Diseño e implementación de evaluaciones estandarizadas de logros de aprendizaje

PERÚ Ministerio de Educación

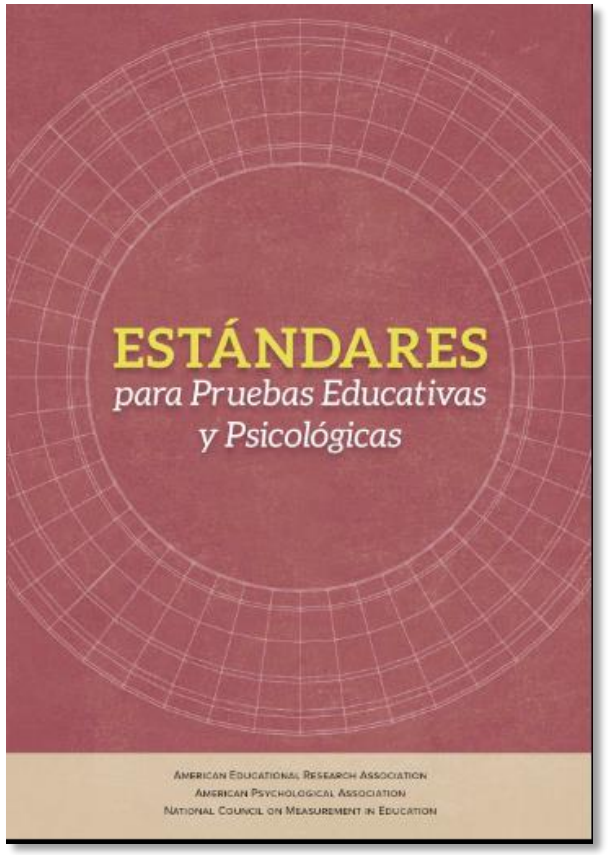
UMC
Oficina de Medición de la Calidad de los Aprendizajes

Principios del diseño y de la construcción de pruebas estandarizadas de logros de aprendizaje

PERÚ Ministerio de Educación

UMC
Oficina de Medición de la Calidad de los Aprendizajes

¿Qué principios rigen el desarrollo de evaluaciones estandarizadas?



Validez

Grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para los usos propuestos de las pruebas.

Confiabilidad

Coherencia de puntajes entre replicaciones de un procedimiento de evaluación, independientemente de cómo se estime o reporte esta coherencia.

Imparcialidad *(fairness)*

Capacidad de respuesta a características individuales y contextos de evaluación de modo que los puntajes de la prueba arrojen interpretaciones válidas para los usos previstos.

Fuentes de evidencias de validez

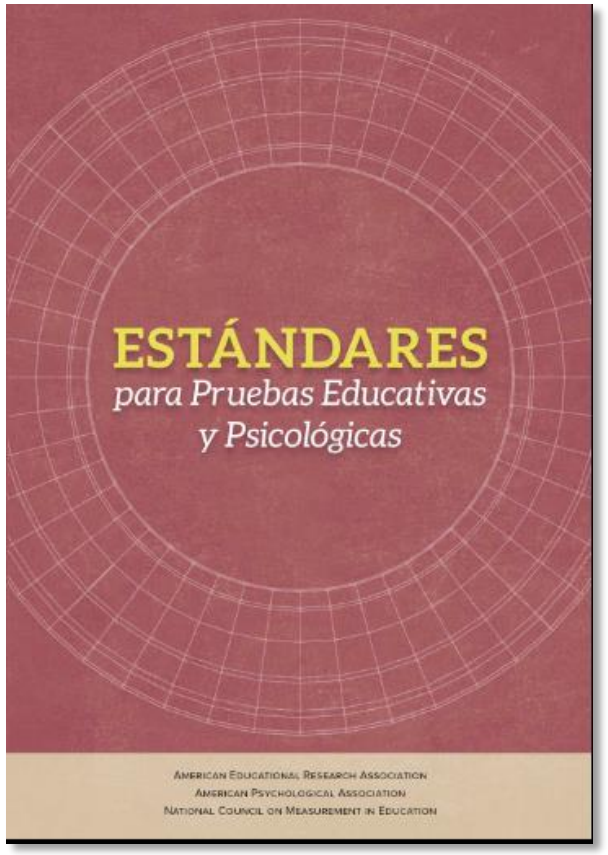
1. **Contenido**: congruencia entre el contenido del test y el dominio que se supone debe medir, adecuación al grupo.

2. **Proceso de respuesta**: consistencia entre las actividades que el test demanda a los examinados y el proceso que se supone representa.

3. **Estructura interna**: consistencia entre la estructura del constructo y las relaciones entre los ítems y/o subescalas del test. Se incluyen evidencias de funcionamiento diferencial de los ítems.

4. **Consecuencias de los tests**: evidencias referidas a los efectos de la evaluación. Análisis de los efectos positivos / negativos implicados en las decisiones basadas en el uso de los tests.

¿Qué principios rigen el desarrollo de evaluaciones estandarizadas?



Validez

Grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para los usos propuestos de las pruebas.

Confiabilidad

Coherencia de puntajes entre replicaciones de un procedimiento de evaluación, independientemente de cómo se estime o reporte esta coherencia.

Imparcialidad *(fairness)*

Capacidad de respuesta a características individuales y contextos de evaluación de modo que los puntajes de la prueba arrojen interpretaciones válidas para los usos previstos.

Confiabilidad

1. Tradicionalmente: estabilidad de los resultados.

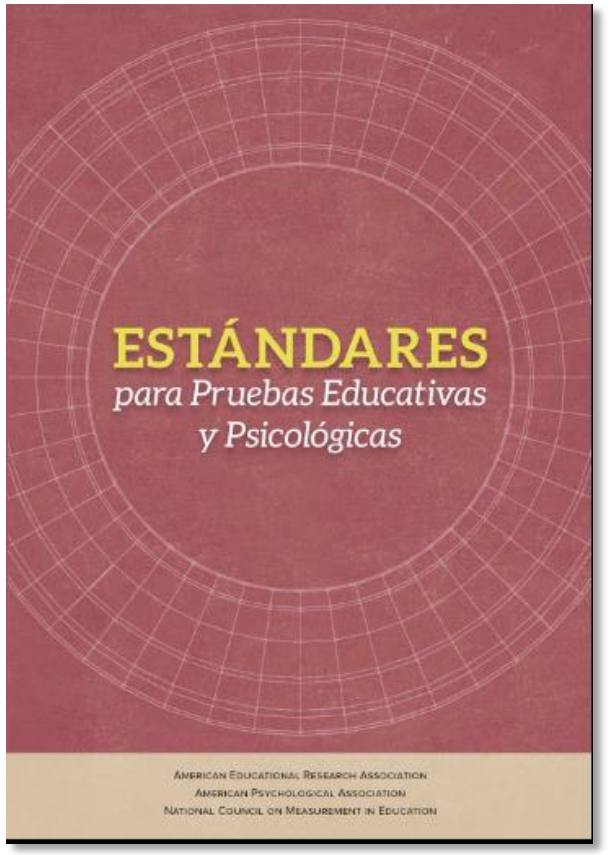
2. Modelo TCT: X = V + e
   - ¿Qué tan bien representa el puntaje observado al puntaje verdadero?

3. ¿Qué explica la variabilidad en los resultados? ¿El rasgo latente o el error?

4. La confiabilidad es una propiedad de los puntajes y del uso que se le dará a los puntajes, no es una propiedad del test en sí.

5. Algunos modelos (como el de Rasch) usan el índice de separación de personas.

¿Qué principios rigen el desarrollo de evaluaciones estandarizadas?



Validez

Grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para los usos propuestos de las pruebas.

Confiabilidad

Coherencia de puntajes entre replicaciones de un procedimiento de evaluación, independientemente de cómo se estime o reporte esta coherencia.

Imparcialidad *(fairness)*

Capacidad de respuesta a características individuales y contextos de evaluación de modo que los puntajes de la prueba arrojen interpretaciones válidas para los usos previstos.

Imparcialidad (*fairness*)

1. La imparcialidad pretende garantizar que la prueba estandarizada sea justa, sin prejuicios ni sesgos, de tal modo que las medidas que se obtienen de ella sean resultado de la comparación de un rasgo latente en condiciones de equidad contextual.

2. Algunas variables pueden inducir a respuestas no objetivas, con las que se producen medidas erróneas y apreciaciones injustas a personas de un grupo específico, en función de **género, grupo etario, nivel socioeconómico, antecedentes culturales, pertenencia religiosa o étnica**, entre otras. En ese sentido, los evaluadores deben cuidar que el lenguaje, las situaciones y el contexto de los ítems no vulneren la dignidad de las personas, que no induzcan la movilización de rasgos latentes no previstos que pudieran favorecer que se movilicen actitudes positivas o negativas en ciertos grupos o individuos.

Imparcialidad (*fairness*)

3. Los estándares (AERA, APA y NCME, 2014) señalan la especificación del **contenido** del test como una fuente potencial de sesgo en una medición. Ello podría ocurrir tanto por **conocimiento de contenidos específicos**, como por **aspectos motivacionales** respecto de tales contenidos.

4. 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.

Imparcialidad (*fairness*)

5. Asimismo, 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.

6. El **análisis de sesgo** debe hacerse ***a priori***, al definir el **objeto y las especificaciones de diseño de la prueba** y ***a posteriori*** con **técnicas estadísticas** avanzadas para detectarlo, medirlo y realizar ajustes matemáticos de cambio de escala e igualación de los resultados obtenidos por los grupos potencialmente afectados por dicho sesgo. **Análisis de funcionamiento diferencial del ítem** (DIF).

Diseño de evaluaciones de logros de aprendizaje

PERÚ Ministerio de Educación

UMC
Oficina de Medición de la Calidad de los Aprendizajes

¿Para qué evaluar? – Propósitos de la evaluación

1. ¿Será una evaluación normativa o criterial?
2. ¿Tendrá altas o bajas consecuencias?
3. ¿Qué alcance tendrá? ¿Censal o muestral?
4. ¿Con qué periodicidad se implementará? ¿Anual? ¿Bienal? ¿Trienal?

¿Para qué evaluar? – Propósitos de la evaluación

1. Diagnosticar y monitorear la calidad de los aprendizajes en el sistema educativo con el fin de establecer una agenda de prioridades de política educativa.

2. Servir de insumo para la ejecución y evaluación de políticas educativas.

3. Colaborar en la rendición de cuentas compartida y con justicia, de los distintos actores educativos sobre su papel en la mejora de la calidad del servicio educativo.

4. Seleccionar a un conjunto de estudiantes.

¿Qué evaluar? – Constructo a evaluar

1. ¿Qué áreas y competencias evaluar? ¿Es recomendable y posible evaluar todas las áreas y competencias del currículo?

2. Las competencias de un área, ¿se evaluarán por separado o como un solo constructo? Modelos unidimensionales o modelos multidimensionales

3. ¿Cómo definimos el constructo a evaluar?

4. ¿Qué evidencias de validez se recogerán acerca del constructo?

¿A quién evaluar? – Población y muestra

1. ¿Cómo definimos a la población? ¿Hay algún criterio de exclusión? ¿Cuál es el marco poblacional?

2. Si la evaluación fuera muestral, ¿qué nivel de desagregación tendrá la información que se genere?

3. ¿Cuál es el marco muestral? ¿Qué tamaño tendrá la muestra?

4. ¿Qué técnica de muestreo se usará?

¿Cómo evaluar? – Metodología

1. ¿Cuál es el modelo de evaluación? ¿Qué dimensiones tiene?

2. ¿Qué formatos de ítems se utilizarán? ¿Existen los recursos humanos y económicos para codificar ítems de respuesta construida?

3. ¿Cómo serán el diseño de bloques y el ensamblaje de las formas?

4. ¿Cuál será el modelo de medición para procesar los resultados? ¿Se enmarca en la teoría clásica de los test, modelos de un parámetro o modelos de dos o tres parámetros?

5. ¿Cuáles serán los principales parámetros de ajuste al modelo, así como los principales estadísticos que el modelo obtendrá? ¿De qué forma alimentarán las decisiones y los análisis esos parámetros y estadísticos?

¿Cómo usar las evaluaciones? - Uso de resultados

1. ¿El sistema está en capacidad de usar la información proveniente de las evaluaciones?

2. ¿Quiénes tendrán acceso a los resultados?

3. ¿Qué estrategias y herramientas de difusión se usarán?

4. ¿Qué se espera que los usuarios hagan con la información?

Referencias

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

Brennan, R. L. (Ed.). (2006). *Educational measurement (*4ª ed.*)*. Wesport, CT: American Council on Education y Praeger Publishers.

Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory*. Masson, OH: Cengage Learning.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01130.x

Lissitz, R. (Ed.) (2009). *The concept of validity. Revisions, new directions, and applications*. Charlotte, NC: IAP.

Referencias

Muñiz, J. (1999). *Teoría Clásica de los Tests* (2ª. ed.). Madrid: Pirámide.

Newton, P. E., y Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*(3), 301–319. http://dx.doi.org/10.1037/a0032969.

Sijtsma, K. (2009). On the use, misuse and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120. http://dx.doi.org/10.1007/s11336-008-9101-0

Thomson, B. (Ed.) (2003). *Score reliability. Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage Publications.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Hoboken, NJ: Wiley.

UMC

Oficina de Medición de la Calidad de los Aprendizajes