modo, comprometerían la validez de los puntajes para algunos individuos.

indicación/indicación de ítem/indicación escrita: Pregunta, estímulo o instrucción que suscita la respuesta de un examinando.

indicador: Marca adjuntada al puntaje de una prueba, a un ítem o a otra entidad para indicar una condición especial. En general, un puntaje de prueba con indicador significa que el puntaje se obtuvo a partir de una prueba modificada, con el consiguiente cambio en el constructo subyacente medido por la prueba. Es posible que los puntajes con indicador no sean comparables a los puntajes sin indicador.

índice de rendición de cuentas: Número o etiqueta que refleja un conjunto de reglas para

la combinación de puntajes y otros datos con la finalidad de extraer conclusiones e informar el proceso de toma de decisiones en un sistema de rendición de cuentas.

reporte interpretativo preparado por computadora: Interpretación programada de los resultados de un examinando basada en los datos empíricos y/o en el juicio de un experto, y que utiliza varios formatos como narraciones, tablas y gráficos. En ocasiones se le denomina *puntaje automatizado o informe narrativo*.

infrarrepresentación de constructo: Grado en el cual una prueba no logra capturar aspectos importantes del dominio de constructo que se pretende medir, lo que se traduce en puntajes de prueba que no representan totalmente ese constructo.

interpretación de puntaje conforme a criterios: Significado de un puntaje de prueba para un individuo (o de un puntaje promedio para un grupo definido) que indica el nivel de desempeño de los individuos o grupos en relación con un dominio de criterios definido. Ejemplos de interpretaciones conforme a criterios incluyen comparaciones para puntajes de corte, interpretaciones basadas en tablas de expectativas e interpretaciones de puntaje conforme a dominios. Compárese con *interpretación de puntaje conforme a normas*.

interpretación de puntaje conforme a normas: Interpretación de puntaje basada en una comparación del desempeño de un examinando con la distribución del desempeño en una población de referencia definida. Compárese con *interpretación de puntaje conforme a criterios*.

intérprete: Alguien que facilita la comunicación intercultural mediante la conversión de conceptos de un idioma a otro (incluyendo el lenguaje de signos).

intervalo de confianza: Intervalo en el cual estará incluido el parámetro de interés con una probabilidad especificada.

inventario de personalidad: Inventario que mide una o más características que, por lo general, se consideran como atributos psicológicos o tendencias interpersonales.

inventario: Cuestionario o lista de comprobación que obtiene información sobre las opiniones, intereses, actitudes, preferencias, características personales, motivaciones o reacciones típicas de un individuo ante situaciones y problemas.

ítem: Enunciado, pregunta, ejercicio o tarea de una prueba en el que el examinando debe seleccionar o construir una respuesta, o realizar una tarea. Véase *indicación*.

ítems de anclaje: Ítems administrados con cada uno de dos o más formularios alternativos de una prueba con la finalidad de equiparar los puntajes obtenidos en estos formularios alternativos.

ítems, tareas o ejercicios de respuesta construida: Ítems, tareas o ejercicios cuyas respuestas o productos propios deben crear los examinandos, en lugar de elegir una respuesta de un conjunto definido. Los ítems de respuestas cortas requieren como respuesta unas pocas palabras o un número; los ítems de respuestas extendidas requieren al menos unas pocas frases y pueden incluir diagramas, pruebas matemáticas, ensayos o soluciones de problemas como, por ejemplo, reparaciones de red u otros productos de trabajo.

laboratorio cognitivo: Método de estudio de los procesos cognitivos que los examinandos usan cuando llevan a cabo tareas como, por ejemplo, resolver un problema matemático o interpretar un texto, y que por lo general comporta que el examinando piense en voz alta mientras responde la tarea o responda a preguntas de entrevista después de realizar la tarea.

licencia: Concesión (por lo general, por parte de una agencia gubernamental) de autorización o permiso legal para la práctica de una ocupación o profesión. Véase *certificación, acreditación*.

manual de la prueba: Publicación preparada por los desarrolladores o editores de la prueba para proporcionar información sobre la administración, el puntaje y la interpretación de la prueba, y para facilitar datos técnicos seleccionados sobre

las características de la prueba. Véase *guía de usuario, manual técnico*.

manual técnico: Publicación preparada por los desarrolladores o editores de la prueba para facilitar información técnica o psicométrica sobre una prueba.

medición de desempeño laboral: Medición del desempeño laboral observado del titular de un cargo, evaluado mediante una prueba de trabajo, una evaluación de conocimientos o calificaciones del desempeño real en el trabajo del titular. Véase *prueba de trabajo*.

meta-análisis: Método estadístico de investigación en el cual se combinan los resultados de estudios comparables e independientes para determinar la dimensión de un efecto global o el grado de relación entre dos variables.

modelos de crecimiento: Modelos estadísticos que miden el progreso de los estudiantes en las pruebas de rendimiento mediante la comparación de los puntajes de los mismos estudiantes a lo largo del tiempo. Véase *modelos de valor añadido*.

modelos de valor añadido: Estimación de la contribución de las escuelas o profesores individuales al desempeño de los estudiantes a través de técnicas estadísticas complejas que usan datos de resultados de varios años, los cuales suelen ser puntajes de pruebas estandarizados. Véase *modelos de crecimiento*.

moderación: Proceso de relacionar puntajes de pruebas diferentes de manera que los puntajes tengan el mismo significado relativo.

modificación de prueba: Cambios hechos en el contenido, formato o procedimiento de administración de una prueba para aumentar la accesibilidad de la prueba para los examinandos que no pueden realizar la prueba original bajo condiciones estándar. A diferencia de las adecuaciones de las pruebas, las modificaciones cambian en cierto grado el constructo que mide la prueba y, por lo tanto, cambian las interpretaciones del puntaje. Véase *adaptación/adaptación de prueba, modificación/modificación de prueba.* Compárese con *adecuación/adecuaciones de prueba*.

modificación/modificación de prueba: Cambio en el contenido de la prueba, el formato (incluido los formatos de las respuestas) o las condiciones de administración, y que se aplica para aumentar la accesibilidad de algunas personas, pero que también afecta al constructo medido y, en consecuencia, a los resultados de los puntajes que difieren en significado de los puntajes de evaluaciones no modificadas.

monitor: En administración de pruebas, la persona responsable de supervisar el proceso de la prueba y de implementar los procedimientos de administración de la prueba.

muestra aleatoria estratificada: Conjunto de muestras aleatorias, cada una de tamaño definido, que provienen de diferentes conjuntos considerados como estratos de una población. Véase *muestra aleatoria, muestra*.

muestra aleatoria: Selección a partir de una población definida de entidades según un proceso aleatorio, con la selección de cada entidad independiente de la selección de otras entidades. Véase *muestra*.

muestra: Selección de un número definido de entidades, denominadas *unidades de muestreo* (examinandos, ítems, etc.), a partir de un conjunto especificado más grande de entidades posibles, denominado población. Véase *muestra aleatoria, muestra aleatoria estratificada*.

muestreo de dominio o contenido: Proceso de selección sistemática de ítems de prueba para representar el conjunto total de ítems que miden un dominio.

muestreo de matriz: Formato de medición en el que un gran conjunto de ítems de una prueba se organiza en un número de conjuntos de ítems relativamente pequeños, cada uno de los cuales se asigna aleatoriamente a una submuestra de examinandos, evitando así la necesidad de administrar todos los ítems a todos los examinandos. No se presume la equivalencia de los conjuntos de ítems pequeños o subconjuntos.

nivel de desempeño: Etiqueta o breve enunciado que clasifica la competencia del examinando en

un dominio concreto, por lo general, definido por un rango de puntajes de una prueba. Por ejemplo, etiquetas como "básico" a "avanzado" o "principiante" a "experto" constituyen rangos generales para la clasificación de la destreza. Véase *niveles de rendimiento, puntaje de corte, descriptor de nivel de desempeño, fijación de estándar*.

nivel de participación: Grado en el que un examinando participa de forma apropiada en la ejecución de la prueba.

niveles de rendimiento/niveles de destreza: Descripción de los niveles de competencia de los examinandos en un área específica de conocimientos o capacidad; por lo general, se define en términos de categorías ordenadas en un continuum, por ejemplo, de "básico" a "avanzado," o "principiante" a "experto". Las categorías constituyen rangos generales para la clasificación del desempeño. Véase *puntaje de corte*.

normas de usuario: Estadísticas descriptivas (incluyendo los rangos de percentil) para un grupo de examinandos que no representa una población de referencia bien definida, por ejemplo, todas las personas evaluadas durante un determinado periodo de tiempo o un conjunto de examinandos autoseleccionados. Véase *normas locales, normas*.

normas locales: Normas por las cuales los puntajes de una prueba se remiten a una población de referencia limitada y específica de interés particular para el usuario de la prueba (p. ej., la población de una localidad, organización o institución). Las normas locales no pretenden ser representativas de las poblaciones más allá del contexto limitado.

normas: Estadísticas o datos tabulares que resumen la distribución o frecuencia de puntajes de prueba para uno o más grupos definidos (por ejemplo, examinandos de diversas edades o grados), diseñados por lo general para representar poblaciones más grandes, a las que se denomina *poblaciones de referencia*. Véase *normas locales*.

oportunidad de aprendizaje: Grado de exposición de los examinandos a los constructos evaluados a través de los programas educativos y/o grado de exposición o de experiencia con el idioma o la cultura mayoritaria requeridos para entender la prueba.

orientación: Actividades de instrucción planificadas a corto plazo para los posibles examinandos, facilitadas antes de la administración de la prueba con el propósito principal de mejorar sus puntajes en las pruebas. Por lo general, las actividades que aproximan la instrucción proporcionada por los planes de estudio escolar o los programas de capacitación ordinarios no se suelen considerar orientación.

parámetro de capacidad: En teoría de respuesta al ítem (IRT, por sus siglas en inglés), valor teórico que indica el nivel de un examinando respecto de la capacidad o rasgo medido por la prueba; análogo al concepto de puntaje verdadero en la teoría clásica de los tests.

percentil: Puntaje de una prueba por debajo del cual se produce un porcentaje determinado de puntajes para una población específica.

población de referencia: Población de examinandos con la que se comparan los examinandos individuales a través las normas de prueba. La población de referencia se puede definir en términos de edad, grado, estado clínico del examinando en el momento de la prueba, o por otras características. Véase *normas*.

porfolio: En evaluación, una recopilación sistemática de productos educativos o de trabajo que se han reunido o acumulado a lo largo del tiempo, de acuerdo con un conjunto específico de principios o reglas.

precisión de la clasificación: Grado de precisión de la asignación de examinandos a categorías específicas; grado en que se evitan las clasificaciones de falsos positivos y falsos negativos. Véase *sensibilidad, especificidad*.

precisión de medida: Impacto de un error de medida en los resultados de la medida. Véase *error estándar de medida, error de medida, confiabilidad/precisión*.

programa educativo individualizado (IEP, por sus siglas en inglés): Plan documentado que

perfila los servicios de educación especial para estudiantes con necesidades especiales y que incluye las adaptaciones necesarias en el aula habitual o en las evaluaciones, y los programas o servicios especiales adicionales.

protocolo de respuesta: Registro de las respuestas dadas por un examinando a una prueba específica.

proyección: Método de vinculación de puntajes en el cual los puntajes de una prueba se usan para predecir los puntajes de otra prueba para un grupo de examinandos, con frecuencia, usando metodología de regresión.

prueba adaptable computarizada: Prueba adaptable administrada mediante computadora. Véase *prueba adaptable*.

prueba adaptable: Forma secuencial de pruebas individuales en la que se seleccionan ítems sucesivos de la prueba, o conjuntos de ítems, para su administración, basándose principalmente en sus propiedades y contenidos psicométricos, en relación con las respuestas del examinando a ítems anteriores.

prueba administrada por computadora: Prueba administrada mediante computadora; los examinandos responden mediante el uso del teclado, el ratón u otros dispositivos de respuesta.

prueba basada en computadora: Véase *prueba administrada por computadora*.

prueba de alto riesgo: Prueba usada para obtener resultados que tienen consecuencias directas y significativas para las personas, programas o instituciones que participan en la prueba. Compárese con *prueba de bajo riesgo*.

prueba de anclaje: Conjunto de ítems de anclaje usado para la equiparación.

prueba de bajo riesgo: Prueba usada para obtener resultados que solo tienen consecuencias menores o indirectas para las personas, programas o instituciones que participan en la prueba. Compárese con *prueba de alto riesgo*.

prueba de campo: Administración de una prueba que se utiliza para comprobar la idoneidad de los procedimientos de la evaluación y las características estadísticas de nuevos ítems o formularios de la prueba. Por lo general, una prueba de campo es más extensa que una prueba piloto. Véase *prueba piloto*.

prueba de cribado: Prueba que se utiliza para establecer categorizaciones amplias de examinandos como primer paso en decisiones de selección o procesos de diagnóstico.

prueba de destreza basada en computadora: Prueba administrada mediante computadora que indica si el examinando ha conseguido un nivel determinado de competencia en un dominio específico, en lugar del grado de rendimiento del examinando en ese campo. Véase *prueba de destreza*.

prueba de destreza: Prueba diseñada para indicar si un examinando ha alcanzado un nivel previsto de competencia o destreza en un dominio. Véase *puntaje de corte, prueba de destreza basada en computadora*.

Prueba de grupo: Prueba para grupos de examinandos; por lo general, en un contexto grupal, con procedimientos de administración estandarizados y supervisados por un monitor o administrador de la prueba.

prueba de inteligencia: Prueba diseñada para medir el nivel de funcionamiento cognitivo de un individuo de acuerdo con una teoría de inteligencia reconocida. Véase *evaluación cognitiva*.

prueba de rendimiento: Prueba para medir el nivel de conocimientos o capacidad logrado por un examinando en un dominio de contenido sobre el cual ha recibido instrucción.

prueba de tiempo: Prueba administrada a los examinandos a los que se asigna un lapso de tiempo prescrito para responder a la prueba.

prueba de trabajo: Prueba de la capacidad de una persona para realizar las tareas que comprende un trabajo. Véase *medición de desempeño laboral*.

prueba piloto: Prueba administrada a una muestra de examinandos para probar algunos aspectos

o ítems de la prueba, por ejemplo, las instrucciones, los límites de tiempo, los formatos de respuesta o las opciones de respuesta a ítems. Véase *prueba de campo*.

prueba unidimensional: Prueba que solo mide una dimensión o solo una variable latente.

prueba: Dispositivo de evaluación o procedimiento en el cual se obtiene y puntúa una muestra sistemática del comportamiento de un examinando en un dominio específico, a través de un proceso estandarizado.

pruebas psicológicas: Uso de pruebas o inventarios para evaluar las características particulares de una persona.

psicodiagnóstico: Formalización o clasificación del estado de salud mental basada en evaluaciones psicológicas.

puesto: En contextos de empleo, la unidad organizativa más pequeña, un conjunto de deberes y responsabilidades asignados que una persona lleva a cabo dentro de una organización.

puntaje agregado: Puntaje total formado por la combinación de puntajes relacionados con la misma prueba o con diversos componentes de la prueba. Los puntajes pueden ser brutos o estandarizados. Los componentes del puntaje agregado se pueden ponderar o no, en función de la interpretación que se dé al puntaje agregado.

puntaje analítico: Método de puntuar respuestas construidas (por ejemplo, ensayos) en el que cada dimensión crítica de un desempeño específico se evalúa y califica por separado, y los valores resultantes se combinan para obtener un puntaje general. En algunos casos, los puntajes de distintas dimensiones se pueden usar para interpretar el desempeño. Compárese con *puntaje holístico*.

puntaje automático: Procedimiento por el cual los ítems de respuestas construidas se califican por computadora usando un método basado en reglas.

puntaje bruto: Puntaje de una prueba que se calcula mediante el recuento del número de respuestas correctas, o de forma más general, la suma u otra combinación de puntajes de ítems.

puntaje compuesto: Puntaje que combina varios puntajes de acuerdo con una fórmula definida.

puntaje de corte: Punto definido en una escala de puntaje. Los puntajes que coinciden o son superiores a ese punto se reportan, interpretan o gestionan de forma diferente a los puntajes inferiores a ese punto.

puntaje de escala: Puntaje obtenido mediante la transformación de puntajes brutos. Los puntajes de escala se suelen usar para facilitar la interpretación.

puntaje de ganancia: En pruebas, la diferencia entre dos puntajes obtenidos por un examinando en una misma prueba o en dos pruebas equiparadas realizadas en diferentes ocasiones, con frecuencia, antes y después de un tratamiento.

puntaje de universo: En la teoría de generabilidad, el valor esperado sobre todas las replicaciones posibles de un procedimiento para el examinando. Véase *teoría de generabilidad*.

puntaje holístico: Método para obtener un puntaje en una prueba, o ítem de una prueba, basándose en un juicio del desempeño general y usando criterios definidos. Compárese con *puntaje analítico*.

puntaje verdadero: En teoría clásica de los tests, promedio de los puntajes que obtendría un individuo en un número ilimitado de formularios estrictamente paralelos de la misma prueba.

puntaje: Cualquier número específico resultado de la evaluación de una persona, por ejemplo, puntaje bruto, puntaje de escala, una estimación de una variable latente, un recuento de producción, un registro de ausencia, un grado escolar o una calificación.

puntajes/calificaciones ponderadas: Método de calificación de una prueba en el que se otorga un diferente número de puntos a una respuesta correcta (o diagnósticamente pertinente) en diferentes ítems. En algunos casos, la fórmula de

calificación otorga un distinto número puntos a cada respuesta diferente del mismo ítem.

rango de percentil: Rango de un puntaje determinado basado en el porcentaje de puntajes de una distribución definida de puntajes que están por debajo del puntaje que se califica.

repetición de la prueba: Administración repetida de una prueba, usando la misma prueba o un formulario alternativo, a veces con capacitación o instrucción adicional entre las administraciones.

restricción de rango o variabilidad: Reducción de la varianza del puntaje observado de una muestra de examinandos comparada con la varianza de toda la población de examinandos, como consecuencia de las restricciones del proceso de muestreo de examinandos. Véase *validez ajustada o coeficiente de confiabilidad*.

rúbrica de puntajes: Criterio establecido (incluyendo reglas, principios e ilustraciones) que se usa para puntuar las respuestas construidas a tareas individuales y agrupamientos de tareas.

rúbrica: Véase *rúbrica de puntajes*.

seguridad de la prueba: Protección del contenido de una prueba de una versión o uso no autorizado, a fin de proteger la integridad de los puntajes de manera que sean válidos para el uso previsto.

selección descendente: Selección de solicitantes sobre la base de puntajes ordenados por clasificación, de la más alta a la más baja.

selección: Aceptación o rechazo de solicitantes de una oportunidad laboral o educativa concreta.

sensibilidad: En clasificación, diagnóstico y selección, proporción de casos que se evalúan como satisfactorios –o que se prevé que satisfagan los criterios– y los que, en realidad, satisfacen los criterios.

sesgo de respuesta: Tendencia de los examinandos a responder de una forma o estilo particular a los ítems de una prueba (p. ej., asentimiento, elección de opciones socialmente deseables, elección de las opciones "verdaderas" en una prueba de verdadero-falso) que genera sistemáticamente errores irrelevantes de constructo en los puntajes de la prueba.

sesgo predictivo: Predicción sistemática excesiva o deficiente del desempeño de un criterio para personas pertenecientes a grupos diferenciados por características no relevantes al desempeño del criterio.

sesgo: 1. En imparcialidad de pruebas, infrarrepresentación del constructo o componentes irrelevantes de constructo en los puntajes de las pruebas que afectan diferencialmente el desempeño de distintos grupos de examinandos y, en consecuencia, la confiabilidad/precisión y la validez de los resultados y usos de sus puntajes. 2. En estadísticas o medición, error sistemático en un puntaje de prueba. Véase *infrarrepresentación del constructo, varianza irrelevante de constructo, imparcialidad, sesgo predictivo*.

sistema de rendición de cuentas: Sistema que aplica incentivos o sanciones en función del desempeño del estudiante a instituciones (como escuelas o sistemas escolares) o a personas (como profesores o proveedores de servicios de salud mental.

subgrupo relevante: Subgrupo de la población al cual se dirige una prueba y que es identificable de alguna manera relevante para la interpretación de los puntajes para los fines previstos.

teoría clásica de los tests: Teoría psicométrica basada en la idea de que el puntaje observado de un individuo en una prueba es la suma de un componente de puntaje verdadero del examinando y de un componente de error aleatorio independiente.

teoría de generabilidad: Modelo metodológico para la evaluación de la confiabilidad/precisión en el cual se calculan varias fuentes de varianza de error a través de la aplicación de técnicas estadísticas de análisis de varianza. El análisis indica la generabilidad de los puntajes por encima de la muestra específica de los ítems, las personas y las condiciones de observación que se estudiaron. También denominada teoría G.

teoría de respuesta al ítem (IRT, por sus siglas en inglés): Modelo matemático de la relación funcional entre el desempeño en un ítem de prueba, las características del ítem y la situación del examinando respecto del constructo sometido a medición.

trabajo/clasificación del trabajo: Grupo de puestos de trabajo con suficiente parecido en deberes, responsabilidades, características requeridas y otros aspectos relevantes, de manera que se pueden colocar bajo el mismo título de puesto laboral.

uso operativo: Uso real de una prueba, después finalizado el desarrollo inicial de la prueba, para informar una interpretación, decisión o acción, basándose total o parcialmente en los puntajes de la prueba.

usuario de la prueba: Persona o entidad responsable de la elección y administración de una prueba, de la interpretación de los puntajes producidos en un contexto dado y de cualquier decisión o acción que se base, en parte, en los puntajes de una prueba.

Validación cruzada: Procedimiento en el que un sistema de puntaje para la predicción del desempeño, derivado de una muestra, se aplica a una segunda muestra para investigar la estabilidad de la predicción de ese sistema.

validación: Proceso mediante el cual se investiga la validez de una interpretación propuesta de los puntajes de una prueba para los usos previstos.

validez ajustada o coeficiente de confiabilidad: Coeficiente de validez o confiabilidad —con mayor frecuencia, una correlación producto-momento— que ha sido ajustado para compensar los efectos de las diferencias en la variabilidad de puntajes, la variabilidad de criterios o la falta de confiabilidad de los puntajes de las pruebas o criterios. Véase *restricción de rango o variabilidad*.

validez: Grado en que la evidencia acumulada y la teoría respaldan una interpretación específica de los puntajes de una prueba para un uso determinado. Si se prevén varias interpretaciones del puntaje de una prueba para diferentes usos, serán necesarias evidencias de validez para cada interpretación.

variable moderadora: Variable que afecta a la dirección o intensidad de la relación entre dos variables diferentes a aquella.

varianza irrelevante de constructo: Varianza en puntajes de examinandos atribuible a factores extrínsecos que distorsionan el significado de los puntajes y, por lo tanto, reducen la validez de la interpretación propuesta.

vinculación/vinculación de puntajes: Proceso de relacionar puntajes de pruebas. Véase *formularios alternativos, equiparación, calibración, moderación, proyección, escalamiento vertical*.

ÍNDICE