Evaluaciones educativas estandarizadas: desde el marco a los resultados[1]

Standardized educational assessments: from framework to outcomes

Avaliações educacionais padronizadas: do marco conceitual aos resultados

DOI: https://doi.org/10.18861/cied.2025.16.2.4015

Pilar Rodríguez
Centro Universitario Regional del Este,
Universidad de la República
Uruguay
prodriguez@cure.edu.uy
https://orcid.org/0000-0003-1929-4961

Juan Soca
Facultad de Derecho, Universidad de la República
Uruguay
juansoca@gmail.com
https://orcid.org/0000-0001-9083-2932

Mauricio Castillo
Facultad de Derecho, Universidad de la República
Uruguay
castillomega@gmail.com
https://orcid.org/0000-0003-2126-3697

Mario Luzardo
Facultad de Psicología, Universidad de la República
Uruguay
mluzardo@psico.edu.uy
https://orcid.org/0000-0002-9360-2806

Recibido: 30/11/24
Aprobado: 26/03/25

Cómo citar:
Rodríguez, P., Soca, J., Castillo, M., & Luzardo, M. (2025). Evaluaciones educativas estandarizadas: desde el marco a los resultados. *Cuadernos de Investigación Educativa, 16*(2). https://doi.org/10.18861/cied.2025.16.2.4015

Resumen

El artículo analiza las Evaluaciones Educativas Estandarizadas (EEE) como herramientas útiles para medir logros educativos y mejorar la calidad de los sistemas educativos. Se brinda un panorama actualizado sobre su desarrollo, implementación y análisis. La investigación se basa en una revisión teórica de tipo bibliográfico. Se analizaron estudios recientes que abordan contextos educativos globales y regionales. El análisis encuentra avances metodológicos importantes, como la incorporación de la teoría de respuesta al ítem, los test adaptativos informatizados y el uso de inteligencia artificial para la generación de ítems, lo que mejora la precisión y relevancia de las mediciones. Sin embargo, se identificaron limitaciones como la desconexión entre el diseño técnico de las EEE y su aplicación práctica, la falta de formación técnica de los docentes y el escaso aprovechamiento de las bases de datos generadas para la investigación educativa y la toma de decisiones informadas. Los hallazgos subrayan la necesidad de diseñar EEE que consideren contextos desfavorables y estudiantes con discapacidades. Se concluye que para maximizar el impacto de las EEE es crucial fortalecer la capacitación en su interpretación y fomentar su uso en investigaciones educativas, contribuyendo así a sistemas educativos más equitativos y eficaces.

Palabras clave:
evaluación educativa, evaluación a gran escala, evaluación estandarizada, psicometría, metodología.

Abstract

This article analyzes standardized educational assessments (SEA) as valuable tools for measuring educational achievements and improving the quality of educational systems. It offers an updated overview of their development, implementation, and analysis methods. The research is based on a theoretical bibliographic review, examining recent studies that address global and regional educational contexts, with a particular emphasis on Latin America. The analysis identifies significant methodological advances, such as the incorporation of item response theory, computerized adaptive testing, and the use of artificial intelligence for item generation, all of which enhance the accuracy and relevance of measurements. However, limitations were also identified, including the disconnect between the technical design of SEA and their practical application, the lack of technical training for teachers, and the underutilization of the databases generated for educational research and informed decision-making. The findings highlight the need to design SEA that consider disadvantaged contexts and students with disabilities. The conclusion emphasizes that maximizing the impact of SEA requires strengthening training in their interpretation and promoting their use in educational research, thereby contributing to more equitable and effective educational systems.

Keywords:
educational assessment, large-scale assessment, standardized assessment, psychometrics, methodology.

Resumo

O artigo analisa as Avaliações Educacionais Padronizadas (AEPs) como ferramentas úteis para medir resultados e melhorar a qualidade dos sistemas educacionais. Apresenta uma visão atualizada sobre o desenvolvimento, a implementação e as formas de análise dessas avaliações. A pesquisa baseia-se em uma revisão teórica de caráter bibliográfico. Foram analisados estudos recentes que abordam contextos educacionais globais e regionais. A análise identifica avanços metodológicos importantes, como a incorporação da Teoria de Resposta ao Item, os testes adaptativos informatizados e o uso de inteligência artificial para a geração de itens, o que melhora a precisão e a relevância das medições. No entanto, também foram identificadas limitações, como a desconexão entre o desenho técnico das AEPs e sua aplicação prática, a falta de formação técnica dos professores e o aproveitamento limitado das bases de dados geradas para a pesquisa educacional e a tomada de decisões informadas. Os resultados destacam a necessidade de desenhar AEPs que considerem contextos desfavoráveis e estudantes com deficiências. Conclui-se que, para maximizar o impacto das AEPs, é crucial fortalecer a capacitação em sua interpretação e fomentar seu uso em pesquisas na área da educação, contribuindo assim para sistemas mais equitativos e eficazes.

Palavras-chave:
avaliação educacional, avaliação em larga escala, avaliação padronizada, psicometria, metodologia.

Introducción

La evaluación se ha desarrollado como un instrumento que provee información relevante para la valoración de la consecución de los objetivos planteados. Por este motivo, los sistemas educativos de cada país se han autoimpuesto la tarea de evaluarse. De esta forma nace el desarrollo de las evaluaciones estandarizadas o también llamadas evaluaciones a gran escala.

Estas evaluaciones surgen para responder a dos grandes necesidades: rendir cuentas sobre el desempeño educativo y seleccionar candidatos para el ingreso a la universidad. Sin embargo, con el tiempo, su enfoque ha evolucionado hacia el suministro de información útil para la mejora de la educación. Este artículo busca proporcionar un panorama actualizado sobre las Evaluaciones Educativas Estandarizadas (EEE), abordando su evolución, los avances metodológicos recientes y las proyecciones futuras.

Antecedentes

Las EEE han evolucionado significativamente, desde la evaluación de aprendizajes hacia la valoración de programas, centros educativos y prácticas docentes. Su propósito central es proporcionar información válida, fiable y comparable sobre logros de aprendizaje o eficacia de programas y currículos. Esto implica la sistematización de instrumentos, criterios de corrección y análisis de información, permitiendo la comparación de los sistemas educativos (Jornet, 2017).

En Estados Unidos y Europa, las EEE están consolidadas, mientras que en Latinoamérica han enfrentado debates y resistencias. A pesar de esto, representan una actividad científico-técnica orientada a mejorar los sistemas educativos (Soca, 2018). Aún presentan desafíos metodológicos que han impulsado avances en psicometría, como la teoría de respuesta al ítem (Carlson & von Davier, 2013).

Las EEE permiten oportunidades de investigación únicas (Teig & Steinmann, 2023). Esto ha producido un gran desarrollo de investigación educativa utilizando las bases de datos de evaluaciones como Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), Programme for International Student Assessment (PISA) entre otras (Rutkowski, *et al.*, 2014).

Estados Unidos lideró las EEE a principios de los años 60, con el First International Mathematics Study (FIMS) conducido por la International Association for the Evaluation of Educational Achievement (IEA) (Rutkowski *et al.*, 2014). Para la década del 90 casi todos los estados contaban con un sistema de evaluación a gran escala.

El Laboratorio Latinoamericano para la Evaluación de la Calidad de la Educación (LLECE) fue pionero en EEE en nuestro continente. El LLECE realizó su Primer Estudio Regional Comparativo y Explicativo (PERCE) en 13 países en 1997 (Wagemaker, 2014).

Popham (1999) identificó dos usos principales para las EEE: rendición de cuentas y mejora educativa. Lamentaba que predominara el enfoque centrado en la

rendición de cuentas, lo que generó mala fama a las evaluaciones y presiones políticas para justificar el gasto educativo. Pérez Juste (2006) subrayó la necesidad de que los educadores comprendan los conceptos básicos de las EEE para maximizar su impacto.

El desarrollo de las EEE ha llevado a la creación de estándares internacionales, como los propuestos por el Joint Committee on Standards for Educational Evaluation (JCSEE), que se organizan en cinco criterios: utilidad, viabilidad, honradez, precisión y responsabilidad (JCSEE, 2010).

Metodología

Se propone una revisión teórica actualizada sobre los principales procesos involucrados en el desarrollo de las EEE, ejemplificándolos con casos recientes y contextualizados. La investigación bibliográfica incluyó una exploración en bases de datos como Science Direct y Education Resources Information Center (ERIC).

Proceso de selección de fuentes

Para garantizar un enfoque riguroso y representativo, se implementó un proceso de búsqueda estructurado en tres fases:

1. **Definición de términos clave:** Se utilizaron palabras clave en inglés y español, tales como *"large scale assessment", "evaluation methods", "educational assessment".*
2. **Filtrado temporal:** Se estableció un criterio de inclusión basado en la actualidad de las publicaciones, limitando la búsqueda a publicaciones de los últimos cinco años (2018-2023). En los casos en que no existían referencias recientes se incluyeron fuentes anteriores.
3. **Selección de artículos relevantes:** A partir de la búsqueda inicial, se obtuvieron 312 artículos y libros. Posteriormente, se aplicaron criterios de selección y eliminación detallados a continuación.

Criterios de inclusión y exclusión

Para asegurar la calidad y pertinencia de las fuentes seleccionadas, se establecieron los siguientes criterios:

Criterios de inclusión:

- Estudios que analicen la evaluación educativa a gran escala desde enfoques teóricos, metodológicos o comparativos.
- Investigaciones que empleen técnicas psicométricas avanzadas.
- Documentos técnicos de organismos internacionales y nacionales.
- Publicaciones en revistas indexadas que incluyan análisis empíricos con muestras representativas.

Criterios de exclusión:

- Investigaciones centradas exclusivamente en contextos alejados del ámbito latinoamericano.
- Estudios excesivamente técnicos.
- Artículos de opinión o editoriales sin respaldo empírico.

Luego de aplicar estos filtros, se seleccionaron 78 estudios y libros. Adicionalmente, se incorporaron referencias complementarias para describir los procesos de desarrollo de diferentes métodos y técnicas. También se consultaron los sitios web de los institutos de evaluación de América Latina y de organizaciones que desarrollan EEE a nivel internacional, lo que permitió incluir reportes técnicos relevantes. En total, el análisis se basó en 91 publicaciones.

El proceso de una evaluación educativa estandarizada

El marco teórico

El marco de la evaluación debe reflejar el constructo a medir. Si bien es una construcción teórica a partir de la producción escrita de referentes en el área, no se debe soslayar el aporte de los implicados y otros referentes en la temática. Por eso, es importante contemplar un proceso de validación del marco teórico, donde se someta el producto elaborado a la opinión de comités de expertos y grupos de discusión con los implicados. Este proceso está descrito en los marcos utilizados por el Instituto Nacional de Evaluación Educativa (INEEd).

En el ámbito educativo, las variables a medir, llamadas variables latentes o constructos, no son directamente observables. Se trata de atributos que se reflejan en el desempeño en las pruebas. La precisión de la medición adquiere gran relevancia, ya que el significado de una medida depende de una teoría del constructo y de las hipótesis sobre lo que se está midiendo. Los instrumentos de evaluación actúan como puente entre la teoría y la observación, generando puntuaciones que representan estos constructos (Jackson Stenner *et al.*, 2022).

La operacionalización de las tablas de dimensiones

Luego de elaborado el marco, es necesario conectar los conceptos teóricos con términos medibles, proceso denominado operacionalización, que es la construcción de una matriz de conceptos y dimensiones para organizar la información sobre los aspectos a evaluar.

La definición operativa se basa en el marco teórico, pero también requiere la consulta a expertos, utilizando técnicas cualitativas como grupos de discusión o paneles delphi (Muñiz & Fonseca-Pedrero, 2019). Este proceso, fundamental para la validación de contenido, asegura que los constructos teóricos estén correctamente representados en el instrumento. En pruebas educativas, esta definición se plasma en una tabla de

doble entrada, donde las filas representan las áreas de contenido y las columnas los procesos cognitivos necesarios para resolver las tareas (Abad *et al.*, 2011).

En los documentos de diversos institutos de evaluación se pueden observar estas tablas. Se brindan dos ejemplos: uno internacional y otro nacional. El National Assessment of Educational Program (NAEP), que realiza evaluaciones del logro académico de los estudiantes de Estados Unidos desde preescolar a bachillerato, publica los marcos teóricos y tablas de especificaciones para cada dominio (National Assessment Governing Board, 2022). En el informe de resultados de la evaluación nacional de logros realizada por el INEEd se da cuenta de las tablas de especificaciones (INEEd, 2017).

En los sistemas educativos, los acuerdos sobre lo que los estudiantes deben saber y poder hacer en cada dimensión y nivel se definen como estándares de contenido, objetivos de aprendizaje o metas de aprendizaje.

El Grupo de Trabajo sobre Estándares y Evaluación del Programa de Promoción de Reforma Educativa en América Latina y el Caribe (PREAL) recomendó, a comienzos del nuevo milenio, que los gobiernos establecieran estándares educacionales y desarrollaran pruebas para medir los resultados (Ferrer, 2006). Actualmente todos los países, incluido Uruguay, se encuentran trabajando para el establecimiento de metas de aprendizaje.

El diseño de los instrumentos

En las EEE se emplean diversos instrumentos para recopilar información, como cuestionarios, pruebas o test, y pautas de observación. Los cuestionarios se utilizan frecuentemente para evaluar habilidades socioemocionales, patrones de comportamiento y características psicológicas, que posibilitan un mejor desempeño educativo (Blair & Razza, 2007). Por otro lado, las pruebas o test son los más comunes para medir dominios cognitivos (Dehaene, 2019).

El diseño de estos instrumentos se basa en la tabla de especificaciones, asegurando que cada dimensión y subdimensión esté representada mediante preguntas, tareas o afirmaciones, conocidas como ítems. El conjunto de ítems conforma un banco, que debe seguir principios básicos de representatividad, relevancia, diversidad, claridad y comprensibilidad (Muñiz & Fonseca-Pedrero, 2019).

La redacción de ítems es fundamental, ya que la violación de lineamientos puede tener efectos negativos en la evaluación. Por eso, se recomienda seguir directrices estrictas de escritura, entrenar a los redactores y dedicar jornadas a la revisión de los ítems (Haladyna & Rodríguez, 2013).

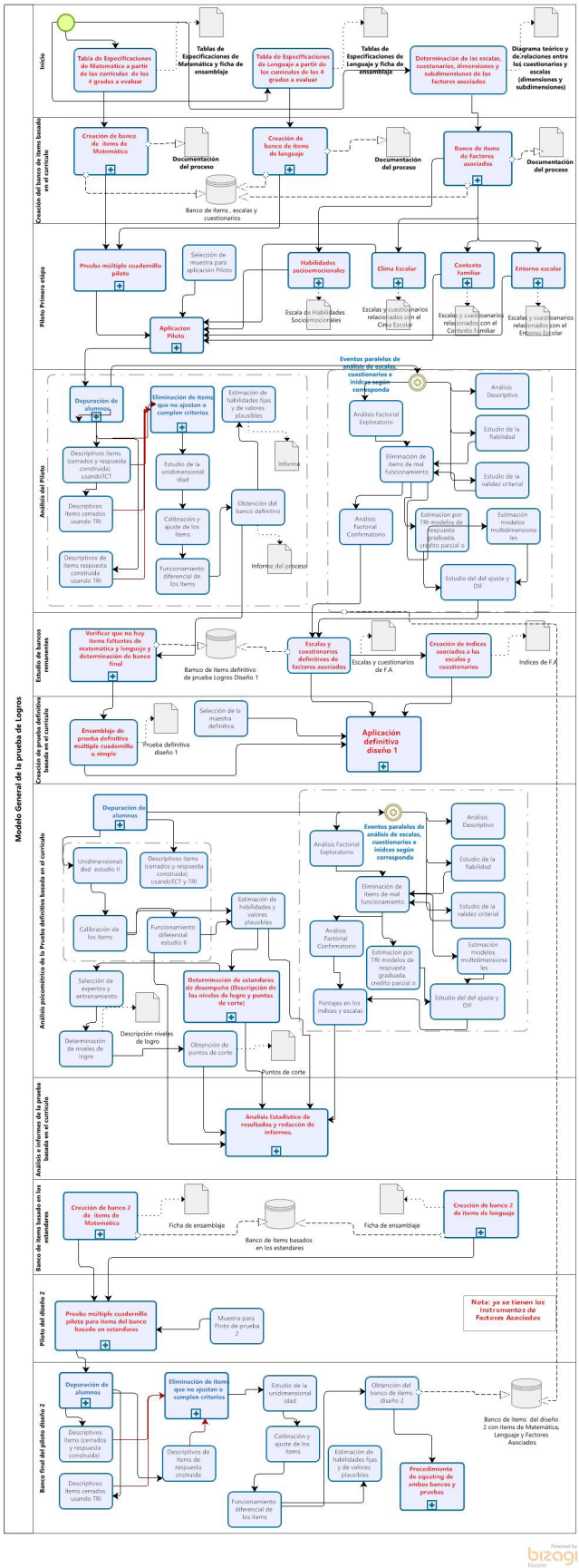
Un desafío en la generación de ítems radica en la capacidad de representar constructos complejos. Dado que los constructos evaluados son de naturaleza mental, el ítem debe evocar una respuesta que permita medir de manera válida y confiable el constructo (Gierl *et al.*, 2021). Esto plantea una tensión entre la viabilidad operativa de la evaluación y la fidelidad con que se representa el constructo a medir.

Como la generación de ítems supone un esfuerzo considerable, se han comenzado a utilizar herramientas como la inteligencia artificial generativa y los grandes modelos de lenguaje (LLM) (Gierl & Haladyna, 2013), como el Psychometric Item Generator (PIG) usando GPT-2 para generar ítems (Götz *et al.*, 2023). También integran modelos

cognitivos para resolver los problemas evaluados (Falcão *et al.*, 2023). La Figura 1 muestra el proceso completo de generación de ítems.

Figura 1

Proceso de creación de ítems de logros en matemática y lenguaje en un país latinoamericano.



El ensamblaje de la prueba

Conformado el banco de ítems, se procede al ensamblaje de la prueba, que implica seleccionar y ordenar los ítems para su versión definitiva. En una ficha de ensamblaje se especifican los pesos porcentuales asignados a cada dimensión de la tabla de especificaciones. Detalles sobre este proceso pueden consultarse en Mahias Finger & Polloni Erazo (2019). Ejemplos específicos se encuentran en Wry & Mullis (2023) para la prueba de lectura de PIRLS y en Rodríguez Morales (2017) para pruebas diagnósticas de matemática al ingreso a la universidad.

En EEE es habitual diseñar cuadernillos múltiples para cubrir los diferentes dominios de la tabla de especificaciones de manera muestral, considerando factores como la extensión de la prueba, el tiempo disponible para los estudiantes, los efectos de posición de los ítems y el contexto de presentación. Además, es necesario garantizar la futura equiparación de puntuaciones entre cuadernillos (Fernández Alonso & Muñiz Fernández, 2011).

El diseño de cuadernillos generalmente utiliza técnicas de diseño experimental, como los diseños de bloques balanceados incompletos (National Assessment of Educational Progress, 2023). Los factores de bloque suelen incluir el cuadernillo y la posición de los ítems, eliminando sesgos relacionados con estos aspectos. Esta metodología asegura que los ítems se distribuyan de forma aleatoria en diferentes posiciones y cuadernillos. Ejemplos de esta técnica pueden observarse en evaluaciones nacionales (Ministerio de Educación del Perú, 2024) y transnacionales como PIRLS (Martin *et al.*, 2015).

El ensamblaje incluye tanto bloques comunes, conocidos como bloques de anclaje, como bloques no comunes, que aparecen en algunos cuadernillos. A través de técnicas de equiparación, estos bloques permiten obtener puntuaciones en una misma métrica, garantizando comparabilidad entre cuadernillos y versiones de la prueba.

Paradigmas de análisis

Se pueden usar dos paradigmas de análisis complementarios: la Teoría Clásica de los Test (TCT) y la Teoría de Respuesta al Ítem (TRI).

Teoría Clásica de los Test

La TCT asume que las puntuaciones observadas son la suma de una puntuación verdadera y un error aleatorio (Muñiz, 2018). En el contexto de una EEE, los análisis basados en la TCT son esenciales para evaluar la calidad del banco de ítems, depurarlos y validar las pruebas. Los análisis que deben considerarse incluyen:

a. **Dificultad de los ítems:** Se define como la proporción de respuestas correctas sobre el total. Ítems con índices menores a 0.1 (muy difíciles) o mayores a 0.9 (muy fáciles) se eliminan. Este índice depende de la muestra evaluada, lo que limita su aplicabilidad (Muñiz, 2018).

b. **Discriminación de los ítems:** Mide la capacidad para diferenciar entre participantes con distintos niveles de desempeño. Se analizan las proporciones de respuestas correctas en dos grupos extremos y la correlación entre cada ítem y el puntaje total del test. Para ítems dicotómicos que miden una variable continua, se utiliza la correlación biserial puntual, eliminando ítems con correlaciones menores a 0.1 y revisando aquellos con valores entre 0.1 y 0.2 (Crocker & Algina, 2008). Cuando los ítems reflejan una variable normal dicotomizada, se usa la correlación biserial. Si ítems y pruebas son dicotómicos, se emplea el coeficiente phi o la correlación tetracórica. Finalmente, corresponde hallar el índice de validez del ítem que mide su correlación con un criterio externo, reflejando su relación con el constructo evaluado (Muñiz, 2018).

c. **Distractores:** Se analiza la distribución de las alternativas incorrectas (distractores), evaluando qué proporción de estudiantes elige cada una. La opción correcta debe ser preferida por quienes tienen puntajes altos, y los distractores, por quienes tienen puntajes bajos. Los test con distractores cuya correlación biserial puntual supere 0.1 deben ser reelaborados si cumplen con los demás criterios esperados (Muñiz, 2018).

d. **Incidencia del ítem en la fiabilidad de la escala:** Se analiza el impacto del ítem en el alfa de Cronbach observando cómo impacta al eliminarlo. Si el coeficiente aumenta, el ítem se revisa y puede ser reelaborado.

e. **Tiempos de respuesta:** El análisis de respuestas con bajo esfuerzo implica evaluar los tiempos por ítem para definir umbrales y establecer criterios de eliminación de alumnos o ítems. La suposición de que los estudiantes siempre realizan un esfuerzo completo (AERA *et al*., 2014) es frecuentemente violada. Este efecto varía entre subgrupos, distorsionando las brechas de rendimiento y afectando especialmente a estudiantes que se están desconectando de la escuela (Soland, 2018, Soland & Kuhfeld, 2019). Este problema puede detectarse analizando los tiempos de respuesta. Métodos como el umbral normativo (Wise & Ma, 2012), la proporción acumulada (Guo *et al*., 2016) y la mezcla de log-normales (Guo & Ríos, 2020) son los más comunes.

Teoría de Respuesta al Ítem

La TRI analiza cada ítem individualmente, a diferencia de la TCT, que evalúa el test como un todo. Se basa en la probabilidad que un individuo acierte un ítem condicionada a su habilidad, llamada curva característica del ítem (CCI) (Crocker & Algina, 2008, Van der Linden, 2018). Rasch desarrolló un primer modelo que considera la habilidad del sujeto y la dificultad del ítem (von Davier, 2016). Posteriormente, se introdujeron modelos de dos y tres parámetros que ampliaron las capacidades de la TRI (Van der Linden, 2018).

La TRI se ha utilizado ampliamente en EEE, como en el Graduate Record Examination (GRE) y el Scholastic Assessment Test (SAT) en Estados Unidos. En Latinoamérica, destaca su uso en las pruebas del LLECE para las ediciones del ERCE (OREALC/UNESCO, 2016). En México, se aplicó en las pruebas Planea (Plan Nacional para la Evaluación de los Aprendizajes) (Instituto Nacional para la Evaluación de la Educación, 2019); en Colombia, en las pruebas SABER del ICFES (Instituto Colombiano para la

Evaluación de la Educación) (ICFES, 2011); en Chile, en el Sistema de Medición de la Calidad de la Educación (SIMCE) (Agencia de Calidad de la Educación, 2014); en Perú, en las evaluaciones del Ministerio de Educación del Perú (2024); y en Brasil, en las pruebas nacionales del Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2024). En Uruguay, utilizan TRI las EEE del Sistema de Evaluación de Aprendizajes (SEA) de la Administración Nacional de Educación Pública (ANEP) y en la Evaluación Nacional de Logros (Aristas) en todos sus componentes (INEEd, 2020).

Adicionalmente, la TRI es la base para los test adaptativos computarizados (TAI), donde los ítems se seleccionan en función del nivel de habilidad estimado del examinado, lo que permite optimizar la evaluación, mejorando la precisión y reduciendo el número de preguntas necesarias (van der Linden & Glas, 2000, Olea & Ponsoda, 2013). En Uruguay son adaptativas las pruebas de matemática y lengua realizadas por la División de Investigación, Evaluación y Estadística (DIEE) de ANEP, la Prueba Nacional Adaptativa de Inglés desarrollada por Ceibal en Inglés y la prueba diagnóstica de matemática al ingreso de la universidad (Rodríguez *et al.*, 2017) aplicadas en la plataforma SEA.

Aunque frecuentemente se usan ítems dicotómicos, los ítems politómicos son útiles en testlets, cuestionarios de habilidades no cognitivas y rúbricas de corrección. Los modelos multidimensionales han ganado relevancia al considerar que los ítems pueden depender de múltiples rasgos latentes (Reckase, 2009). Una aproximación a la TRI para comprender su utilidad en la construcción de escalas se puede leer en Hidalgo-Montesinos & French (2016).

Los principales modelos de TRI son:

1. El modelo de Rasch, que considera solo la dificultad del ítem para calcular la probabilidad de acierto, asumiendo igual discriminación para todos los ítems. La dificultad corresponde al nivel de habilidad que tiene una probabilidad 0.5 de acierto (Bond & Fox, 2013, von Davier, 2016). Proporciona una escala logarítmica (logits) común para la habilidad y la dificultad, facilitando comparaciones precisas entre sujetos e ítems (Linacre, 2012).
2. El modelo de 2 parámetros, que incluye como parámetros la dificultad y la discriminación, es más flexible y mejora su ajuste. Ítems con alta discriminación distinguen mejor entre habilidades cercanas a su dificultad, siendo útil en evaluaciones con ítems no uniformes (de Ayala, 2009).
3. El modelo de 3 parámetros añade el pseudoazar. Es adecuado para pruebas de opción múltiple donde puede haber adivinación y donde los distractores también evalúan la habilidad del participante (van der Linden, 2016).
4. Los modelos para ítems politómicos son útiles en evaluaciones que complementan las pruebas cognitivas, como cuestionarios de contexto, rúbricas y testlets. Permiten analizar categorías intermedias de respuesta y mejorar la validez al captar matices en las respuestas, como el modelo de respuesta graduada (Samejima, 2016) el modelo de crédito parcial (Masters, 2016) y el modelo nominal (Thissen & Cai, 2016).
5. Los modelos no paramétricos son una alternativa a los modelos tradicionales. No asumen formas específicas para las CCI, estimándolas directamente desde los datos con métodos flexibles y sin imponer restricciones sobre la forma (Sijstma & Molenaar, 2016, Xu & Douglas, 2006).

6. Los modelos multidimensionales analizan ítems influenciados por múltiples habilidades. En los modelos compensatorios, la probabilidad de acierto combina linealmente las habilidades, reconociendo que muchos ítems evalúan más de un rasgo (Swaminathan & Rogers, 2016). Ejemplos incluyen el modelo 2PL multidimensional (Reckase, 2016), y los modelos no paramétricos multidimensionales (Luzardo & Rodríguez, 2015), y el modelo isótono (Luzardo, 2019).

Luego de seleccionado el modelo, deben calibrarse los ítems mediante métodos como máxima verosimilitud, bayesianos o no paramétricos y evaluar su ajuste a los datos. En pruebas de múltiple cuadernillo, se requiere equiparar los ítems usando calibración concurrente usando ítems de anclaje u otras metodologías. Debe verificarse la invarianza y detectar sesgos o funcionamiento diferencial de los ítems (DIF) para evitar favorecer subgrupos, como género, etnia o región, utilizándose Mantel-Haenszel, índice de estandarización, área bajo la curva o test de hipótesis sobre los parámetros del ítem.

Es fundamental evaluar la calidad del banco de ítems, asegurando que los parámetros de dificultad cubran todo el rango de habilidades y que la información sea alta, garantizando precisión en la estimación.

Ambos paradigmas de análisis presentan limitaciones importantes que deben considerarse. La TCT depende de las características de la muestra, lo que restringe su capacidad para generalizar resultados más allá de la población evaluada. La TRI requiere tamaños muestrales considerables y supuestos estadísticos estrictos que no siempre se cumplen en todos los contextos educativos.

Fiabilidad de los instrumentos

La fiabilidad mide la consistencia y precisión de un instrumento para reflejar lo que pretende evaluar. Se define como la correlación entre la puntuación empírica y la verdadera, con valores entre 0 y 1 (Muñiz, 2018). En educación, garantiza que las puntuaciones reflejen principalmente habilidades o conocimientos, minimizando el impacto de errores como el ambiente, el estado emocional o la ambigüedad de los ítems.

Con la finalidad de estudiar la fiabilidad se utilizan los siguientes coeficientes:

a. El test-retest mide la estabilidad de una prueba al aplicarla dos veces a los mismos sujetos, calculando la correlación entre las puntuaciones. Un coeficiente mayor a 0.6 es aceptable.

b. El coeficiente alfa de Cronbach mide la consistencia interna del test bajo el principio de tau-equivalencia, siendo una cota inferior de la fiabilidad. Depende del número de ítems, las opciones de respuesta y la varianza total, considerándose adecuado un valor mayor a 0.7.

c. El coeficiente omega de McDonald, con valores aceptables entre 0.7 y 0.9, es una alternativa moderna al alfa de Cronbach. Basado en cargas factoriales, es más estable y representativo de la fiabilidad real, sin depender del número de ítems (McDonald, 1999, Loken & Gelman, 2017, Flora, 2020).

d. El coeficiente de correlación intraclase (CCI) mide el acuerdo entre mediciones repetidas. Valores cercanos a 1 indican alta fiabilidad. Según el diseño, puede analizar mediciones individuales, promedios o modelos de efectos (Correa-Rojas, 2021).

Validez de los instrumentos

La validez evalúa si un instrumento mide lo que afirma medir y si las interpretaciones de sus resultados son apropiadas (AERA *et al.*, 2014; Muñiz, 2018). Garantiza que el test evalúe el constructo de interés y no características no deseadas.

La literatura reciente identifica cinco fuentes de evidencia para validar instrumentos de evaluación (Abad *et al.*, 2011, Muñiz, 2018, Sireci & Benítez, 2023):

- Evidencia basada en el contenido: analiza si los ítems representan adecuadamente el constructo a medir, generalmente a través de juicios de expertos (Beck, 2020, Reynolds & Moncaleano, 2021).
- Evidencia basada en la estructura interna: verifica si la estructura del test (factores o dimensiones) coincide con el constructo teórico subyacente, utilizando análisis factorial exploratorio y confirmatorio (Ferrando *et al.*, 2022).
- Evidencia basada en la relación con otras variables: evalúa si las puntuaciones del instrumento se relacionan con otras medidas según lo esperado teóricamente, a través de la validez convergente, validez discriminante y validez predictiva . Se usan métodos como correlaciones, regresión lineal y curvas ROC (Muñiz, 2018, Abad *et al.*, 2011).
- Evidencia basada en el proceso de respuesta: examina si los procesos cognitivos o conductuales al responder son consistentes con la teoría subyacente. Se utilizan entrevistas cognitivas y observación directa de los participantes (Engelhardt & Goldhammer, 2019, Lee & Winke, 2018).
- Evidencia basada en las consecuencias del uso del test: analiza los efectos, previstos o no, del instrumento en su contexto de aplicación, utilizando técnicas cualitativas, como entrevistas cognitivas, grupos focales y consultas a tomadores de decisiones, o cuantitativas, como análisis de impacto adverso y curvas de Pareto para ponderar puntuaciones en pruebas de selección (Dumas *et al.*, 2022).

Cómo interpretar los resultados

El establecimiento de estándares de desempeño

Las pruebas cognitivas deben fundamentarse en los estándares de contenido. Estos estándares son establecidos por las autoridades educativas competentes de cada país y describen lo que los estudiantes deben saber y ser capaces de hacer para alcanzar determinados niveles de competencia.

Por otro lado, los estándares de desempeño se definen como descripciones del nivel de logro alcanzado por los estudiantes en diferentes categorías de desempeño. Cada

categoría provee información sobre los conocimientos o habilidades que integran el nivel de desempeño (Cizek *et al.*, 2004).

Estos estándares son una herramienta clave para reportar el desempeño de grupos de estudiantes y proporcionar información sobre su progreso. Pueden ser elaborados al mismo tiempo que se desarrollan los estándares de contenido o pueden desarrollarse por el equipo de expertos que determine los puntos de corte.

Diferentes ejemplos para pruebas internacionales se encuentran en Cizek & Bunch (2007), Jornet & González (2009) y Tourón (2009). Un ejemplo adaptado al currículo de Uruguay es presentado en los informes del INEEd (2018, 2021).

En la Tabla 1, donde se muestra la clasificación de los niveles de desempeño en distintas pruebas, se refleja la heterogeneidad en la denominación y estructuración de los niveles de desempeño según cada prueba.

Tabla 1

Niveles de desempeño en distintas pruebas

Prueba	Fuente	Niveles de desempeño
K-12 Achievement Testing Programs	National Assessment of Educational Progress (NAEP)	Por debajo del nivel básico, básico, competente, avanzado.
Programme for International Student Assessment (PISA)	Programme for International Student Assessment	6 niveles de tipo numérico.
Advanced Placement (AP) Examinations	College Board	No aprobado, posiblemente calificado, calificado, bien calificado, extremadamente bien calificado.
Achievement Test Ohio State	Departamento de Educación del Estado de Ohio	Limitado, básico, competente, acelerado, avanzado.
Examen para la Calidad y el Logro Educativo (EXCALE)	Instituto Nacional para la Evaluación de la Educación (INEE), México	Por debajo del nivel básico, básico, medio, avanzado.
Pruebas Saber	Instituto Colombiano para la Evaluación de la Educación (ICFES)	Insuficiente, mínimo, satisfactorio, avanzado.
Pruebas ERCE	UNESCO, LLECE	Nivel I, Nivel II, Nivel III, Nivel IV.
Evaluación Nacional de Logros Educativos (Aristas)	Instituto Nacional de Evaluación Educativa (INEEd), Uruguay	Nivel 1, Nivel 2, Nivel 3, Nivel 4, Nivel 5.

Nota. Adaptado de Jornet & González (2009), Linn (2003), ICFES (2011), INEE (2004), UNESCO-OREALC (2016), INEEd (2021).

Métodos para el establecimiento de estándares de desempeño

El establecimiento de estándares de desempeño o puntos de corte es fundamental para clasificar a los estudiantes según su nivel en pruebas cognitivas. Aunque existen numerosos métodos, no hay consenso sobre cuál es el mejor, ya que cada enfoque

tiene ventajas y limitaciones dependiendo del contexto (Linn, 2003). Estos métodos pueden clasificarse en tres categorías principales: centrados en el test, centrados en las personas y de compromiso.

Métodos centrados en el test

Estos procedimientos para fijar el punto de corte o establecer el estándar de desempeño se basan en las valoraciones de los jueces sobre los ítems del test, por eso también se los denomina métodos valorativos. Entre estos métodos se encuentran el de Nedelsky, el de Ebel, el del consenso directo, el de correspondencia con el ítem descriptor y el de Angoff entre los clásicos.

El Método del Marcador (Bookmark), que ordena los ítems por dificultad en un librillo, permite que los jueces coloquen marcas para definir el punto de corte, de ahí su nombre (Mitzel *et al.*, 2001). Es ampliamente usado por su flexibilidad, su adaptación a evaluaciones complejas, por facilitar la tarea de los jueces y fundamentarse en la TRI (Cizek & Bunch, 2007). Por estas razones es el método por excelencia utilizado en las EEE.

El método de García *et al.* (2013) construye un banco de ítems basado en estándares, estima las CCI promedio y conjuntas para los ítems en el mismo nivel de desempeño y sobre esta base calcula los puntos de corte. Este enfoque supera limitaciones de métodos como Bookmark donde se depende más del juicio de expertos y de las propiedades empíricas del test (North & Jones, 2009).

El método Cloud Delphi ponderado integra el método anterior con el modelo de nube normal que relaciona conceptos cualitativos con datos cuantitativos mediante probabilidad y lógica difusa mejorando la precisión y validez en la clasificación del desempeño estudiantil (Rodríguez & Luzardo, 2019). El método se aplicó en una prueba diagnóstica universitaria que evalúa lectura y matemática (Rodríguez, 2017) y en las pruebas nacionales de logro (Aristas) del INEEd (2018).

Métodos centrados en las personas

Estos métodos se centran en el juicio de los expertos sobre las competencias de los estudiantes. Entre ellos se encuentran los métodos de los grupos contrastantes y del grupo límite y los métodos holísticos (Cizek & Bunch, 2007).

Métodos de compromiso

En estos métodos se incluyen, además del juicio absoluto, la información relativa al grupo, para llegar a un compromiso entre ambos datos. Se destacan los métodos de Hofstee y Beuk (Muñiz, 2018).

Si bien los métodos de establecimiento de estándares de desempeño han avanzado en precisión, continúan enfrentando desafíos metodológicos significativos. Uno de los principales problemas radica en la subjetividad inherente a los juicios de expertos, que pueden verse influenciados por sesgos cognitivos y diferencias en la interpretación de los criterios de desempeño.

Se debería incorporar al establecimiento de los puntos de corte técnicas estadísticas que permitan reducir la dependencia exclusiva de la opinión de los jueces, proporcionando estimaciones más robustas. Sin embargo, es fundamental garantizar que estas innovaciones se utilicen como herramientas complementarias y no como sustitutos del juicio experto, ya que la interpretación de los estándares debe mantener un equilibrio entre rigor técnico y pertinencia educativa.

Presentación de resultados

Para hacer más amigables los resultados suele transformarse la escala de habilidades mediante una transformación lineal. En particular, cuando se presentan resultados agregados es recomendable generar series de valores plausibles para realizar los análisis estadísticos (Marsman, 2014). Si bien estas transformaciones facilitan la interpretación de los resultados, no siempre garantizan una mejor interpretación pedagógica.

Para presentar las puntuaciones en pruebas no cognitivas es posible emplear baremos o índices derivados de la TRI definidos en un grupo normativo. Ciertas características de la muestra pueden influir en los puntajes derivados dependiendo de la composición del grupo de referencia, lo que introduce sesgos en la interpretación de los resultados. Por eso, la selección adecuada de este grupo resulta crucial (Abad *et al.*, 2011). Además, es recomendable complementar estas medidas con enfoques de evaluación criterial.

Desafíos futuros

La evaluación educativa ha avanzado significativamente, pero enfrenta una serie de desafíos emergentes. Uno de los principales retos es la transición de modelos de evaluación rígidos y generalistas hacia enfoques más dinámicos, adaptativos e inclusivos. La integración de pruebas adaptativas informatizadas y la personalización de los instrumentos mediante inteligencia artificial y *machine learning* representan oportunidades clave para optimizar la medición del desempeño, permitiendo evaluaciones más precisas y relevantes para cada estudiante (Falcão *et al.*, 2023). Sin embargo, esto plantea interrogantes sobre la equidad y el acceso a estas tecnologías en contextos de desigualdad social y digital.

Otro desafío crítico es la consolidación de modelos evaluativos que no solo midan habilidades cognitivas ligadas exclusivamente al currículo, sino que también capturen aspectos del individuo que son mediadores del aprendizaje como la motivación, la autorregulación o la resiliencia. La combinación de estos factores con datos de desempeño académico permite un enfoque más integral del aprendizaje (Engelhardt & Goldhammer, 2019).

Por otra parte, es necesario incorporar métodos que permitan plantear intervenciones para la mejora. En este sentido, las ciencias cognitivas y las neurociencias han proporcionado evidencia fundamental para comprender los procesos cognitivos subyacentes al aprendizaje. Han destacado el papel crítico de las variables socioeconómicas, como el acceso a recursos educativos y la calidad del entorno

familiar, en la configuración tanto del desarrollo cerebral como del rendimiento académico (Benaros *et al.*, 2010, De la Torre & Minchen, 2014, García *et al.*, 2014, Vladisauskas & Goldin, 2020).

En términos metodológicos, la expansión del uso de big data en la evaluación educativa abre nuevas posibilidades, pero también desafíos éticos y técnicos. El análisis de tiempos de respuesta revela fluidez, comprensión sobre los procesos cognitivos y detecta falseamiento en evaluaciones virtuales mediante herramientas estadísticas (Rodríguez & Luzardo, 2020, Sanz *et al.*, 2020).

La aplicabilidad de los resultados de las EEE sigue siendo un área de preocupación. La transferencia efectiva de los hallazgos a la práctica educativa y a la toma de decisiones políticas es aún limitada. La capacitación de docentes y responsables educativos en la interpretación y uso de los datos derivados de estas evaluaciones es crucial para maximizar su impacto en la mejora de los aprendizajes y la reducción de desigualdades (Heyneman & Lee, 2014).

Es necesario avanzar en la identificación de sesgos utilizando DIF y en la utilización de TAI para obtener pruebas más justas, precisas y personalizadas (Cuellar *et al.*, 2021, Russell, 2011). La creación de modelos de evaluación adaptados al contexto y técnicamente rigurosos es una tarea imprescindible que requiere mayor investigación e innovación en el diseño de pruebas (Hambleton & Zenisky, 2011, Muñiz *et al.*, 2013).

El fortalecimiento de una perspectiva ética en el diseño y uso de las EEE es un desafío ineludible. La transparencia en los procesos de evaluación, la participación de múltiples actores en su desarrollo y la rendición de cuentas sobre su impacto deben ser principios centrales en la agenda futura (Backhoff, 2018).

La evolución de las EEE dependerá de su capacidad para adaptarse a las necesidades cambiantes de los sistemas educativos, promoviendo enfoques más justos, inclusivos y efectivos para la mejora de la calidad de la educación.

Discusión y conclusiones

Los aportes de este artículo subrayan el valor estratégico de las EEE como herramientas para la mejora de la calidad educativa. Sin embargo, su implementación y aplicación presentan limitaciones significativas que pueden restringir su efectividad e impacto real en la mejora de los sistemas educativos. El análisis del proceso de creación, implementación y análisis de estas pruebas refleja avances significativos en la precisión metodológica, la inclusión de nuevas dimensiones y la adopción de enfoques tecnológicos.

Un aspecto crítico identificado es la desconexión frecuente entre el diseño técnico de las evaluaciones y su aplicación práctica, especialmente en el contexto latinoamericano. Esto puede limitar el impacto de las EEE al reducir su utilidad para la toma de decisiones basadas en evidencia.

Además, el potencial investigativo de las bases de datos generadas por estas evaluaciones está subexplotado en la región. Su uso requiere una mayor capacitación técnica (Heyneman & Lee, 2014). En muchos países, la información generada no se traduce en estrategias concretas para mejorar la enseñanza y el aprendizaje, lo que reduce su impacto (Engelhardt & Goldhammer, 2019).

En conclusión, aunque las EEE han evolucionado significativamente y continúan siendo herramientas clave en la evaluación del aprendizaje, sus limitaciones deben ser abordadas con enfoques más integrales y equitativos.

Es necesario desarrollar estrategias para garantizar que los datos generados sean utilizados de manera efectiva en la toma de decisiones y la mejora de los sistemas educativos. Solo así será posible transformar las evaluaciones educativas estandarizadas en instrumentos que realmente contribuyan a la equidad y calidad educativa.

Notas:

[1] Esta publicación se desarrolló en el marco del Programa Grupos I+D financiado por la Comisión Sectorial de Investigación Científica (CSIC) de la Udelar.

Aprobación final del artículo:
Dra. Verónica Zorrilla de San Martín, editora responsable de la revista.

Contribución de autoría:
Pilar Rodríguez: conceptualización, curación de datos, investigación, diseño de metodología, administración, supervisión, visualización, escritura del borrador y revisión del manuscrito.
Juan Soca: conceptualización, curación de datos, investigación, diseño de metodología, visualización, escritura del borrador y revisión del manuscrito.
Mauricio Castillo: curación de datos, investigación, diseño de metodología, visualización, escritura del borrador y revisión del manuscrito.
Mario Luzardo: conceptualización, curación de datos, investigación, diseño de metodología, supervisión, visualización, escritura del borrador y revisión del manuscrito.

Disponibilidad de los datos:
El conjunto de datos utilizados para este artículo se encuentran en las bases de datos utilizadas y mencionadas en la metodología.

Referencias

ABAD, F. J., OLEA, J., PONSODA, V., & GARCÍA, C. (2011). *Medición en ciencias sociales y de la salud*. Síntesis.

AGENCIA DE CALIDAD DE LA EDUCACIÓN (2014). *Informe Técnico SIMCE 2012*. Agencia de Calidad de la Educación.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (2014). *Standards for educational and psychological testing*. American Educational Research Association.

BACKHOFF, E. (2018). Evaluación estandarizada de logro educativo: contribuciones y retos. *Revista Digital Universitaria*, *19*(6), 1-15. http://doi.org/10.22201/codeic.16076079e.2018.v19n6.a3

BECK, K. (2020). Ensuring content validity of psychological and educational tests—the role of experts. *Frontline Learning Research, 8*(6), 1-37. https://doi.org/10.14786/flr.v8i6.517

BENAROS, S., LIPINA, S. J., SEGRETIN, M. S., HERMIDA, M. J., & COLOMBO, J. A. (2010). Neurociencia y educación: hacia la construcción de puentes interactivos. *Revista de Neurología*, *50*(3), 179-186.

BLAIR, C., & RAZZA, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child development*, *78*(2), 647–663.

BOND, T. G., & FOX, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

CARLSON, J. E., & VON DAVIER, M. (2013). Item Response Theory. *ETS Research Report Series*, *2013*(2), i-69. https://doi.org/10.1002/j.2333-8504.2013.tb02335.x

CIZEK, G. J., & BUNCH, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications.

CORREA-ROJAS, J. (2021). Coeficiente de Correlación Intraclase: Aplicaciones para estimar la estabilidad temporal de un instrumento de medida. *Ciencias Psicológicas*, *15*(2), 1-12. https://doi.org/10.22235/cp.v15i2.2318

CROCKER, L., & ALGINA, J. (2008). *Introduction to classical and modern test theory*. CENGAGE Learning.

CUELLAR, E., PARTCHEV, I., ZWITSER, R., & BECHGER, T. (2021). Making sense out of measurement non-invariance: how to explore differences among educational systems in international large-scale assessments. *Educational Assessment, Evaluation and Accountability,* 33, 9-25. https://doi.org/10.1007/s11092-021-09355-x

DE AYALA, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

DE LA TORRE, J., & MINCHEN, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89-97. https://doi.org/10.1016/j.pse.2014.11.003

DEHAENE, S. (2019). ¿Cómo aprendemos?: Los cuatro pilares con los que la educación puede potenciar los talentos de nuestro cerebro. Siglo XXI Editores.

DUMAS, D., DONG, Y., & MCNEISH, D. (2022). How fair is my test: A ratio statistic to help represent consequential validity. *European Journal of Psychological Assessment, 39*(6), 416-423. https://doi.org/10.1027/1015-5759/a000724

ENGELHARDT, L., & GOLDHAMMER, F. (2019). Validating test score interpretations using time information. *Frontiers in Psychology*, *10*, 1131. https://doi.org/10.3389/fpsyg.2019.01131

FALCÃO, F., PEREIRA, D. M., GONÇALVES, N., DE CHAMPLAIN, A., COSTA, P., & PÊGO, J. M. (2023). A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation. *Advances in Health Sciences Education*, *28*(5), 1441-1465.

FERNÁNDEZ ALONSO, R., & MUÑIZ FERNÁNDEZ, J. (2011). Diseño de cuadernillos para la evaluación de las competencias básicas. *Aula abierta*, *39*(2), 3-34.

FERRANDO, P. J., LORENZO SEVA, U., HERNÁNDEZ DORADO, A., & MUÑIZ, J. (2022). Decalogue for the factor analysis of test items. *Psicothema*, *34*(1), 7-17. https://doi.org/10.7334/psicothema2021.456

FERRER, G. (2006). *Estándares en educación. Implicancias en América Latina*. PREAL.

FLORA, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484-501.

GARCÍA, P. E., ABAD, F. J., OLEA, J., & AGUADO, D. (2013). A new IRT-based standard setting method: Application to eCat-Listening. *Psicothema*, *25*(2), 238-244.

GARCÍA, P. E., OLEA, J., & DE LA TORRE, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema*, *26*(3), 372-377. https://doi.org/10.7334/psicothema2013.322

GIERL, M. J., & HALADYNA, T. M. (2013). *Automatic item generation: Theory and practice*. Routledge.

GIERL, M. J., LAI, H., & TANYGIN, V. (2021). *Advanced methods in automatic item generation*. Routledge.

GÖTZ, F. M., MAERTENS, R., LOOMBA, S., & VAN DER LINDEN, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, *29*(3), 494–518. https://doi.org/10.1037/met0000540

GUO, H., RÍOS, J. A., HABERMAN, S., LIU, O. L., WANG, J., & PAEK, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173–183.

HALADYNA, T. M., & RODRIGUEZ, M. C. (2013). *Developing and validating test items*. Routledge.

HAMBLETON, R. K., & ZENISKY, A. L. (2011). Translating and adapting tests for cross-cultural assessments. En D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46-70). Cambridge University Press.

HEYNEMAN, S., & LEE, B. (2014). The impact of international studies of academic achievement on policy and research. En L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment. Background, Technical Issues and Methods of Data Analysis* (pp. 37-72). CRC Press.

HIDALGO-MONTESINOS, M. D., & FRENCH, B. F. (2016). Una introducción didáctica a la Teoría de Respuesta al Ítem para comprender la construcción de escalas. *Revista de Psicología Clínica con Niños y Adolescentes*, *3*(2), 13-21.

INSTITUTO COLOMBIANO PARA LA EVALUACIÓN DE LA EDUCACIÓN (2011). *Informe técnico de las pruebas Saber 5.to y 9.no 2009*. ICFES.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (2024). *Saeb 2023: detalhamento da população e resultados: nota técnica n.º 18/2023/CGMEB/DAEB*. INEP.

INSTITUTO NACIONAL DE EVALUACIÓN EDUCATIVA (2017). *Aristas. Marco de lectura en tercero y sexto de primaria*. INEEd.

INSTITUTO NACIONAL DE EVALUACIÓN EDUCATIVA (2018). *Aristas. Marco general de la evaluación*. INEEd.

INSTITUTO NACIONAL DE EVALUACIÓN EDUCATIVA (2020). *Aristas 2018. Informe de resultados de tercero de educación media*. INEEd.

INSTITUTO NACIONAL DE EVALUACIÓN EDUCATIVA (2021). *Aristas 2020. Primer informe de resultados de tercero y sexto de educación primaria*. INEEd.

INSTITUTO NACIONAL PARA LA EVALUACIÓN DE LA EDUCACIÓN (2004). *El Aprendizaje del español y las Matemáticas en la educación básica en México. Sexto de primaria y tercero de secundaria.* INEE.

INSTITUTO NACIONAL PARA LA EVALUACIÓN DE LA EDUCACIÓN (2019). *Manual técnico del Plan Nacional para la Evaluación de los Aprendizajes PLANEA 2015. Educación media superior*. INEE.

JACKSON STENNER, A., SMITH III, M., & BURDICK, D.S. (2022). Toward a Theory of Construct Definition. En W. P. Fisher & P. J. Massengill (Eds.), *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement* (pp. 43-55). Springer.

JOINT COMMITTEE ON STANDARDS FOR EDUCATIONAL EVALUATION (JCSEE) (2010). *The Program Evaluation Standards*. Sage.

JORNET MELIÁ, J. M. (2017). Evaluación estandarizada. *Revista Iberoamericana de Evaluación Educativa (RIEE), 10*(1), 5-8.

JORNET MELIÁ, J. M., & GONZÁLEZ-SUCH, J. (2009). Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios sobre educación, 16*, 103-123.

LEE, S., & WINKE, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing*, *35*(2), 239-269. https://doi.org/10.1177/0265532217704009

LINACRE, J. M. (2012). *Winsteps Rasch Measurement Computer Program User's Guide*. Winsteps.

LINN, R. (2003). Performance Standards: Utility for Different Uses of Assessments. *Education Policy Analysis Archives*, *11*(31).

LOKEN, E., & GELMAN, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584-585.

LUZARDO, M. (2019). Item Selection Algorithms in Computerized Adaptive Test Comparison Using Items Modeled with Nonparametric Isotonic Model. En M. Wiberg, S. Culpepper, R. Janssen, J. González & D. Molenaar (Eds.), *Quantitative Psychology* (pp. 95-105). Springer International Publishing. https://doi.org/10.1007/978-3-030-01310-3_6

LUZARDO, M., & RODRÍGUEZ, P. (2015). A nonparametric estimator of a monotone item characteristic curve. En L. A. van der Ark, D. Bolt, W. C. Wang, A. Douglas & S. M. Chow (Eds.), *Quantitative Psychology* (pp. 99–108), Springer.

MAHIAS FINGER, P., & POLLONI ERAZO, M. P. (2019). *Cuadernillo técnico de evaluación educativa Desarrollo de instrumentos de evaluación: pruebas.* Centro de Medición MIDE UC; INEE.

MARSMAN, M. (2014). *Plausible values in statistical inference* [Tesis doctoral, University of Twente].

MARTIN, M. O., MULLIS, I. V. S., & FOY, P. (2015). Assessment Design for PIRLS, PIRLS Literacy, and ePIRLS in 2016. En I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2016 Assessment Framework*. TIMSS & PIRLS International Study Center.

MASTERS, G. N. (2016). Partial Credit Models. En W. J. van der Linden (Ed.), *Handbook of modern item response theory*. CRC Press.

MCDONALD, R. P. (1999). *Test Theory: A Unified Treatment.* Erlbaum.

MINISTERIO DE EDUCACIÓN DEL PERÚ (2024). *Reporte técnico de la Evaluación Nacional de Logros de Aprendizajes de Estudiantes 2023* (ENLA). MINEDU.

MITZEL, H. C., LEWIS, D. M., PATZ, R. J., & GREEN, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Lawrence Erlbaum.

MUÑIZ, J. (2018). *Introducción a la Psicometría: Teoría Clásica y TRI.* Pirámide.

MUÑIZ, J., & FONSECA-PEDRERO, E. (2019). *Diez pasos para la construcción de un test*. *Psicothema*, *31*(1), 7–16.

MUÑIZ, J., ELOSUA, P., & HAMBLETON, R. K. (2013). Directrices para la traducción y adaptación de los test: segunda edición. *Psicothema*, *25*(2), 151-157. https://doi.org/10.7334/psicothema2013.24

NATIONAL ASSESSMENT GOVERNING BOARD (2022). *Mathematics Assessment Framework for the 2022 to 2024 National Assessment of Educational Progress*. NAGB.

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (2023). *Technical Documentation: Student Test Form and Booklet Block Design*. NAEP.

NORTH, B., & JONES, N. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling.* Council of Europe.

OLEA, J., & PONSODA, V. (2013). *Test adaptativos informatizados*. Editorial UNED.

OREALC-UNESCO (2016). *Reporte Técnico. Tercer Estudio Regional Comparativo y Explicativo (TERCE)*.

PÉREZ JUSTE, R. (2006). *Evaluación de programas educativos*. La Muralla.

POPHAM, W. J. (1999). Where Large Scale Educational Assessment Is Heading and Why It Shouldn't. *Educational Measurement: Issues and Practice*, *18*(3), 13-17. https://doi.org/10.1111/j.1745-3992.1999.tb00268.x

RAYKOV, T. (2007). Reliability if deleted, not 'alpha if deleted': Evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology*, *60*(2), 201-216. https://doi.org/10.1348/000711006X115954

RECKASE, M. D. (2009). *Multidimensional Item Response Theory*. Springer.

RECKASE, M. D. (2016). Multidimensional logistic models. En W. J. van der Linden (Ed.), *Handbook of Item Response Theory: Models* (pp. 189-210). CRC Press.

REYNOLDS, K. A., & MONCALEANO, S. (2021). Digital module 26: Content alignment in standards-based educational assessment. *Educational Measurement: Issues & Practice*, *40*(3), 127-128. https://doi.org/10.1111/emip.12405

RÍOS, J. A., & GUO, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, *33*(4), 263-279.

RODRÍGUEZ MORALES, P. (2017). Creación, Desarrollo y Resultados de la Aplicación de Pruebas de Evaluación basadas en Estándares para Diagnosticar Competencias en Matemática y Lectura al Ingreso a la Universidad. *Revista Iberoamericana de Evaluación Educativa*, *10*(1), 89-107. https://doi.org/10.15366/riee2017.10.1.005

RODRÍGUEZ MORALES, P., & LUZARDO VERDE, M. (2020). Cómo asegurar evaluaciones válidas y detectar falseamiento en pruebas a distancia síncronas. *Revista Digital de Investigación en Docencia Universitaria*, *14*(2), e1240.

RODRÍGUEZ, P., & LUZARDO, M. (2019). A Modification of the IRT-Based Standard Setting Method. En M. Wiberg, S. Culpepper, R. Janssen, J. González & D. Molenaar (Eds.), *Quantitative Psychology* (pp. 65-74), Springer Nature. https://doi.org/10.1007/978-3-030-01310-3_6

RODRÍGUEZ, P., PÉREZ, G., & LUZARDO, M. (2017). Desarrollo y aplicación del primer test adaptativo informatizado (TAI) de Matemática para orientar trayectorias en la Universidad. En N. Peré (Comp.), *La Universidad Se Investiga* (pp. 1041-1048). CSE-ANEP.

RUSSELL, M. (2011). Personalizing assessment. En T. Gray & H. Silver-Pacuilla (Eds), *Breakthrough teaching and learning* (pp. 111–126). Springer.

RUTKOWSKI, D., RUTKOWSKI, L., & VON DAVIER, M. (2014). A brief introduction to modern international large scale assessment. En L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment. Background, Technical Issues and Methods of Data Analysis* (pp. 3-10). CRC Press.

SAMEJIMA, F. (2016). Graded Response Models. En W. J. van der Linden (Ed.), *Handbook of modern item response theory*. CRC Press.

SANZ, S., LUZARDO, M., GARCÍA, C., & ABAD, F. J. (2020). Detecting cheating methods on unproctored Internet tests. *Psicothema*, *32*(4), 549-558. https://dx.doi.org/10.7334/psicothema2020.86

SIJSTMA, K., & MOLENAAR, I. W. (2016). Mokken models. En W. J. van der Linden (Ed.), *Handbook of modern item response theory*. CRC Press.

SIRECI, S., & BENÍTEZ, I. (2023). Evidence for test validation: a guide for practitioners. *Psicothema*, *35*(3), 217-226. https://dx.doi.org/10.7334/psicothema2022.477

SISTEMA DE MEDICIÓN DE LA CALIDAD DE LA EDUCACIÓN (SIMCE) (2010) *Resultados Nacionales SIMCE 2009*. Agencia de Calidad de la Educación.

SOCA, J. M. (2018). *Tendencias de Investigación e Innovación en Evaluación Educativa.* CONACyT – INEE.

SOLAND, J. (2018). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, *120*(12).

SOLAND, J., & KUHFELD, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment*, *24*(4), 327–342.

SWAMINATHAN, H., & ROGERS, H. J. (2016). Normal-ogive multidimensional models. En W. J. van der Linden (Ed.), *Handbook of Item Response Theory: Models* (pp. 167–188). CRC Press.

TEIG, N., & STEINMANN, I. (2023). Leveraging large-scale assessments for effective and equitable school practices: the case of the nordic countries. *Large-scale Assessments in Education, 11*, 11-21. https://doi.org/10.1186/s40536-023-00172-w

THISSEN, D., & CAI, L. (2016). Nominal Categories Models. En W. J. van der Linden (Ed.), *Handbook of modern item response theory*. CRC Press.

TOURÓN, J. (2009). *El establecimiento de estándares de rendimiento en los sistemas educativos*. *Estudios sobre Educación*, *16*, 127–146.

VAN DER LINDEN, W. J. (2016). Unidimensional Logistic Response Models. En W. J. van der Linden (Ed.), *Handbook of modern item response theory* (pp. 19-30). CRC Press.

VAN DER LINDEN, W. J. (2018). *Handbook of item response theory.* CRC Press.

VAN DER LINDEN, W. J., & GLAS, C. A. (2000). *Computerized adaptive testing: Theory and practice*. Kluwer Academic.

VLADISAUSKAS, M., & GOLDIN, A. P. (2020). 20 años de entrenamiento cognitivo: una perspectiva amplia. *Journal of Neuroeducation*, *1*(1), 130-135.

VON DAVIER, M. (2016). Rasch Models. En W. J. Van der Linden (Ed.), *Handbook of item response theory* (pp. 31-45). CRC Press.

WAGEMAKER, H. (2014). International large-scale assessments: from research to policy. En L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment. Background, Technical Issues and Methods of Data Analysis* (pp. 11-36). CRC Press.

WISE, S. L., & MA, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada.

WRY, E., & MULLIS, I. V. S. (2023). Developing the PIRLS 2021 achievement instruments. En M. von Davier, I. V. S. Mullis, B. Fishbein & P. Foy (Eds.), *Methods and Procedures: PIRLS 2021 Technical Report* (pp. 1-24). Boston College; TIMSS; PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2101.kb7549

XU, X., & DOUGLAS, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, *71*, 121-137.