Cuadernillo técnico
de evaluación educativa

2

Confiabilidad, validez e imparcialidad en evaluación educativa

Centro UC
Medición - MIDE

Textos de divulgación

INEE
Instituto Nacional para la Evaluación de la Educación
México

Cuadernillo técnico
de evaluación educativa

2

Confiabilidad, validez e imparcialidad en evaluación educativa

Centro UC
Medición - MIDE

Textos de divulgación

INEE
Instituto Nacional para la Evaluación de la Educación
México

Confiabilidad, validez e imparcialidad en evaluación educativa

Av. Vicuña Mackenna 4860
Macul, Santiago, Chile, cp 7820436

Barranca del Muerto 341, col. San José Insurgentes,
Alcaldía Benito Juárez, Ciudad de México, cp 03900

Autora
Andrea Abarzúa Morasso, MIDE UC

Editora
María Rosa García González, MIDE UC

Corrección de estilo
Arturo Cosme Valadez, INEE
Lissette Sepúlveda Cepeda, MIDE UC

Coordinación General
Adriana Guadalupe Aragón Díaz, INEE
Marcela Cuevas Ossandón, MIDE UC
Marcela Ramírez Jordán, INEE

Diseño
www.iunta.cl

Índice

Presentación 1

Resumen 2

Los tres pilares fundamentales en una medición: confiabilidad, validez e imparcialidad 3

Confiabilidad: ¿qué significa que una medición sea confiable? 7

Validez: ¿de qué hablamos cuando hablamos de validez? 11

Imparcialidad: ¿cómo podemos asegurar la imparcialidad de nuestra medición? 22

Consideraciones finales: ideas fuerza 31

Referencias 32

Presentación

El Instituto Nacional para la Evaluación de la Educación de México, INEE, y el Centro de Medición MIDE UC, de la Pontificia Universidad Católica de Chile, han gestado una colaboración para el desarrollo y fortalecimiento de capacidades en evaluación educativa, en profesionales del Instituto y de los equipos responsables de los Programas Estatales de Evaluación y Mejora Educativa (PROEME) y del Proyecto Nacional de Evaluación y Mejora Educativa de Escuelas Multigrado (PRONAEME), en el marco del Sistema Nacional de Evaluación Educativa (SNEE), en México.

El documento que a continuación presentamos constituye un material de consulta que forma parte de una serie de nueve cuadernillos, cuyo propósito es orientar la comprensión de los conceptos centrales de la medición y la evaluación educativas y su impacto en el diseño de instrumentos; considerando que el proceso evaluativo es una suma de decisiones que deben cuidar la coherencia de cada uno de los elementos y fases que lo componen.

Este material se ha organizado en una serie de cuadernillos con base en las siguientes temáticas:

1. Nociones básicas en medición y evaluación en el contexto educativo.
2. Confiabilidad, validez e imparcialidad en evaluación educativa.
3. Definición del marco de referencia de la evaluación.
4. Desarrollo de instrumentos de evaluación: pruebas.
5. Desarrollo de instrumentos de evaluación: cuestionarios.
6. Desarrollo de instrumentos de evaluación: pautas de observación.
7. Desarrollo de instrumentos de evaluación: tareas de desempeño y rúbricas.
8. Análisis y uso de resultados.
9. Uso de resultados y retroalimentación.

Esperamos que este material resulte de utilidad para los profesionales que se desempeñan en el contexto de la medición y evaluación educacional. En los cuadernillos encontrarán nociones y conceptos fundamentales, además de recomendaciones prácticas, y sugerencias bibliográficas para quienes deseen profundizar en cada una de las temáticas trabajadas.

Confiabilidad, validez e imparcialidad en evaluación educativa

Resumen

El presente capítulo aborda los tres pilares fundamentales de la calidad técnica de una medición, como son: su nivel de confiabilidad respecto de los puntajes que entrega, la evidencia de validez de los usos de la información que genera para cumplir sus propósitos intencionados, y su imparcialidad o ecuanimidad en relación con los examinados.

En un primer apartado se presentan los tres pilares de forma introductoria, y se señala brevemente su historia, enfatizando el hecho de que estos pilares se encuentran en permanente revisión y enriquecimiento para estar al día con las nuevas metodologías en medición como un campo en permanente desarrollo.

En los tres apartados siguientes se aborda cada uno de los pilares, detallando su conceptualización según los estándares establecidos por la American Educational Research Association [AERA], la American Psychological Association [APA] y el National Council on Measurement in Education [NCME] el año 2014. Para cada estándar se presentan ejemplos tomados de mediciones estandarizadas a gran escala, tanto en Latinoamérica como a nivel mundial.

Mediante este acercamiento a los estándares, se espera que el lector pueda comprender las nociones fundamentales de los tres pilares y contar con herramientas para la lectura comprensiva de documentación técnica sobre las mediciones estandarizadas a gran escala.

I. Los tres pilares fundamentales en una medición: confiabilidad, validez e imparcialidad

Origen de los estándares, historia y evolución

El propósito declarado de los estándares para la medición educativa y psicológica es proveer de un conjunto de criterios sobre el desarrollo y evaluación de mediciones y prácticas de medición, así como entregar guía con el fin de juzgar la validez de las interpretaciones de los puntajes para los usos intencionados de estas [AERA], [APA] y [NCME], 2014).

Los estándares más recientes se publicaron el año 2014, en su quinta versión. El primer set de estándares fue publicado el año 1954 por la APA, en un documento llamado *Technical Recommendations for Psychological Tests and Diagnostic Techniques.* Al año siguiente, AERA y NCME publicaron un set llamado *Technical Recommendations for Achievement Tests*. En el año 1966 se formó el Comité Conjunto, con la finalidad de consolidar y revisar estos estándares en un único set. Este referente para la medición fue revisado en 1974, 1985, 1999 y, por último, en 2014, dando paso a la versión actual (AERA, APA y NCME, 2014).

Los estándares publicados en 2014 presentan dos cambios respecto de la versión de 1999, que se constituyen en temas centrales en este documento. El principal consiste en una consolidación de la ecuanimidad como un tema distinto, y del mismo nivel de importancia, que los pilares tradicionales de validez y confiabilidad. En segundo lugar, ofrece una conceptualización de la confiabilidad realzando su significado como precisión de los puntajes entregados por un test. En este cuadernillo se profundizará en las nociones de confiabilidad, validez y ecuanimidad, siguiendo la revisión actualizada de los estándares.

Introducción a los pilares fundamentales y ejemplos en los que mediciones internacionales los han abordado

A modo de introducción, se presentará el siguiente resumen esquemático de los tres pilares fundamentales de las mediciones: confiabilidad, validez y ecuanimidad, y se explicará brevemente a qué refiere cada uno de ellos, acompañándolos de ejemplos concretos disponibles en la literatura especializada.

Primera premisa

Toda medición busca responder a un propósito específico, sea de diagnóstico, de acreditación u otros. Una medición es, por tanto, una herramienta para un uso previamente intencionado. La correspondencia entre los resultados de una medición y su empleo intencionado debe ser respaldada por evidencias teórica y empírica. A estas les llamamos *evidencia de validez* de una medición. En el ejemplo se muestra cómo una medición internacional ha recabado evidencia de validez.

Ejemplo de evidencia de validez basada en el contenido de test: ICCS 2009

Panel de expertos

El panel de expertos es una estrategia basada en el trabajo en equipo para la revisión de los materiales que forman parte de una medición (preguntas, instrucciones, ítems). Es un riguroso mecanismo de control de calidad que se lleva a cabo durante la fase de desarrollo de los materiales de la medición. El uso de un panel de expertos es un proceso que reconoce la importancia de exponer el material a múltiples puntos de vista. Durante este proceso, un pequeño grupo (entre tres y seis) de desarrolladores de test revisa en conjunto los materiales que uno o más de ellos han desarrollado. Esta revisión conduce a la aceptación, modificación o rechazo de los materiales. Los integrantes del panel comparan sus respuestas con las preguntas o ítems y las problematizan. La discusión acerca de los potenciales problemas de los ítems debe ser fuertemente sustentada para poder asegurar que los ítems seleccionados se comporten de la forma intencionada. Las siguientes preguntas sintetizan los cuestionamientos que son foco central en la evaluación de los ítems desarrollados para International Civic and Citizenship Study (ICCS). La relevancia de cada cuestionamiento varía de acuerdo con las características del material que está siendo sometido a revisión.

Validez de contenido

- ¿Cómo se relaciona el material con las especificaciones de la prueba ICCS?
- Las preguntas, ¿miden el contenido y procesos cognitivos descritos en el marco de evaluación?
- Las preguntas, ¿apuntan a aspectos esenciales del estímulo o se orientan a aspectos tangenciales de este?
- ¿Cómo sería juzgado este material en caso de ser sometido a escrutinio público (incluyendo tanto al equipo que forma parte del proyecto, como a las personas no relacionadas directamente con esta medición)?

Fuente: Traducido al español por la autora, desde Schulz, Ainley y Fraillon (2011).

Segunda premisa

Una medición no puede ser válidamente usada si no se cuenta con evidencia acerca de la credibilidad de los puntajes que genera. Si los puntajes que entrega una medición no son suficientemente precisos, o muestran alto grado de error de medición, ningún uso que pueda hacerse de ellos será argumentable como válido. Al nivel de confianza sobre los puntajes que entrega un instrumento le llamamos *confiabilidad.* En el ejemplo se muestra evidencia de confiabilidad de una medición internacional.

Ejemplo de confiabilidad de una medición: TIMSS 2015

La tabla 1 muestra los coeficientes de confiabilidad de las pruebas de Matemáticas y Ciencias de cuarto grado para cada país. Estos coeficientes corresponden a la mediana de la confiabilidad (Alpha de Cronbach) de los cuadernillos que forman parte de la medición de las Trends in International Mathematics and Science Study TIMSS 2015 en cuarto grado. En general, las confiabilidades fueron relativamente altas. La mediana de las confiabilidades de los países fue de 0.83 para Matemáticas y 0.78 para Ciencias.

TABLA 1

MEDIANA DE LAS CONFIABILIDADES DE LOS CUADERNILLOS PARA CADA PAÍS Y MEDIANA INTERNACIONAL

País	Matemáticas	Ciencias
Alemania	0.82	0.77
Arabia Saudita	0.76	0.80
Australia	0.86	0.79
Baréin	0.81	0.82
Bulgaria	0.86	0.85
Canadá	0.82	0.79
Chile	0.80	0.76
China	0.83	0.77
Chipre	0.85	0.77
Corea	0.82	0.75
Croacia	0.81	0.73
Dinamarca	0.84	0.76
Emiratos Árabes	0.87	0.85
Eslovenia	0.82	0.78
España	0.80	0.77
Estados Unidos	0.87	0.82
Finlandia	0.81	0.74
Francia	0.82	0.78
Georgia	0.82	0.76
Hong Kong SAR	0.81	0.77
Hungría	0.88	0.82
Indonesia	0.76	0.76
Inglaterra	0.86	0.77
Irán	0.83	0.80

País	Matemáticas	Ciencias
Irlanda	0.84	0.77
Irlanda del Norte	0.87	0.77
Italia	0.82	0.75
Japón	0.83	0.77
Kasajistán	0.86	0.81
Kuwait	0.76	0.78
Lituania	0.83	0.77
Marruecos	0.76	0.78
Noruega	0.83	0.72
Nueva Zelanda	0.85	0.82
Omán	0.83	0.84
Países Bajos	0.77	0.71
Polonia	0.83	0.78
Portugal	0.84	0.72
Qatar	0.84	0.82
República Checa	0.83	0.78
República Eslovaca	0.84	0.82
Rusia	0.84	0.77
Serbia	0.87	0.80
Singapur	0.88	0.83
Suecia	0.81	0.79
Turquía	0.87	0.81
Promedio Internacional	**0.83**	**0.78**

Fuente: Adaptado y traducido al español por la autora, desde Martin, Mullis y Hooper (2016).

Tercera premisa

Una medición, para cumplir su propósito adecuadamente, debe permitir al examinado demostrar su potencial real, independientemente de su género, idioma u otras características personales que no se relacionan con el propósito de la medición. Si ello se encuentra interferido por prácticas del proceso de implementación de la medición, o bien, por características de los ítems o tareas propuestas a los examinados, los resultados que se obtengan beneficiarán o perjudicarán a ciertos examinados por sobre otros. Reducir este riesgo corresponde al aseguramiento de la *imparcialidad* de una medición. A continuación, el ejemplo muestra un proceso, seguido por una medición internacional para asegurar imparcialidad.

Ejemplo de proceso para asegurar imparcialidad de una medición: PIRLS 2016

Documentando las adaptaciones nacionales

Todos los ajustes de la versión internacional de la medición de lectura y sus cuestionarios de contexto fueron documentados en las Formas Adaptadas Nacionales. Para cada instrumento se completó un formulario que enumeraba los cambios realizados y, en los casos excepcionales de preguntas no administradas en un país, se indicó la razón detrás de esta decisión. Estos formularios fueron actualizados luego de cada etapa de verificación del proceso.

Traduciendo los materiales de evaluación de PIRLS

Cada traductor y revisor recibió la versión internacional del set de materiales de evaluación del Progress in International Reading Literacy Study (PIRLS) 2006 para ser traducidos. También recibieron información con el fin de familiarizarse con la medición PIRLS y los procedimientos de traducción y Formularios de Adaptaciones Nacionales. Los traductores usaron estos materiales para traducir cada uno de los instrumentos, siguiendo las indicaciones de adaptación descritas previamente en este capítulo. Si hubo más de un traductor dedicado a un idioma, el traductor que trabajó en un fragmento específico también se hizo cargo de las preguntas correspondientes a ese fragmento. Durante el proceso de traducción, los traductores debieron documentar todos los cambios hechos en los textos originales en una versión electrónica de las Formas Adaptadas Nacionales. Este set de materiales traducidos luego se entregó a un revisor, cuya tarea fue asegurar que las traducciones fuesen apropiadas para la población de interés de la medición. Las sugerencias de los revisores luego fueron incorporadas por los traductores en los materiales, y las formas se actualizaron en función de ello.

Fuente: Traducido al español por la autora desde Martin, Mullis y Hooper (2017).

¿Cómo hacerse cargo de los estándares?

AERA, APA y NCME (2014) señalan de manera explícita que someter a juicio una medición y/o práctica evaluativa es un ejercicio profesional y, en ese sentido, el interés de los estándares es proporcionar un marco de referencia que permita asegurar que todos los aspectos clave sean considerados. Así, podemos encontrar estándares que cubren aspectos técnicos, profesionales y operacionales de distintas formas de medición, y que son usados en variados contextos. Estos estándares aplican tanto a desarrolladores de test, como a

quienes los financian, los publican y a sus usuarios, entregando criterios específicos para la evaluación de los test mismos, las prácticas evaluativas y los efectos del uso de los test que componen una determinada medición.

II. Confiabilidad: ¿qué significa que una medición sea confiable?

Noción global de confiabilidad: precisión y replicabilidad

La confiabilidad de una medición es la propiedad más importante de un test, siendo un requisito para alcanzar los otros dos pilares fundamentales: validez e imparcialidad. Conceptualmente, la confiabilidad se opone al concepto de error de una medición, es decir, una medición que entrega resultados afectados por error de medición no puede ser argumentada como válida ni como imparcial (AERA, APA y NCME, 2014).

Los estándares de AERA, APA y NCME (2014) advierten que el término confiabilidad ha sido usado de múltiples maneras, confundiéndose con los distintos procedimientos de estimación que se han desarrollado, y aludiendo a las distintas facetas de la medición en las cuales tal propiedad se puede analizar. Para unificar esta noción, los estándares proponen conceptualizar la confiabilidad como precisión de la medición a nivel de los examinados. Esto enfatiza la noción de replicabilidad de los puntajes a nivel individual que entrega un determinado instrumento. Así, un buen análisis de confiabilidad permite estudiar las posibles amenazas a la replicabilidad de un test.

Fuentes de imprecisión y dificultades para la replicabilidad en distintos formatos de medición

La confiabilidad es comúnmente estudiada y reportada en mediciones, sin embargo, rara vez se explicita la pertinencia de los estimadores de confiabilidad que se reportan, siendo muy infrecuente encontrar argumentos acerca de la adecuación de estimadores específicos para una medición en particular (Cronbach y Shavelson, 2004).

Con el fin de decidir la estimación de confiabilidad apropiada para una medición, deben tenerse presente las fuentes de error o imprecisión propias del tipo de medición en particular. Recordemos que la medición educacional y psicológica se orienta a la medición de constructos inobservables (latentes), a través de manifestaciones conductuales (observables) de estos constructos. Las manifestaciones conductuales pueden ser categorizadas en dos grandes tipos para un conocimiento X:

A) Al examinado se le presenta una pregunta o problema y, en función de su conocimiento de X, debe identificar la respuesta en un set de opciones de respuesta.

B) Al examinado se le presenta una instrucción o tarea, en función de la cual debe dejar un registro (escrito, filmación) en el que demuestre que puede resolverla correctamente dado su conocimiento de X.

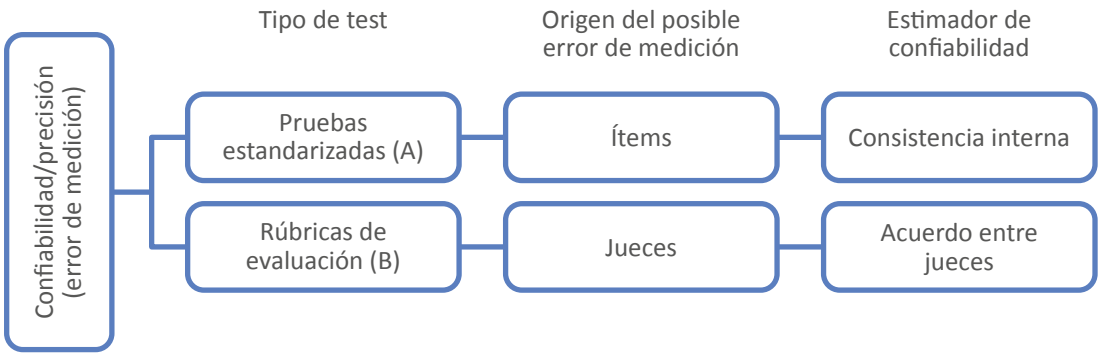
En ambos casos, cada pregunta o tarea busca ser una muestra de su nivel de conocimiento X, y es cuidadosamente seleccionada para dar cuenta de manera consistente acerca de ese conocimiento, pero la consistencia se encuentra limitada por la tarea en sí.

En el caso de las manifestaciones de tipo A, un ítem cuidadosamente construido puede hacer una diferenciación gruesa de los examinados, ordenándolos en dos grupos: quienes aciertan y quienes fallan. Tal ordenamiento es, por definición, impreciso y por tanto no es confiable o, dicho en otros términos, ordena a los examinados cometiendo una gran cantidad de errores. Solamente la acumulación de muchas tareas como A, y empíricamente consistentes con A, permitirá reunir suficiente información acerca del conocimiento X para generar un ordenamiento de los examinados que sea preciso y confiable. Un caso típico de este tipo de manifestación conductual es la respuesta a ítems de opción múltiple.

En el caso de las manifestaciones de tipo B, la tarea se especifica de forma que el examinado pueda realizarla sin dificultades, pero la medición de su nivel de conocimiento recae en manos de un experto que juzga la manifestación conductual con base en una serie de criterios previamente descritos. Esto agrega una capa de complejidad al problema anterior, ya que el experto es una nueva fuente de error de medición (puede verse afectado por cansancio o distracción, puede ver influenciado su juicio por razones imprevistas, puede tener un error de procedimiento al decidir el puntaje a asignar al examinado, entre otras). En este caso, no se puede afirmar que las mediciones de los examinados son precisas, a menos que se estudie el nivel de confluencia de juicios de jueces independientes en relación con la evidencia. Un caso típico de este tipo de manifestación conductual es la medición con base en rúbricas de evaluación. La figura 1 ilustra la distinción descrita.

FIGURA 1

RESUMEN ESQUEMÁTICO DE LAS FUENTES DE ERROR Y ESTIMADORES DE CONFIABILIDAD



Fuente: elaboración propia.

Distintas formas de estimar la confiabilidad y pertinencia según el tipo de medición

En mediciones compuestas por tareas de tipo A, la forma más habitual de estudiar la confiabilidad es el índice de consistencia interna. Se trata de una medida de replicabilidad, bajo el supuesto de que un test altamente consistente resistiría variaciones en la cantidad de ítems (entendidas como muestras representativas del mismo test), entregando estimaciones de habilidad estables o invariantes. Entre estas medidas, la más popular es el coeficiente Alpha de Cronbach (1951) y su correcta interpretación consiste en reportarlo como el porcentaje de las diferencias individuales encontradas (descritas como la varianza de los puntajes obtenidos por los examinados en el test) que es atribuible a la varianza de sus puntajes verdaderos en el atributo medido (Cronbach y Shavelson, 2004).

El índice de consistencia interna o Alpha de Cronbach (1951), como medida de confiabilidad, adopta valores entre 0 y 1. Se interpreta como la proporción de examinados correctamente ordenados, es decir, libre de error. Convencionalmente se considera deseable que la confiabilidad se encuentre en el valor de 0.85, entonces diríamos que el 85% del ordenamiento resultante se encuentra libre de error (es preciso y replicable). Inversamente, diríamos que el 15% del ordenamiento resultante de los examinados es atribuible a error.

En cambio, en mediciones compuestas por tareas de tipo B, la primera fuente de error de medición que debe ser aminorada es la variación entre jueces. En este tipo de instrumentos, la replicabilidad se sustenta cuando el puntaje obtenido por el examinado es invariante al juez que revise su evidencia y asigne puntuaciones. Para estudiar si los jueces son apegados a la rúbrica, se puede recurrir a:

- El porcentaje de acuerdo interjueces.
- Índices de acuerdo entre jueces (Kappa, T-index).
- Índices de consistencia entre jueces (ICC).
- Teoría de Generalizabilidad.

El índice de acuerdo interjueces es la medida más directa para estudiar la confiabilidad de una medición con base en juicios expertos. Se expresa como porcentaje. Si dos jueces muestran un acuerdo de 80%, significa que 20% de los puntajes que generan tiene error. Es prácticamente imposible que en un proceso de evaluaciones independientes el acuerdo interjueces sea de 100%. Por esta razón, en algunas mediciones se recurre a intervalos de acuerdo considerados como razonables.

Por otra parte, los índices de acuerdo interjueces tales como Kappa de Cohen y el T-index de Lawlis y Lu (1972) se sustentan sobre una prueba de significación estadística basada en una distribución Chi Cuadrado, siendo el T-index más apropiado en el caso de variables de tipo ordinal (Tinsley y Weiss, 2000).

Los índices de consistencia entre jueces, típicamente estimados mediante una correlación intraclase, dan cuenta del grado en que distintos jueces se asemejan cuando son expresados como desviaciones de sus medias. Puede ser el caso que existan altos niveles de consistencia, pero con muy bajo nivel de acuerdo entre jueces (Tinsley y Weiss, 2000).

Finalmente, la teoría de la Generalizabilidad (Brennan, 1992) es un marco metodológico y conceptual que busca descomponer las distintas fuentes de error en un procedimiento de medición. Se basa en la teoría clásica de test y en la metodología de análisis de varianza (ANOVA), y se ha usado típicamente en el estudio de la confiabilidad de las mediciones de habilidades de escritura, en las cuales es relevante distinguir qué proporción de la varianza de puntajes observadas es atribuible a las evidencias sometidas a evaluación, y que proporción puede ser explicada por otras facetas del procesos como el momento de la medición, el evaluador, u otros.

Ejemplo de confiabilidad de la corrección de preguntas abiertas: TIMSS 2015

En la prueba TIMSS 2015, el TIMSS y PIRLS International Study Center revisó el comportamiento estadístico de los ítems en cada país participante, con el fin de asegurar que los ítems se comportaban de manera suficientemente comparable entre países. En el caso de las preguntas abiertas, se consideró la eliminación de ítems de las bases de datos del estudio cuando la confiabilidad de la corrección de estas preguntas al interior de un país mostraba un porcentaje de acuerdo inferior al 70%.

Fuente: Martin, Mullis y Hooper (2016).

III. Validez: ¿de qué hablamos cuando hablamos de validez?

Definición más general: de los usos e interpretaciones del test

Las conceptualizaciones acerca de la validez han cambiado y muy posiblemente sigan modificándose. La primera definición, o definición clásica de validez, es la desarrollada por Ruch (1924; citado en AERA, APA y NCME, 2014), quien indica que la validez es el grado en que un test mide aquello que se propone medir.

En los años cincuenta, emerge el concepto de **validez de constructo**, que enfatiza el hecho de que la medición psicológica y educacional busca dar cuenta de atributos no directamente observables, sino que son medidos a partir de sus manifestaciones. Esto transforma la visión anterior, poniendo el acento en la acumulación de evidencia en torno al significado de los puntajes que entrega un test, lo que trae como consecuencia la conceptualización de la validez como algo no dicotómico ni definitivo, sino como una pregunta que se encuentra siempre abierta a la acumulación de nueva evidencia. Se abandona, por tanto, la pretensión de contar con un índice de validez.

Los estándares, en su versión de 1985, entregan por primera vez un concepto unitario sobre validez, entendiéndola como la adecuación, significación y utilidad de las inferencias hechas a partir de los puntajes de un test. La validación de un test se entiende, entonces, como el proceso de acumular evidencia para apoyar tales inferencias (APA, AERA y NCME, 1985; citado en AERA, APA y NCME, 2014).

En base al desarrollo anterior, en la siguiente revisión de los estándares ya se distinguen y agregan tres tipos de validez a la validez de constructo: la de contenido, la concurrente y la predictiva. En estos estándares la conceptualización es la que sigue (APA, AERA y NCME, 1999; citado en AERA, APA y NCME, 2014):

- **Validez de constructo:** Se evalúa investigando las propiedades psicológicas que son medidas por una prueba, requiriendo simultáneamente de una aproximación lógica y empírica.
- **Validez de contenido:** Se evalúa analizando si el contenido de la prueba es una buena muestra de las situaciones o el contenido sobre el que se pretende sacar conclusiones. Los estándares explícitamente indican que este tipo de validez es especialmente importante en casos de mediciones de logro.
- **Validez predictiva:** Se evalúa analizando qué tan bien se confirman las predicciones basadas en la prueba por evidencia recolectada después de dicha prueba. Se menciona

explícitamente el uso predictivo en casos de inteligencia, vocacionales y resultados terapéuticos.

- **Validez concurrente:** Se evalúa analizando qué tan bien los resultados de la prueba corresponden a otras pruebas tomadas al mismo tiempo.

En esos mismos años, Kane (1992) y Messick (1995) entregan una reconceptualización de la noción de validez no como propiedad intrínseca a la medición, sino contextual y fuertemente ligada a los usos que se planea dar a una medición específica. Kane (1992) introduce el concepto de **argumento de validez**, mientras que Messick (1995) señala:

> La validez no es una propiedad de un test o medición, sino de los significados de los puntajes de un test. Estos puntajes dependen no solo de los ítems o estímulos presentes en el test, sino que también de las personas que responden, así como del contexto de la medición. De este modo, lo que requiere ser válido es el significado e interpretación de los puntajes, así como cualquier implicancia práctica que derive del test (p. 5).

Producto de estas reconceptualizaciones, la definición actual de validez refiere al grado en que la evidencia y la teoría respaldan el uso de los puntajes del test para determinados propósitos (AERA, APA y NCME, 2014). Por esta razón, la validez que se pone en tela de juicio no es la de un instrumento en sí, sino la de pertinencia de este para un uso en particular, mediante un proceso de acumulación de evidencia y argumentos acerca de la pertinencia de un test con respecto a su o sus usos intencionados. Por ello, se dice también que un test puede ser válido para un grupo o para un contexto determinado, pero inválido en otros grupos o contextos.

Sintéticamente, los estándares (AERA, APA y NCME, 2014) proponen que:

- Quienes desarrollen el test deben establecer claramente cómo interpretar y usar los puntajes. La población para la cual el test es apropiado tiene que estar claramente delimitada y se debe describir con precisión el constructo que el test pretende medir.

- Se debe presentar una justificación para cada interpretación recomendada y para el uso de los puntajes del test, junto con un resumen comprensivo de la evidencia y la teoría que subyacen a estos usos interpretativos.

- Si la validez para alguna interpretación común de un test no ha sido evaluada, o si dicha interpretación es inconsistente con la evidencia existente, ello se debe explicitar

y generar claras advertencias a potenciales usuarios acerca de interpretaciones que no tengan sustento.

- Si un test es usado en alguna forma que no ha sido previamente validada, es responsabilidad del usuario justificar este nuevo uso y recabar nueva evidencia en caso de ser necesario.

- Cuando se afirma o asume que la interpretación recomendada de puntajes de un test para un determinado uso conduce a determinado resultado, las bases para esperar tal resultado se deben presentar, así como la evidencia que sea relevante.

- Cuando se recomienda el uso de un test afirmando que la medición producirá algún beneficio adicional, asociado a la utilidad de la información proveniente de la interpretación de sus puntajes, quien hace la recomendación debe explicitar el fundamento que lleva a esperar tales beneficios. Se deben aportar argumentos lógicos o teóricos, además de evidencia que apoye dichos beneficios. Por otro lado, se debe ponderar adecuadamente cualquier evidencia contradictoria acerca de los beneficios, incluyendo resultados que revelen otras consecuencias, diferentes de las esperadas.

En el proceso de acumulación de evidencia, los estándares definen cinco fuentes específicas de evidencia de validez, las cuales se desarrollarán a continuación.

Las fuentes de evidencia para la validación

Evidencia basada en el contenido del test

Recordemos que la medición educacional y psicológica se orienta a la medición de constructos inobservables. Por esta razón, la medición debe permitir reunir evidencia conductual, u observable, que permita decidir en qué magnitud se encuentra este atributo en cada examinado.

El proceso de validación comienza con la declaración explícita de la interpretación del puntaje de una prueba que se quiere hacer, incluyendo la especificación del constructo que esta pretende medir. Lo más natural es que se atribuyan varios significados a los puntajes de un test, por lo que tanto desarrolladores como usuarios de la prueba tienen la obligación de hacer estas especificaciones, definiendo los aspectos del constructo que serán representados y su alcance. Idealmente, el marco conceptual debe incluir cómo el constructo representado se relaciona y distingue de otros constructos y variables (AERA, APA y NCME, 2014). Tal selección debe ser manifestada por el desarrollador de un test, declarando de forma detallada cómo el constructo a medir se expresa de forma concreta,

típicamente en la forma de una tabla de especificaciones (Lane, Raymond, Haladyna y Downing, 2016; Wise y Plake, 2016).

Los estándares de AERA, APA y NCME (2014) proponen como una fuente de evidencia de validez de este tipo, el juicio experto acerca de la pertinencia de cada componente del contenido de un test (temas, vocabulario utilizado, formato de los ítems, tareas y preguntas del test) a la medición del constructo. Koretz (2010) llama a esto **el adecuado muestreo de contenidos a incluir en una evaluación**, destacando el hecho de que un test es una muestra acotada de un amplio conjunto de posibles manifestaciones del constructo, y como cualquier muestra, es de especial interés que sea lo suficientemente representativa de estas diversas manifestaciones del atributo de interés.

Dentro de los problemas que pueden surgir al estudiar la validez de contenido de un test, podemos encontrar:

- **Subrepresentación del constructo:** grado en el cual una prueba ignora (o falla en medir) aspectos relevantes de un constructo. Jueces expertos podrían considerar que las especificaciones del test dejan fuera aspectos relevantes del constructo, y que por tanto, la medición es incompleta.

Ejemplo de definición de especificaciones de un test para la medición de habilidades escritas

Para la medición de habilidades escritas, podemos considerar que existen al menos dos grandes facetas de este constructo: una relacionada con el adecuado uso de las convenciones de escritura (ortografía, redacción) y otra relacionada con la calidad de la expresión de ideas en función del propósito comunicacional. Un juez experto debería juzgar si las especificaciones del test, que darán cuenta de cierto número de indicadores (y, por ende, puntuaciones) equilibra de manera adecuada estas dos grandes facetas del constructo de interés.

Fuente: elaboración propia.

- **Fuentes de varianza irrelevante para el constructo:** grado en que los puntajes de una prueba se ven afectados por procesos ajenos a lo que esta pretende medir. Jueces expertos pueden detectar aspectos de la medición que podrían estar influenciados por otros que no son el constructo de interés, sea inflando los puntajes (aumentando la facilidad) o disminuyéndolos (subestimando la habilidad).

Ejemplo de ítem liberado de Pruebas SEPA* 2009-2010

El siguiente ítem de lectura se orienta a medir la capacidad de identificar información explícita en un texto no literario, y tiene una dificultad media entre estudiantes chilenos de tercer grado de primaria. Según sus propiedades psicométricas, podemos afirmar que este ítem se encuentra correctamente construido para la medición de este indicador, en el grado mencionado.

Eje	Lectura
Indicador	Identificar una relacion de hechos explícitos en un texto no literario.

Los castores son roedores semi-acuáticos nativos de América del Norte y Europa. Son de color café, excepto su cola que es negra. Las hembras tienen entre 3 y 4 crías.

Son grandes arquitectos: talan árboles y embalsan las corrientes de agua para hacer lagos donde se ponen a salvo. Los diques que ellos forman llegan a medir más de 500 m de largo y son tan resistentes que soportan el peso de una persona. Los castores están adaptados a la vida en el agua, ya que tienen patas palmeadas y cola aplanada.

Se alimentan de corteza de árboles y hojas, y almacenan ramas bajo el agua para el invierno. Se alojan en un enorme montículo de ramas, que construyen en el centro del lago. Las entradas se encuentran bajo el agua, de modo que pueden entrar y salir sin ser vistos.

http://es.wikipedia.org

¿Qué características les han permitido a los castores adaptarse para vivir en el agua?

a) Sus firmes pulmones y su fuerza.

b) Sus largas aletas y su gran tamaño.

c) Su piel gruesa y su resistencia al frío.

d) Sus patas palmeadas y su cola aplanada.

Supongamos que este ítem se aplicase a estudiantes de primer grado de primaria. En tal caso, la identificación de información explícita se vería obstaculizada por la extensión del texto. En esta situación hipotética, la dificultad del ítem podría verse aumentada, pero no por la dificultad de reconocer información explícita, sino porque la extensión del texto podría generar cansancio lector en los examinados, al no encontrarse debidamente adecuada a su velocidad de lectura general. La varianza irrelevante para el constructo de interés consiste, en interpretar la falla en este ítem como la dificultad para extraer información explícita, cuando en realidad el resultado se encuentra afectado por una dificultad excesiva de la longitud del texto para los examinados.

Fuente: Centro de Medición MIDE UC (2011). * Sistema de Evaluación de Progreso del Aprendizaje

Evidencia basada en los procesos de respuesta

Los estándares de AERA, APA y NCME (2014) destacan los procesos de respuesta o resolución de una tarea como una fuente de evidencia de validez, que no solo permite respaldar que la medición efectivamente obedece al constructo de interés, sino que puede iluminar y contribuir a describir mejor el constructo a medir. Este tipo de evidencia se puede obtener de diversas formas, por ejemplo: consultando a los examinados cómo les fue posible resolver cada tarea o pregunta presentada, pidiéndoles que den a conocer su estrategia de respuesta, guardando registros de intentos de respuesta previos a la respuesta definitiva, o bien, monitoreando los tiempos de respuesta.

Ejemplo de evidencia de validez basada en los procesos de respuesta. Uso de entrevistas cognitivas para definir el tipo de proceso evaluado mediante un pool de ítems

Uso de entrevistas cognitivas para identificar el tipo de habilidad medida por un grupo de ítems, en el proceso de desarrollo de un banco de ítems del área de salud infantil, para formar parte del examen de titulación de la carrera de Enfermería.

En este estudio se contaba con un banco de 114 ítems, los que debían ser clasificados por jueces expertos en función de su complejidad (conocimiento básico y avanzado), habilidad cognitiva abordada (básica y superior) y coherencia con el propósito de la medición (ver tabla 2).

TABLA 2

TABLA DE CATEGORIZACIÓN DE ÍTEMS SEGÚN PROCESO COGNITIVO EVALUADO Y COHERENCIA CON EL PROPÓSITO DE LA MEDICIÓN

Conocimiento/ habilidades	Factual	Conceptual	Procedural	Metacognitivo
Habilidades básicas: recordar y entender.	H1C1	H1C2	H1C3	H1C4
Habilidades superiores: aplicar, analizar, evaluar y crear.	H2C1	H2C2	H2C3	H2C4
¿Esta pregunta es pertinente para el examen de grado?	Pertinente	Medianamente pertinente	No pertinente	

Se formó un panel de dieciocho jueces, que recibieron una capacitación en la tarea a realizar. Los ítems que no lograron ser identificados claramente por los jueces fueron sometidos a un estudio de los procesos de respuesta, el cual consistió en la realización de entrevistas cognitivas a cinco exalumnos egresados, con al menos un año de desempeño en el área evaluada. La entrevista cognitiva consistió en presentarle al exalumno, de forma individual, confidencial y en un tiempo máximo de 45 minutos, los ítems dudosos y se le solicitó que describiera verbalmente los procesos mentales que debía realizar para responder cada una de las preguntas, información que fue grabada. Con esta nueva información el grupo de profesores del panel de expertos procedió a clasificar el ítem.

Esta forma de reunir evidencia también aplica a los procesos de puntuación realizados por jueces. En este caso, los estándares subrayan la importancia de revisar si los procesos de puntuación que realizan los jueces sobre determinada evidencia son los correctos, en función del atributo de interés, o si se trata de acuerdos alcanzados de forma espuria.

Evidencia basada en la estructura interna del test

Según se establece en los estándares (AERA, APA y NCME, 2014), el análisis de la estructura interna de un test puede indicar el grado en que las relaciones entre los ítems o componentes de un test se organizan acordes al constructo que se espera medir. La premisa, en este caso, es que la validez de los puntajes de una medición puede ser argumentada mediante un análisis de correspondencia entre los resultados de la medición (respuestas observadas) y la cantidad de constructos de interés (dimensiones latentes).

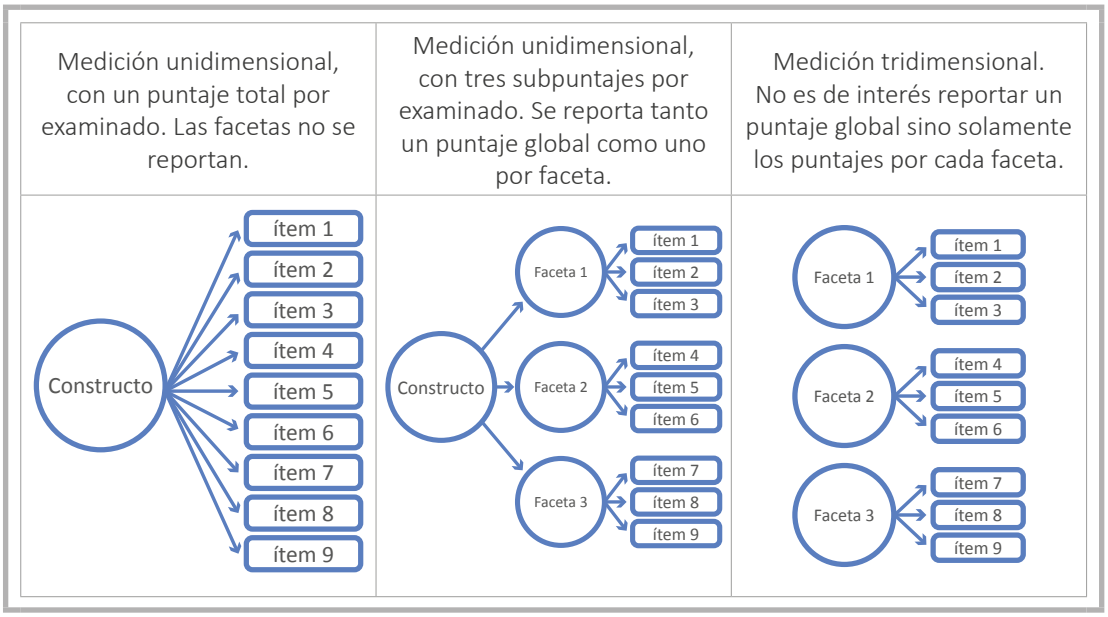
En mediciones estandarizadas a gran escala habitualmente se realizan análisis de dimensiones latentes, o análisis factoriales, que son técnicas de análisis multivariado orientadas a verificar el número y robustez de las variables latentes esperadas, empíricamente, a partir de una matriz de asociaciones entre respuestas a los ítems y tareas de un test. La tarea de esta familia de procedimientos consiste en estimar una estructura subyacente a partir de una matriz de respuestas, mediante una abstracción matemática, y luego determinar si el modelo estadístico resultante es adecuado a la matriz de datos (Brown, 2006).

Para reunir evidencia de este tipo, el primer paso consiste en la especificación del número de constructos a medir, decisión que debe estar alineada con el número de puntajes que se desea reportar a partir de una medición. Al respecto, Haladyna (2016) señala que entre los análisis a realizar para que las interpretaciones de una prueba sean válidas, se encuentran aquellos orientados a determinar si el desempeño o habilidad evaluados consisten en una única dimensión, o si se componen de diversas dimensiones; si un puntaje total es suficiente o si se requieren subpuntajes para llegar a interpretaciones más válidas de la variable medida.

Supongamos que una medición tiene un constructo compuesto por tres facetas. Dependiendo de su definición conceptual y operacional, la estructura podría tomar al menos tres grandes formas (ver figura 2).

FIGURA 2

MODELOS POSIBLES PARA UN MISMO TEST, DEPENDIENDO DE LA ESTRUCTURA QUE SE DESEA REPORTAR E INTERPRETAR



Lo recomendable en este caso sería realizar tantos análisis factoriales como alternativas de medición tenemos, y revisar si alguno de ellos ajusta mejor que otro, o bien, revisar el comportamiento ítem por ítem en caso de que el modelo más apropiado a los usos intencionados no se ajuste adecuadamente.

Ejemplo de evidencia de validez de la estructura interna: Prueba TERCE 2014

En el estudio de factores asociados al logro de aprendizajes del Tercer Estudio Regional Comparativo y Explicativo (TERCE, 2014), un atributo de interés consistía en medir la percepción de seguridad del estudiante en su sala de clases. La escala utilizada para estudiantes de 6° grado fue la siguiente:

¿Algunas de estas cosas te pasan cuando estás en la escuela? (Respuestas Sí/No)

- Mis compañeros me dejan solo.
- Mis compañeros me fuerzan a que haga cosas que yo no quiero hacer.
- Mis compañeros se burlan de mí.
- Temo que uno de mis compañeros me golpee o me haga daño.
- Me siento amenazado por alguno de mis compañeros.
- Tengo miedo de alguno de mis compañeros.

En este estudio, la escala se orienta a medir la sensación de seguridad del estudiante en su sala de clases. Se realizó un análisis comparado de dos estructuras factoriales posibles: una unidimensional y otra bidimensional. En este análisis se puede ver que ambas estructuras cuentan con evidencia a favor, y los índices de ajuste globales resultaron levemente más beneficiosos en la solución de dos factores. Este es un caso en que es una decisión profesional del investigador el reportar una escala o dos escalas por separado, en función del propósito de la medición y la teoría acerca del fenómeno estudiado.

TABLA 3

RESUMEN DE ANÁLISIS FACTORIAL REALIZADO PARA MODELOS UNI Y BIFACTORIALES

Ítem	Factor 1	Factor 2
Mis compañeros me dejan solo.	0.582	NA
Mis compañeros me fuerzan a que haga cosas que yo no quiero hacer.	0.639	NA
Mis compañeros se burlan de mí.	0.711	NA
Temo que uno de mis compañeros me golpee o me haga daño.	0.863	NA
Me siento amenazado por alguno de mis compañeros.	0.883	NA
Tengo miedo de alguno de mis compañeros.	0.868	NA
Ítem	**Factor 1**	**Factor 2**
Mis compañeros me dejan solo.	0.727	-0.005
Mis compañeros me fuerzan a que haga cosas que yo no quiero hacer.	0.452	0.279
Mis compañeros se burlan de mí.	0.545	0.29
Temo que uno de mis compañeros me golpee o me haga daño.	0.002	0.886
Me siento amenazado por alguno de mis compañeros.	0.191	0.746
Tengo miedo de alguno de mis compañeros.	-0.01	0.898

Fuente: elaboración propia.

Evidencia basada en relaciones con otras variables

Los estándares señalan que, en muchos casos, los usos intencionados de un test conllevan de manera implícita una cierta asociación esperada con otras variables o constructos, por tanto, el análisis de estas relaciones -es decir, de las mediciones que entrega el test con otras mediciones externas-, provee una evidencia importante acerca de su validez (AERA, APA y NCME, 2014).

Para reunir este tipo de evidencia, la teoría del constructo de interés permite identificar otras mediciones conceptualmente relacionadas, así como la magnitud y dirección de tales asociaciones. Se puede hablar de **evidencia de validez convergente** cuando se espera que dos mediciones muestren una asociación fuerte, y de **validez discriminante**, cuando se espera que dos mediciones muestren una asociación nula.

Por ejemplo, un test de resistencia al cansancio lector debería correlacionar positivamente con un test que mida capacidad de concentración, y negativamente con una medición de distractibilidad. Ambas serían asociaciones que darían cuenta de validez de tipo convergente, ya que sus resultados van en la misma línea, aunque la escala sea distinta. La correlación no tendría que ser en extremo alta, ya que no se mide exactamente lo mismo.

En cambio, una medición de habilidades escritas no debería correlacionar con una medición de habilidades matemáticas. De existir asociación positiva, esta debería ser pequeña, ya que se trata de dos constructos que, más allá de alguna habilidad académica general, no deberían estar altamente correlacionados. Esto sería evidencia de validez discriminante a favor de la validez de estas mediciones, dado que mostraría atributos diferenciables no solo teórica sino también empíricamente.

Ejemplo de evidencia de validez convergente y discriminante: Pruebas SEPA 2016

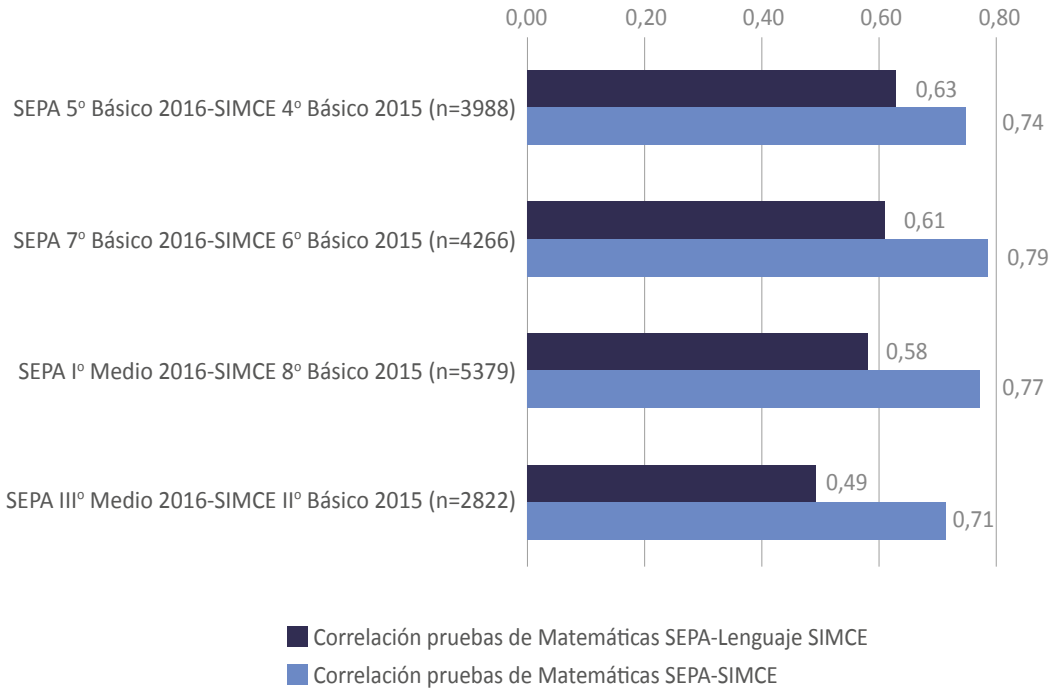
En el caso de las pruebas SEPA, se realizó un estudio correlacional, a nivel de estudiante, con los resultados de las pruebas más importantes a escala nacional: el SIMCE de Lenguaje y Matemáticas.

Dado que SEPA y SIMCE miden el dominio del currículum escolar en Lenguaje y Matemáticas, se espera que en una misma área de aprendizaje se observe una alta asociación entre ambas pruebas (validez convergente). Para estudiar esto, una vez pareados a nivel estudiante los resultados de ambas pruebas, se espera que el coeficiente de correlación entre ambas medidas sea alto y positivo.

En contraposición, se espera que la asociación entre el desempeño en distintas áreas sea menos fuerte que en el caso anterior, aunque positiva, pues el nivel académico general del estudiante explica, al menos en parte, el rendimiento en los distintos sectores de aprendizaje.

FIGURA 3

EVIDENCIA DE VALIDEZ CONVERGENTE Y DISCRIMINANTE PRUEBAS DE MATEMÁTICAS SEPA



Evidencia basada en las consecuencias de un test

Las consecuencias que se derivan del uso de un test dependen de manera directa de la interpretación de sus puntajes para los usos intencionados. En este sentido, el proceso de validación involucra reunir evidencia con el fin de evaluar la sensatez de estas interpretaciones para tales usos (AERA, APA y NCME, 2014). En el caso de mediciones que se usan con la intención de tomar decisiones acerca de personas o grupos de personas, es especialmente relevante cuidar que quienes implementan estas acciones, dispongan de información para minimizar sus posibles efectos negativos de estas (Nkwake, 2015; Taut, Santelices y Stetcher, 2011).

Al realizar este tipo de validación, es necesario contar con una declaración explícita de las consecuencias intencionadas de una medición, lo que es una tarea compleja, dado que, en palabras de Kane (2008), es necesario evaluar las distintas interpretaciones y usos de los puntajes de las pruebas, en tanto sus aplicaciones son diversas y, en ocasiones, fuertemente contextuales.

Ejemplo de estudio de validez consecuencial del Sistema de Evaluación Docente en Chile

El estudio se orientó a recabar las múltiples teorías subyacentes al Sistema de Evaluación Docente chileno. Con la explicitación de estas teorías, se podría contar con una base clara para poder evaluar las consecuencias intencionadas por el sistema, permitiendo a la vez visibilizar las no intencionadas por el mismo. Con tal fin, se analizaron documentos legales y declaraciones oficiales del programa, para luego entrevistar a actores clave en el proceso de diseño e implementación del programa de evaluación.

Los resultados mostraron distintas teorías subyacentes conviviendo en el programa, las que revelaban bastante superposición en efectos intencionados, tales como la movilización de cambios y la instalación de capacidades en los docentes, entre otros.

Fuente: Taut, Santelices, Araya y Manzi (2010).

IV. Imparcialidad: ¿cómo podemos asegurar la imparcialidad de nuestra medición?

Distinguir inequidad educativa de inequidad en una medición

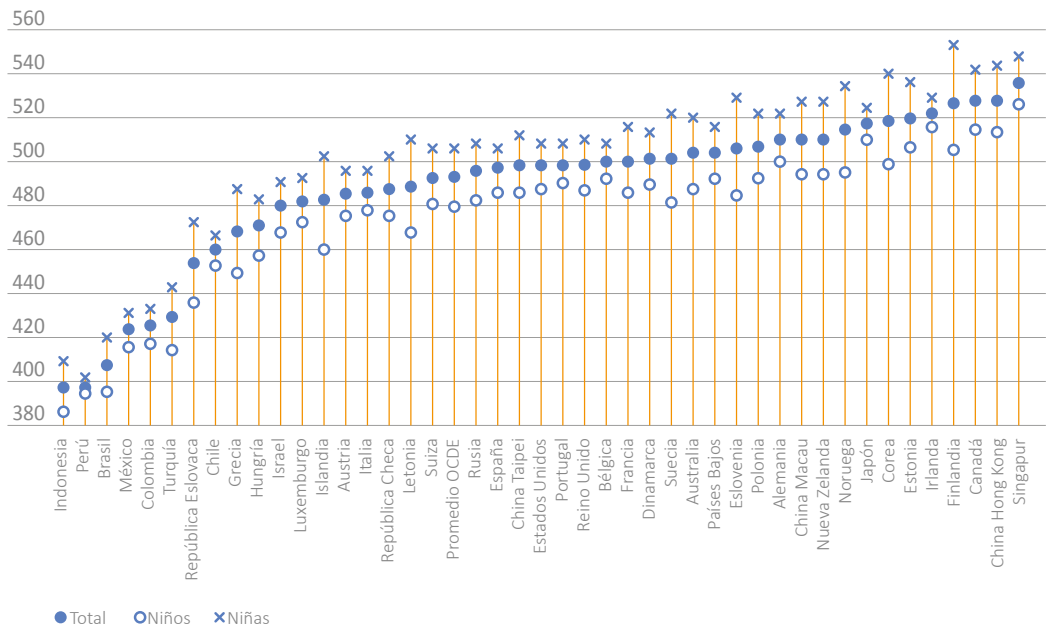
El estudio de la imparcialidad de una medición es especialmente relevante en el contexto de la medición educacional, donde los resultados pueden mostrar brechas importantes entre distintos grupos que participan en la medición, como reflejo de inequidades o diferencias en el acceso a oportunidades educacionales. Esta desigualdad de oportunidades, como problema de orden social, económico y cultural, se traducen en diferencias de desempeño en las mediciones estandarizadas, induciendo la sospecha acerca de si el test refleja tales diferencias, o si la medición contribuye a generarlas.

Una diferencia clásica de rendimiento entre grupos tiene que ver con el desigual rendimiento entre niños y niñas en las mediciones estandarizadas. En el siguiente ejemplo vemos los resultados de hombres y mujeres en todos los países que participaron en PISA 2015.

Ejemplo de resultado de niños y niñas, por país, en la medición de lectura: PISA 2015

FIGURA 4

RESULTADO PROMEDIO SEGÚN GÉNERO PARA CADA PAÍS, RESPECTO DE SU PROMEDIO EN LA MEDICIÓN DE LECTURA DE LA PRUEBA PISA



Fuente: Organización para la Cooperación y el Desarrollo Económico [OCDE], (2018).

Surge la siguiente pregunta: ¿son todas las diferencias entre hombres y mujeres producidas por un sesgo en las pruebas? ¿O son diferencias que provienen de sistemas educacionales y culturales que entregan diferentes oportunidades de aprender a niños y niñas? Esta distinción es relevante para el desarrollador del test, quien debe asegurar que las diferencias entre grupos son genuinas y no provocadas por la prueba.

Formalmente, se habla de sesgo como una forma de invalidez o error sistemático en los puntajes que produce el test para los integrantes de un determinado grupo. Podemos distinguir dos tipos de sesgo en una medición, a nivel de test y a nivel de ítems.

- **Sesgo a nivel de test:** Una prueba es segada para los miembros de un determinado grupo, si en la predicción de un criterio para el que ha sido construido el test, se

observan errores sistemáticos en la predicción para ese grupo. En otras palabras, el test es sesgado si un determinado grupo presenta subpredicción o sobrepredicción (Cleary, 1968).

Ejemplo de evidencia para evaluar equidad en la Prueba de Selección Universitaria en Chile

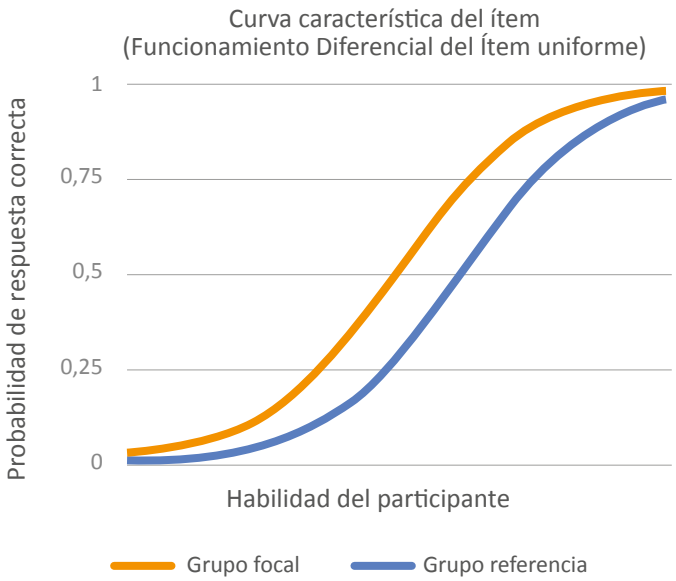
Luego de sus primeros años de implementación, se sometieron a estudio las Pruebas de Selección Universitaria chilenas de Matemáticas, y de Lenguaje y Comunicación, tomando el rendimiento académico de los estudiantes en el primer año de sus estudios universitarios como criterio. Con estos datos se estimaron y compararon indicadores de validez y predicción diferencial según género de los estudiantes. Los resultados, semejantes a lo observado con pruebas de admisión estadounidenses, revelan una leve presencia de validez diferencial, especialmente cuando se considera el género de los estudiantes, evidenciando una consistente pero leve subpredicción del rendimiento de las mujeres.

Fuente: Manzi et al. (2010).

- **Sesgo a nivel de ítems:** Un ítem es sesgado cuando grupos específicos muestran un desempeño diferencial que no se explicaría por su nivel de habilidad en el dominio que se mide, sino por su pertenencia a un grupo específico (Camilli, 2006). En la figura 4 se ilustra un caso donde la probabilidad de respuesta correcta (eje vertical) es mayor para un mismo nivel de habilidad (eje horizontal) en el grupo focal que en el grupo de referencia.

FIGURA 4

CURVA DE PROBABILIDADES DE ACIERTO PARA DOS GRUPOS



Imparcialidad de la medición: aseguramiento en las distintas fases de una medición

Desarrollo: construcción de test e ítems cuidando la imparcialidad de la medición

Los estándares (AERA, APA y NCME, 2014) señalan la especificación del contenido del test como una fuente potencial de sesgo en una medición, es decir, este podría entremezclar la medición del constructo de interés con conocimientos específicos que favorezcan a ciertos subgrupos por sobre otros. Ello podría ocurrir tanto por conocimiento de contenidos específicos, como por aspectos motivacionales respecto de tales contenidos. Esto no implica necesariamente quitarle autenticidad a los ítems y a su contexto, sino abordar el hecho de que estudiantes con distintas características y pertenecientes a distintos contextos responderán una misma prueba.

Los procesos de respuesta a los ítems también pueden estar influenciados por varianza irrelevante para el constructo de interés, en tanto los ítems pueden ser resueltos en formas que no son las intencionadas, o en tanto el formato de respuesta sea menos o más familiar al examinado.

Ejemplo de ítem 5° básico en dos versiones: no contextual y altamente contextual: SEPA (2015)

Ítem muy poco contextual

¿Cómo hacer pegamento casero?

¿Se te acabó el pegamento?, ¿es tarde y ya han cerrado todos los comercios?, ¿estás aburrido y quieres experimentar con pegamentos caseros?, o ¿Tienes que hacer algún trabajo para la escuela y quieres ahorrarte el dinero que te han dado para comprar pegamento? Para esas situaciones u otras, aprende cómo hacer pegamento casero con ingredientes que son fáciles de conseguir y que seguramente encontrarás en la cocina de tu casa.

Materiales: Dos cucharadas de leche en polvo; un cuarto de taza de agua de la llave caliente; una cucharada de vinagre; media cucharada de bicarbonato.

1. Mezcla la leche en polvo y algo de agua caliente (dependiendo de la cantidad de leche en polvo) y revuelve bien.
2. Agrega vinagre a la leche, esto hará que la leche se divida en una parte sólida y, por otra parte, se genera una especie de suero líquido. Sigue revolviendo hasta que esta separación se complete y luego desecha el suero de leche líquido.
3. Luego, asegura una toallita de papel con una banda elástica en el borde de una taza grande y coloca la leche sólida sobre ella. Pon otra toallita de papel sobre la leche y presiona de modo que todo el líquido salga.
4. Ubica la mezcla sólida en otra taza y rómpela en trozos más pequeños.
5. Agrega una cucharada de agua caliente y el bicarbonato. Probablemente se observe algo de espuma debido a la reacción que se produce entre el bicarbonato y el vinagre.
6. Revuelve la mezcla hasta que alcance una buena consistencia, agregando agua en caso de que esté muy espesa.

¿Cuál es el propósito comunicativo del texto anterior?

a) Informar qué hacer en caso de que falte pegamento en la casa.
b) Entretener a través de un experimento sobre pegamento casero.
c) Convencer de que es mejor utilizar pegamento casero que comprado.
d) Dar instrucciones sobre una forma de confeccionar pegamento casero.

Ítem altamente contextual

En función del texto leído en clase, titulado “¿Cómo hacer pegamento casero?”, responde la siguiente pregunta.

¿Cuál es el propósito comunicativo del texto leído?

a) Informar qué hacer en caso de que falte pegamento en la casa.
b) Entretener a través de un experimento sobre pegamento casero.
c) Convencer de que es mejor utilizar pegamento casero que comprado.
d) Dar instrucciones sobre una forma de confeccionar pegamento casero.

Fuente: Centro de Medición MIDE UC (2011).

Como puede verse, en el caso del ítem anterior en su versión poco contextual la forma en que se encuentra expresado permite que el examinado pueda encontrar la respuesta correcta sin recorrer a otra fuente de información que no sea el conocimiento del contenido que está siendo medido. En cambio, en la versión altamente contextual de este ítem, la medición del contenido se encuentra obstaculizada por el recuerdo o memorización de una actividad realizada en clases, pudiendo conducir a error de medición. Este es un ejemplo extremado de un ítem que quizá podría funcionar en una evaluación de aula (a micro escala), pero que no sería admisible en una medición estandarizada, ya que apela a elementos fuera de la prueba, que no son el conocimiento medido, sino información contextual al proceso de enseñanza.

En el desarrollo de rúbricas, es particularmente relevante que el puntaje a asignar se derive del constructo de interés y no de características de la respuesta irrelevantes o tangenciales al constructo, las que pueden pasar inadvertidas en el proceso de desarrollo (AERA, APA y NCME, 2014).

Ejemplo de rúbrica para la calificación de una presentación oral

Instrucción presentada a los estudiantes: Presenta el cuento leído a tus compañeros siguiendo la secuencia de la historia, y cuidando de indicarles quiénes son los personajes principales y secundarios del cuento.

Luego, la presentación es calificada de la siguiente forma. En este caso, el estudiante solamente puede obtener el puntaje completo si va más allá de lo solicitado.

0 puntos	3 puntos	6 puntos	10 puntos
• No cumple con lo mínimo exigido para el nivel inmediatamente superior.	• Sigue correctamente la secuencia de la narración. O bien: • Identifica claramente los personajes principales, y también identifica claramente los personajes secundarios.	• Sigue correctamente la secuencia de la narración. • Identifica claramente los personajes principales y también identifica claramente los personajes secundarios.	• Sigue correctamente la secuencia de la narración. • Identifica claramente los personajes principales y también identifica claramente los personajes secundarios. • Narra el cuento con ritmo y entonación adecuados para facilitar la comprensión de sus compañeros.

En los casos de mediciones que se aplicarán a mediciones aplicadas en diversos idiomas y que, por tanto, deben ser adaptadas localmente para medir de forma adecuada el constructo de interés, surge la problemática por la **invarianza de la medición**. Un instrumento invariante es aquel en que las inferencias acerca de los puntajes de una medición son las mismas con base en instrumentos localmente adaptados (AERA, APA y NCME, 2014), proceso complejo y necesario en cualquier instrumento de medición internacional.

Este problema trasciende a la traducción de los ítems, ya que no se puede afirmar que la traducción no incida en propiedades tales como su dificultad de los ítems o la precisión de la escala. Las traducciones deben mantener no solo el sentido o significado, sino también aspectos culturales que inciden en la familiaridad de los examinados, la que debe ser también interculturalmente similar. Este problema no es únicamente lingüístico, sino implica las imágenes y contextos que se presentan a los examinados, que pueden ser más o menos familiares, no solo a en cuanto al idioma, sino en relación a subgrupos al interior de un país.

Ejemplo de imágenes de balanzas para poner en el contexto de un ítem. ¿Cuál es más familiar para la población de interés de la medición?

Aplicación: importancia de la estandarización de procesos y aseguramiento de acceso universal a la medición para asegurar la imparcialidad

Los estándares (AERA, APA y NCME, 2014) visibilizan cuáles aspectos contextuales del test pueden resultar una fuente de varianza irrelevante, como es la falta de claridad en las instrucciones, o, el uso de claves lingüísticas o culturales específicas como sustento para los enunciados. En los casos en que la medición involucra un contexto interpersonal, la interacción con el examinador también puede ser una fuente de varianza irrelevante, para lo cual deben ser alertados y contar con pautas estandarizadas durante el proceso de medición.

El aseguramiento de un trato comparable entre examinados es relevante como forma de asegurar la imparcialidad de la medición, tanto durante los procesos de administración

del test como de la puntuación. Los primeros deben ser estandarizados al máximo, para evitar que aspectos idiosincrásicos del proceso incidan en el resultado.

Ejemplos de situaciones de examinación para un mismo test

Supongamos que en una medición estandarizada, con aplicadores externos al establecimiento educacional, se olvida instruir a los aplicadores acerca de cómo debe ser abordado el tiempo "de sobra" que puede existir para aquellos estudiantes que finalicen antes del lapso estipulado.

El aplicador de la sala A, cuando advierte que un grupo de alumnos ha terminado antes del tiempo establecido, les indica que pueden ir a recreo.

El aplicador de la sala B, cuando advierte que un grupo de alumnos ha terminado antes del tiempo establecido, les sugiere que revisen sus respuestas, para aprovechar el tiempo disponible.

Como será evidente para el lector, los examinados de la sala A probablemente responderán más rápida (y quizás menos atentamente) la prueba, con tal de terminar antes y recibir un incentivo por finalizar (poder salir a recreo). Esto podría incidir en que sus puntajes sean menores que los de los estudiantes de la sala B, quienes fueron incentivados a revisar sus respuestas y posiblemente detecten errores y los corrijan antes de que se termine el tiempo disponible.

Los procesos de puntuación deben ser también imparciales, lo que se facilita con la automatización de los procesos de digitación y puntuación. Sin embargo, debe existir una adecuada verificación que asegure la fidelidad de estos procesos, así como la seguridad de la información.

Puntuación: estudios de invarianza y sesgo, como forma de alertar posibles problemas de imparcialidad

Cuando existe alerta empírica de sesgo, se puede decir que se está en presencia de una distorsión que afecta la validez de los puntajes, lo cual implica que, en determinadas preguntas, grupos específicos muestran un desempeño diferencial que no se explicaría por su nivel de habilidad en el dominio evaluado (AERA, APA y NCME, 2014; Camilli, 2006). En concreto, se habla de **presencia de funcionamiento diferencial de un ítem (DIF)** cuando dos (o más) grupos de examinados, que poseen un mismo nivel de habilidad, muestran una probabilidad diferente de responder correctamente una determinada pregunta.

La norma actual en mediciones estandarizadas a gran escala es realizar este tipo de verificación mediante un modelamiento de teoría de respuesta al item (IRT), que es especificado con facetas; la faceta de interés suele ser el género del estudiante. Este tipo de modelo permite estimar la dificultad de los ítems considerando un efecto global del género, y así poder identificar si existe una interacción estadística de este (o incidencia sustantivamente relevante) con la probabilidad de acierto. Dicha magnitud se ha expresado

como una interacción de más de 0.3 logits en los estudios internacionales (por ejemplo: Schulz, Ainley y Fraillon, 2011; Fraillon, Schulz, Friedman, Ainley y Gebhardt, 2015, entre otros).

Ejemplo de alerta de funcionamiento diferencial según género del estudiante: ICCS 2009

La tabla 4 muestra los estimadores de comportamiento diferencial de ítems para aquellos que formaron parte del escalamiento de la prueba. Como se puede ver, solamente unos pocos mostraron una limitada evidencia de Funcionamiento Diferencial del Ítem (DIF) (estimadores mayores que 0.3 logits). En general, dado que el comportamiento diferencial según género en los ítems de la prueba ICCS no fue considerado como un problema serio de la medición, se decidió no excluir ítems del escalamiento sobre la base de este análisis.

TABLA 4

ESTIMADORES DE COMPORTAMIENTO DIFERENCIAL SEGÚN GÉNERO EN ÍTEMS PRUEBA ICCS 2009

Ítem	Estimador de DIF según género	Ítem	Estimador de DIF según género	Ítem	Estimador de DIF según género
CI2COM1	0.29	CI2PRM1	-0.07	CI2CEM2	-0.05
CI2MOM1	0.21	CI2CCM1	-0.13	CI2WFO1	0.01
CI2MLM1	0.17	CI2CCM2	-0.02	CI2ORM1	-0.03
CI2MLM2	-0.15	CI2SRM1	0.10	CI2RCM1	-0.05
CI2PDO1	0.17	CI2SRM2	-0.08	CI2PJM1	0.12
CI2RDM2	-0.13	CI2SRM3	0.05	CI2PJM2	0.15
CI2SHM1	-0.12	CI2OMM1	0.07	CI2REM2	0.06
CI2SHM2	0.26	CI2OMM2	0.00	CI2REM3	0.24
CI2TGM1	-0.05	CI2OMM3	0.24	CI101M1	-0.18
CI2TGM2	-0.56	CI2RRO1	0.20	CI109M1	0.05
CI2BPM1	-0.07	CI2DCM1	0.21	CI108M1	-0.20
CI2BPM2	0.20	CI2PFM1	0.05	CI128M1	0.07
CI2GFM1	0.14	CI2PFM2	0.08	CI137M1	0.12
CI2BIO1	0.30	CI2PCM1	-0.02	CI110M1	-0.19
CI2GLM1	-0.11	CI2PCM2	0.19	CI113M1	-0.05
CI2GLM2	0.13	CI2VOM1	-0.02	CI104M1	-0.27
CI2FDM1	0.09	CI2VOM2	0.04	CI115M1	-0.19
CI2FSM1	0.03	CI2VOM3	0.17	CI119M1	-0.13
CI2SCM1	0.00	CI2DLM1	-0.14	CI120M1	0.28
CI2SCM2	-0.14	CI2HRM1	0.04	CI121M1	-0.32
CI2ASM1	-0.32	CI2JOM1	-0.01	CI127M1	-0.48
CI2ASM2	-0.18	CI2WFO2	0.09	CI132M1	0.02
CI2CNM1	-0.03	CI2PGM1	0.12	CI129M1	-0.32
CI2CNM2	0.25	CI2PGM2	0.04	CI130M1	-0.11
CI2ETO1	0.04	CI2ECM1	0.16	CI106M1	0.02
CI2ETM2	-0.24	CI2ECM2	-0.04		

Nota: muestra de calibración internacional: los estimadores bajo -0.3 y sobre 0.3 se encuentran sombreados. Valores negativos muestran DIF a favor de niñas, mientras que valores positivos muestran DIF a favor de los niños.

Consideraciones finales: ideas fuerza

La calidad técnica de una medición se basa en tres pilares fundamentales: la confiabilidad y precisión de un test, la evidencia de su validez para los propósitos específicos de la medición y la imparcialidad que, originalmente concebida como una faceta de la validez, enfatiza la importancia de que una medición muestre brechas que le anteceden, y no que contribuya a generarlas mediante situaciones injustas en las distintas fases de desarrollo, aplicación y puntuación de test.

Los estándares constituyen un referente al cual toda medición debe aspirar y, en cuanto tales, se han concebido como un listado exhaustivo de los temas que deben ser sometidos a un juicio profesional para juzgar la calidad de una medición, y con base en este, priorizados para ser abordados tanto por desarrolladores como por los usuarios de una medición.

La noción de confiabilidad es amplia y la forma específica en que debe ser tratada en una medición tiene relación directa con las fuentes de error propias de cada medición. Una vez cumplido este requisito, es responsabilidad del desarrollador del test recabar evidencia acerca de su validez e imparcialidad, proceso que parte desde su especificación hasta el cálculo de puntajes.

Contar con evidencia de confiabilidad, validez e imparcialidad es de suma relevancia para asegurar que las interpretaciones que se realicen de los resultados entregados por una medición sean adecuados, como también para que pueda ser utilizada conforme a los fines propuestos, lo cual es la esencia de los planteamientos formulados en los éstándares para la medición educativa y psicológica (AERA, APA y NMCE, 2014).

Referencias

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

BARRIOS, S., Urrutia, M., y Catoni, M. I. (2017). Validez de contenido de un banco de ítems en el área de salud del niño. *Educación Médica Superior*, 31(4), 1-9.

BRENNAN, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4), 27-34.

BROWN, T. A. (2006). *Confirmatory factor analysis for applied research*. Nueva York: Guilford Press.

CAMILLI, G. (2006). Test fairness. En R.Brennan (ed.), *Educational Measurement* (pp. 221-256). Westport: American Council on Education and Praeger Publishing.

CENTRO DE MEDICIÓN MIDE UC. (2011). *Preguntas liberadas pruebas SEPA 2009 – 2010*. Recuperado de http://www.sepauc.cl/wp-content/uploads/2014/09/Preguntas-liberadas-2009-2010.pdf

CLEARY, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2), 115-124.

CONTRERAS, J. y Abarzúa, R. A. (2019). Sistema de evaluación de progreso del aprendizaje, SEPA: Evidencia de su confiabilidad y validez. En J. Manzi, M. R. García y S. Taut, (en prensa). *Validez de sistemas de evaluación en Chile y Latinoamérica.* Santiago, Chile: Ediciones UC.

CRONBACH, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

CRONBACH, L. J., y Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3), 391-418.

FRAILLON, J., Schulz, W., Friedman, T., Ainley, J., y Gebhardt, E. (2015). ICILS 2013 Technical Report. Ámsterdam: International Association for the Evaluation of Educational Achievement.

HALADYNA, T. (2016). Item Analysis for Selected-reponse Test Items. En S. Lane, M. Raymond y T. Haladyna (eds.). *Handbook of Test Development*. (2a ed. pp. 392-409). New York: Routledge.

KANE, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.

KANE, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher*, 37(2), 76-82.

KORETZ, D. (2010). *El ABC de la evaluación educativa* (*Measuring Up*). Ciudad de México: CENEVAL.

LANE, S., Raymond, M., Haladyna, T. y Downing, M. (2016). Test development process. En S. Lane, M. Raymond y T. Haladyna (eds.). *Handbook of Test Developpement*. (2a ed., pp. 3-18). Nueva York: Routledge.

LAWLIS, G. F., y Lu, E. (1972). Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin*, 78(1), 17-20.

MANZI, J., Bosch, A., Bravo, D., del Pino, G., Donoso, G., Martínez, M., y Pizarro, R. (2010). Validez diferencial y sesgo en la predictividad de las pruebas de admisión a las universidades chilenas (PSU). *Revista Iberoamericana de Evaluación Educativa*, 3(2), 29-48.

MARTIN, M. O., Mullis, I. V. S., y Hooper, M. (eds.) (2016). *Methods and procedures in* TIMSS *2015*. Recuperado de http://timssandpirls.bc.edu/publications/timss/2015-methods.html

MARTIN, M. O., Mullis, I. V. S., y Hooper, M. (eds.) (2017). M*ethods and procedures in* PIRLS *2016*. Recuperado de https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

MESSICK, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8.

NKWAKE, A. (2015). *Credibility, validity, and assumptions in program evaluation methodology*. Basel: Springer.

ORGANIZACIÓN PARA LA COOPERACIÓN Y EL DESARROLLO ECONÓMICO. (2018). *Reading performance (PISA) (indicator)*. Rescatado de https://data.oecd.org/pisa/reading-performance-pisa.htm

SCHULZ, W., Ainley, J., y Fraillon, J. (2011). ICCS *2009 Technical Report*. Ámsterdam: International Association for the Evaluation of Educational Achievement.

TAUT, S., Santelices, M. V., Araya, C., y Manzi, J. (2010). Theory underlying a national teacher evaluation program. *Evaluation and Program Planning, 33*(4), 477-486.

TAUT, S., Santelices, M. V., y Stecher, B. (2011). *Validation of the Chilean national teacher evaluation system*. Santiago: Sociedad Chilena de Políticas Públicas.

TINSLEY, H. E., y Weiss, D. J. (2000). Interrater reliability and agreement. In *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95-124). San Diego, Academic Press.

WISE, L. y Plake, B. (2016). Test design and development following the standards for educational. En S. Lane, M. Raymond y T. Haladyna (Eds.). *Handbook of Test Development*. (2a ed. pp. 19-39). Nueva York: Routledge.

Instituto Nacional para la Evaluación de la Educación
México

Centro UC
Medición - MIDE