CUADERNOS METODOLÓGICOS

ISSN 2215-681X

Cuaderno 9

Estándares de calidad para pruebas estandarizadas de alto impacto en el contexto académico y profesional costarricense

IIP Instituto de Investigaciones Psicológicas

2022

Miembros del comité editorial

Dr. Luis Rojas Torres. Coordinador
Dra. Mónica Arias Monge
Dr. Andrés Castillo Vargas
M.L. Diana Martínez Alpízar
Dr. Odir Rodríguez Villagra

Diseño gráfico

Licda. María Gabriela González Acuña

Diagramación

Tannya Monge Carvajal

001.4 Cuadernos Metodológicos [recurso electrónico] / Instituto
C961c de Investigaciones Psicológicas (IIP).
V. 9 (2022) . San José, Costa Rica: Instituto de
Investigaciones Psicológicas, IIP, 2022
Un recurso en línea: 3,5 Mb

Publicación anual
ISSN 2215-681X

1. PRUEBAS ESTANDARIZADAS. 2.EVALUACIÓN EDUCATIVA.
3. MEDICIÓN EDUCATIVA. 4. INVESTIGACIÓN.

Para dar crédito utilice la siguiente referencia bibliográfica:

Moreira-Mora, T.E., Alfaro-Rojas, L., Brizuela-Rodríguez, A., Chacón-Vega, C., Gómez-González, E., Jiménez-Alfaro, K., Jiménez-Segura, F., Mena-Castillo, P., Montero-Rojas, E., Picado-Barrantes, H., Rojas-Rojas, G., Rojas-Torres, L., Smith-Castro, V., Solórzano-Salas, M.J. & Villarreal-Galera, M.P. (2022). *Estándares de calidad para pruebas de alto impacto en el contexto académico y profesional costarricense*. Instituto de Investigaciones Psicológicas, Universidad de Costa Rica.

Reconocimiento

Los miembros de la Comisión interinstitucional de estándares de calidad para pruebas estandarizadas de alto impacto en el ámbito académico y profesional de Costa Rica agradecen el apoyo brindado por el Instituto de Investigaciones Psicológicas, en especial a la Dra. Vanessa Smith Castro, por su gestión como directora en el periodo 2013 a 2019. También un reconocimiento a la Dra. Eiliana Montero Rojas por su labor como coordinadora de la comisión.

UNIVERSIDAD DE COSTA RICA

TEC | Tecnológico de Costa Rica

Instituto de Investigaciones Psicológicas (IIP)
Facultad de Ciencias Sociales
Finca 2
Sede Universitaria Rodrigo Facio Brenes
San Pedro, Montes de Oca
11501-2060
Web: iip.ucr.ac.cr

Autores

1. Tania Elena Moreira-Mora
Instituto Tecnológico de Costa Rica
tmoreira@itcr.ac.cr

2. Lucrecia Alfaro-Rojas
Universidad de Costa Rica
lucrecia.alfaro@ucr.ac.cr

3. Armel Brizuela-Rodríguez
Universidad de Costa Rica
armel.brizuelarodriguez@ucr.ac.cr

4. Carolina Chacón-Vega
Ministerio de Educación Pública
carolina.chacon.vega@mep.go.cr

5. Eulalia Gómez-González
Universidad de Costa Rica
eulalia.gomez@ucr.ac.cr

6. Karol Jiménez-Alfaro
Universidad de Costa Rica
karol.jimenez@ucr.ac.cr

7. Flor Jiménez-Segura
Universidad de Costa Rica
flor.jimenez@ucr.ac.cr

8. Pablo Mena-Castillo
Ministerio de Educación Pública
pablo.mena.castillo@mep.go.cr

9. Eiliana Montero-Rojas
Universidad de Costa Rica
eiliana.montero@ucr.ac.cr

10. Helen Picado-Barrantes
Universidad de Costa Rica
helen.picado@ucr.ac.cr

11. Guaner Rojas-Rojas
Universidad de Costa Rica
guaner.rojas@ucr.ac.cr

12. Luis Rojas-Torres
Universidad de Costa Rica
luismiguel.rojas@ucr.ac.cr

13. Vanessa Smith-Castro
Universidad de Costa Rica
vanessa.smith@ucr.ac.cr

14. María Julieta Solórzano-Salas
Universidad de Costa Ricav
julieta.solorzano@ucr.ac.cr

15. Maria Paula Villarreal-Galera
Universidad de Costa Rica
maria.villarreal@ucr.ac.cr

16. Fiscalía del Colegio de Profesionales en Psicología de Costa Rica
fiscaladjunta@psicologiacr.com

Miembros de la Comisión interinstitucional de estándares de calidad para pruebas estandarizadas de alto impacto en el ámbito académico y profesional de Costa Rica

- Eiliana Montero Rojas, Coordinadora de la Comisión, Programa Permanente de la Prueba de Aptitud Académica, Instituto de Investigaciones Psicológicas. Escuela de Estadística, UCR
- Ana Cristina Monge Vargas, Colegio de Profesionales en Psicología de Costa Rica
- Armel Brizuela Rodríguez, Instituto de Investigaciones Psicológicas, UCR
- Carolina Chacón Vega, Dirección de Gestión y Evaluación de la Calidad (DGEC), MEP
- Gerardo Solís Sequeira, Colegio de Abogados y Abogadas de Costa Rica
- Eulalia Ivette Gómez González, Oficina de Recursos Humanos, UCR
- Fernando Chacón Rojas, Dirección General del Servicio Civil (DGSC)
- Guaner Rojas Rojas, Programa Permanente de la Prueba de Aptitud Académica, Instituto de Investigaciones Psicológicas, UCR
- Helen Picado Barrantes, Oficina de Recursos Humanos, UCR
- Karol Jiménez Alfaro, Programa Permanente de la Prueba de Aptitud Académica, Instituto de Investigaciones Psicológicas, UCR
- Lucrecia Alfaro Rojas, Oficina de Becas y Atención Socioeconómica, UCR
- Luis Miguel Rojas Torres, Programa Permanente de la Prueba de Aptitud Académica, Prueba de Habilidades Cuantitativas, Instituto de Investigaciones Psicológicas, UCR
- María Elisa Vásquez Acuña, Escuela de Medicina, UCR
- Manfred Araya Quirós, Colegio de Abogados y Abogadas de Costa Rica
- María Julieta Solórzano Salas, Programa de Posgrado en Educación énfasis en Evaluación Educativa, Escuela de Orientación y Educación Especial, UCR
- María Paula Villarreal Galera, Programa de Posgrado en Especialidades Médicas, Sistema de Estudios de Posgrado. Instituto de Investigación en Educación, UCR
- Mario Medina Céspedes, Escuela de Medicina, UCR
- Pablo José Mena Castillo, Dirección de Gestión y Evaluación de la Calidad (DGEC), MEP
- Tania Elena Moreira Mora, Comité Examen de Admisión, Instituto Tecnológico de Costa Rica
- Vanessa Smith Castro, Instituto de Investigaciones Psicológicas, UCR
- Vilma B. Mesén Madrigal, Colegio de Abogados y Abogadas de Costa Rica

Tabla de contenidos

Presentación

El uso de pruebas estandarizadas de alto impacto implica un proceso de construcción y validación guiado científicamente, utilizando los más altos estándares de calidad para garantizar la confiabilidad y validez de las inferencias que se realicen a partir de sus puntuaciones. Debido a que se trata de exámenes de grandes consecuencias para la vida y el desarrollo de las personas, las pruebas deben ser construidas, adaptadas y aplicadas utilizando la tradición científica en psicometría y medición educativa, que brinda un cuerpo de teorías y métodos para construir y validar este tipo de instrumentos y permite recolectar información empírica y evidencia sólida acerca de la confiabilidad y validez de los mismos.

En este contexto de la evaluación de alto impacto, el presente documento tiene como propósito orientar la labor y los planes de mejoramiento de las instituciones y organizaciones comprometidas con la validez, equidad y ética en la aplicación de pruebas de estandarizadas de alto impacto en contextos académicos y profesionales.

El logro de los estándares propuestos se debe asumir como un proceso formativo, por lo tanto, es posible que en una primera valoración no se cumplan a cabalidad los estándares que aquí se presentan.

Se espera que los programas responsables ejecuten acciones concretas para cumplir, en un plazo determinado, con aquellos estándares para los cuales no hay, en la actualidad, suficiente evidencia de su cumplimiento.

Parte 1

Pruebas estandarizadas de alto impacto en el contexto académico y profesional

María Julieta Solórzano-Salas

Las pruebas estandarizadas se pueden aplicar tanto en contextos académicos como profesionales, para efectos de este documento no se consideran las pruebas que se aplican individualmente a nivel clínico, forense y en evaluaciones con criterios de riesgo. Según la American Education Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME)

> Una prueba es un dispositivo evaluativo o procedimiento mediante el cual se obtiene una muestra de la conducta de las personas examinadas en un dominio especificado, el cual es posteriormente evaluado y calificado mediante un proceso estandarizado (2014, p. 2).

En esta misma línea teórica Hamilton, Stecher y Klein (2002) definen las pruebas estandarizadas de alto impacto como aquellas que se administran a un gran número de personas, por esto se consideran masivas; son desarrolladas por equipos especializadas y administradas como parte de un programa de pruebas. Según, Martínez-Riso (2009) la forma en que las pruebas se administran y califican está predeterminada de manera estándar.

Las pruebas estandarizadas en los contextos académicos, laborales y profesionales pueden tener diversas clasificaciones, dentro de las cuales se destaca la de Martínez, Hernández y Hernández (2006, pp. 18-20).

- En función del impacto para las personas examinadas: pueden ser altas como en el caso de la selección de personal o bajas como cuando se desarrolla un test para la investigación.
- En función del planteamiento del problema y tipo de respuesta: la persona responde a ítems de respuesta seleccionada o construye la respuesta tales como entrevista y productos de desempeño.
- En función del área de comportamiento: cognitivos y no cognitivos, referidos estos últimos a constructos como personalidad, intereses, actitudes, entre otros.
- En función de la modalidad de aplicación: se usan designaciones como individual/colectivo y papel y lápiz/ordenador.
- En función de las demandas temporales: velocidad o potencia.
- En función del grado de las demandas específicas de las dimensiones culturales para la resolución de la prueba.
- En función del modelo estadístico en que se basan las interpretaciones de las puntuaciones e inferencias a partir de la muestra de conductas medidas.
- En función del tipo de interpretación de las puntuaciones, puede ser referido a normas o a criterios

Las pruebas estandarizadas pueden favorecer en algunas de las siguientes decisiones según detallan Martínez, Hernández y Hernández (2006, p.27)

- Relativas al dominio del sujeto en una habilidad o conocimiento determinado.
- Diagnóstico de habilidades, conocimientos y dimensiones afectivas.

- Selección de personas con mayores habilidades en el constructo de interés.
- Asignación de personas dentro de una jerarquía.
- Clasificación de personas dentro ciertas categorías.
- Guiar a la persona sobre los aspectos que debe realizar para el logro de una tarea determinada.

Toda prueba estandarizada debe cumplir con los siguientes requerimientos según Martínez, Hernández y Hernández (2006, p.35)

- Fundamentación teórica del constructo o explicación del dominio que mide la prueba.
- Evidencias de validez para los usos que está destinado.
- Poblaciones a las que está dirigido.
- Instrucciones completas para su administración.
- Coeficientes de fiabilidad o datos de precisión de la medida.
- Sugerencias sobre los posibles usos de las puntuaciones.
- Clasificación de personas dentro ciertas categorías.
- Guiar a la persona sobre los aspectos que debe realizar para el logro de una tarea determinada.

Toda prueba estandarizada debe cumplir con los siguientes requerimientos según Martínez, Hernández y Hernández (2006, p.35)

- Fundamentación teórica del constructo o explicación del dominio que mide la prueba.
- Evidencias de validez para los usos que está destinado.
- Poblaciones a las que está dirigido.
- Instrucciones completas para su administración.
- Coeficientes de fiabilidad o datos de precisión de la medida.
- Sugerencias sobre los posibles usos de las puntuaciones.
- Normas de interpretación de las puntuaciones.

También resulta necesario comprender las diferencias entre las pruebas estandarizadas y de aula, especialmente, en el proceso de construcción, validación, calificación e interpretación de sus resultados como se resume en la Tabla 1.

Tabla 1
Diferencias entre las pruebas estandarizadas, de aula y clínicas.

Aspecto	Estandarizadas	De Aula	Clínicas
Medición de contenidos	La medición de los contenidos debe ser común para todos los centros educativos. Las pruebas rara vez	Consideran aspectos específicos del aprendizaje generado con un cierto proceso de mediación	No aplica

Aspecto	Estandarizadas	De Aula	Clínicas
	enfatizan aspectos locales del currículo.	pedagógica y contexto particular de dinámica del aula	
Calidad de los ítems	En general, si se siguen los procedimientos establecidos por la psicometría y la medición educativa, la calidad técnica es alta. Los ítems son escritos por especialistas, se realizan pruebas piloto y se crea un banco de ítems. La selección es en términos de su efectividad para realizar una medición adecuada.	La calidad de los ítems es desconocida, o al menos, no suele estar documentada. Típicamente la calidad es más baja, por el poco tiempo que dispone el docente y la falta de conocimientos especializados. Pocas veces es revisado por otros docentes, no se hacen pilotajes ni análisis de los ítems.	
Confiabilidad	Alta confiabilidad, comúnmente, es mayor a 0.8 (tanto en la Teoría Clásica de los Test como para el modelo de Rasch)	Usualmente desconocida, puede ser alta si la prueba es cuidadosamente construida, rara vez se calculan medidas empíricas.	Alta confiabilidad
Administración calificación	Procedimientos estandarizados (uniformes). Se brindan instrucciones específicas con protocolos de administración y calificación.	Se favorecen los procedimientos uniformes, pero hay laxitud. Muchas veces no hay protocolos o son poco específicos	

Aspecto	Estandarizadas	De aula	Clínicas
Interpretación puntajes	Los puntajes pueden ser comparados con grupos normativos o con niveles de desempeño (enfoque criterial). Utilizan diferentes modelos de análisis, manuales de la prueba y otras guías para la interpretación y el uso de los resultados.	Las comparaciones de puntajes e interpretaciones se limitan a la situación local del centro educativo o del aula.	
Número examinados	Aplicación masiva, usualmente varios cientos o miles.	Aplicación reducida, grupos pequeños en un contexto circunscrito.	Aplicación individual

En los siguientes apartados del documento se profundiza en los distintos procesos técnicos relativos a la construcción, validación, calificación de las pruebas estandarizadas.

Referencias

American Educational Research Association, American Psychological Association y National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Gronlund N.E. & Linn R.L. (1990). *Measurement and Evaluation in Teaching*. Macmillan.

Hamilton, L. S., Stecher, B. M., & Klein S. P. (2002). *Making sense of test-based accountability in education*. Rand Corporation. https://www.rand.org/content/dam/rand/pubs/monograph_reports/2002/MR1554.pdf

Martínez, M., Hernández, M. y Hernández, M. (2006). *Psicometría*. Madrid: Alianza Editorial.

Martínez-Riso, F. (2009). *Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado*. Revista Electrónica de Investigación Educativa, 11(2), 1-18. http://www.scielo.org.mx/pdf/redie/v11n2/v11n2a2.pdf

Sección 1.1

Las pruebas estandarizadas en contextos académicos

María Julieta Solórzano-Salas
y Flor Jiménez-Segura

Las pruebas estandarizadas de rendimiento académico se utilizan para evaluar la calidad y progreso educativo de un país, para tomar decisiones de política educativa y rendir cuentas a la sociedad.

Según Martínez-Rizo, (2009), en América Latina, México y Costa Rica comenzaron a emprenderse evaluaciones a gran escala en educación básica desde las décadas de 1970 y 1980, pero sólo Chile desarrolló un verdadero sistema de evaluación antes de 1990. En la última década del siglo XX y en la primera del XXI, casi todos los países lo han hecho.

El Consejo Nacional de Educación (2013) menciona que las evaluaciones estandarizadas de alto impacto del rendimiento escolar son, generalmente, iniciativas costosas que requieren del trabajo dedicado de profesionales altamente especializados durante largos períodos, desde la concepción inicial hasta la elaboración de reportes para diferentes actores educativos.

Entre los aportes asociados a las pruebas estandarizadas de alto impacto se deben mencionar los avances técnicos y la formación de especialistas de buen nivel; la creciente conciencia ciudadana del derecho a conocer los resultados de las evaluaciones y el que algunas autoridades educativas, comiencen a hacer uso de los resultados de las evaluaciones para la toma de decisiones (Martínez-Riso, 2010, párr. 15).

A través de los años la rendición de cuentas ha adquirido gran preponderancia debido a: el convencimiento de los países de que una mejor educación traerá un mayor desarrollo económico y una menor desigualdad social; la creencia de los gobernantes y de un sector de la sociedad de que la exposición pública de los resultados educativos sirve como motor o incentivo, por sí mismo, para el mejoramiento de la calidad educativa; y la exigencia creciente de la sociedad por conocer los

resultados de los programas de gobierno, entre los que se encuentran los educativos (Contreras-Roldán y Backhoff-Escudero, 2014).

Con respecto a las pruebas de logro con funciones de rendición de cuentas Contreras-Roldán y Backhoff-Escudero (2014) mencionan que se consideran evaluaciones de alto impacto por las consecuencias que están asociadas a ellas. Este tipo de evaluaciones tiende a responsabilizar a las instituciones (autoridades, organismos descentralizados y centros educativos) y a las personas (directores y docentes) de los resultados de aprendizaje del grupo de estudiantes.

Para Hamilton, Stecher y Klein (2002, p. 4-5) las pruebas estandarizadas de alto impacto pueden ser usadas para diferentes propósitos, tales como: proporcionar evidencia de la calidad de la educación para la revisión pública; proporcionar información al profesorado para mejorar las prácticas de enseñanza; monitorear el progreso del estudiante; premiar a estudiantes y docentes. También contribuyen a identificar las escuelas con mejores rendimientos, evaluar la eficacia de los esfuerzos de la reforma de un plan de estudios o

programas; juzgar si el estudiantado debe ser promovido a un grado superior; ubicarlos en los programas educativos especializados (por ejemplo: clases bilingües, estudiantes sobredotados o que necesiten clases de recuperación) y determinar si recibirán un diploma. La implementación de pruebas estandarizadas puede generar situaciones negativas que, con frecuencia, son usadas en los discursos para atacar sus usos y costos. Más en el contexto actual donde tiene un gran peso en la evaluación del desempeño de estudiantes, docentes y centros educativos; así como el uso inapropiado de los resultados y su excesivo peso en el diseño de las políticas públicas (Martínez-Riso, 2010, párr. 15).

Uno de los riesgos es que la presión por obtener mejores puntuaciones induce a que el profesorado reduzca la enseñanza del currículo a los contenidos que evalúan las pruebas, dediquen gran parte del tiempo de clase a la preparación para la prueba, permitan la copia entre estudiantes o eviten que estudiantes con menor rendimiento sean evaluados (Contreras-Roldán y Backhoff-Escudero, 2014 y Amrein y Berliner, 2003).

Esto implica que los centros educativos puedan verse obligados a abandonar un currículo enriquecedor por la presión que hay sobre el puntaje de las pruebas estandarizadas, empleando una cantidad de tiempo significativa para preparar al estudiantado.

Asimismo, el hecho de que el profesorado se compare con otros, a partir de los resultados de pruebas estandarizadas, puede darle un fuerte incentivo para mejorar el desempeño de sus estudiantes. No obstante, para Jones, Jones, Hardin, Chapman, Yarbrough y Davis (1999), tanto el estudiantado como el personal docente pueden sentirse abrumados por la presión para lograr un buen desempeño, por lo que la población estudiantil puede desarrollar ansiedad ante los exámenes y obtener un bajo desempeño; además pueden ver los resultados de las pruebas como única medida de su valor académico. Por su parte, el profesorado puede sentir presión para centrarse en las calificaciones obtenidas en lugar de los logros académicos más generalizados. Finalmente, siempre existe la amenaza al cumplimiento de la equidad, debido a la diversidad de contextos y ambientes de aprendizaje.

Referencias

Armein, A.L., & Berliner, D.C. (2003). *The effects of high-stakes testing on student motivation and learning*. Educational Leadership, 60(5), 32-38. http://eric.ed.gov/?id=EJ660880

Consejo Nacional de Educación. (2013). *Evaluaciones Estandarizadas del Rendimiento Escolar*. Arte Perú.

Contreras-Roldán, S., & Backhoff-Escudero, E. (2014). *Tendencias en el aprendizaje de la educación en México: Una comparación entre ENLACE, Excale y Pisa*. Nexos. http://www.nexos.com.mx/?p=22749

Gronlund N.E. & Linn R.L. (1990). *Measurement and Evaluation in Teaching*. Macmil lan.

Hamilton, L. S., Stecher, B. M., & Klein S. P. (2002). *Making sense of test-based ac countability in education*. Rand Corporation. https://www.rand.org/content/dam/rand/pubs/monograph_reports/2002/MR1554.pdf

Jones, M.G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T., y Davis, M. (1999). *The impact of High-Stakes Testing on Teachers and Students in North Carolina*. The Phi Delta Kappan, 81 (3), 199-203. www.jstor.org/stable/20439620

Martínez, M.R, Hernández, M.J. & Hernández, M.V. (2006). *Psicometría*. Alianza Editorial.

Martínez-Riso, F. (2009). *Evaluación formativa en aula y evaluación a gran escala: ha cia un sistema más equilibrado*. Revista Electrónica de Investigación Educativa, 11(2), 1-18. http://www.scielo.org.mx/pdf/redie/v11n2/v11n2a2.pdf

Martínez-Riso, F. (2010). *Las pruebas de rendimiento y sus consecuencias*. Altablero, (55). http://www.mineducacion.gov.co/1621/article-241800.html

Sección 1.2

Validez y psicometría

Eiliana Montero-Rojas

Ante el uso creciente de pruebas de alto impacto en contextos académicos, laborales y profesionales, resulta evidente la necesidad de generar mediciones de constructos educativos y psicológicos de una forma válida para alimentar la toma de decisiones. Tal es el caso de seleccionar, entre los aspirantes de primer ingreso a la universidad, a aquellos que poseen el perfil que permita un adecuado desempeño académico en la Universidad, o bien, las pruebas de certificación académica para evidenciar de manera objetiva los logros de aprendizaje para un programa educativo específico.

Los constructos, en general, son difíciles de operacionalizar y de medir, por ello, se han desarrollado diversas metodologías para su medición. La psicometría es, sin lugar a duda, una de las propuestas científicas más exitosas en términos de brindar herramientas útiles para emprender esta tarea de medición de constructos en educación y psicología (Martínez et al, 2006; Nunnally & Bernstein, 1995). Particularmente, ha contribuido a evidenciar las dos propiedades fundamentales de una "buena" medición: la validez y la confiabilidad (Nunnally & Bernstein, 1995; Martínez et al 2006).

(*) **Nota aclaratoria:** Este apartado es una adaptación hecha por la autora de su publicación: Referentes conceptuales y metodológico sobre la noción moderna devalidez de instrumentos de medición: implicaciones para el caso de personas con necesidades educativas especiales.

Validez

El concepto de validez sufrió, a partir de los años 1990, una importante transformación conceptual gracias al trabajo de Samuel Messick (1989a; 1989b). Ante la definición tradicional de validez que prácticamente se refería a una tautología, "un instrumento es válido si mide lo que con él se pretende medir", Messick provocó una pequeña revolución en la comunidad de la medición educativa al proponer la validez como el grado de propiedad de las inferencias e interpretaciones derivadas de los puntajes de los tests, incluyendo las consecuencias sociales que se derivan de la aplicación del instrumento (Padilla, Gómez, Hidalgo y Muñiz, 2006).

Una de las mayores contribuciones de Messick (1989a, 1989b) fue el concepto unitario de validez, el cual fue adoptado formalmente en los estándares de la American Educational Research Association, American Psychological Associaton y National Council on Measurement in Education (1999), y que puede considerarse el "ISO 9000" internacional en cuanto a estándares de calidad de las pruebas educativas y psicológicas.

Así, en lugar de buscar diferentes tipos de validez, Messick (1989a) indica que la idea es recolectar diferentes evidencias de validez de constructo, de acuerdo con los propósitos y usos de los instrumentos, dentro de las cuales se pueden citar las evidencias de:

- Contenido: son especialmente relevantes en pruebas educativas de conocimientos que miden resultados de procesos de aprendizaje formales.
- Predictivas: se refieren a instrumentos que intentan estimar comportamientos futuros, tal es el caso de las pruebas de admisión a la educación superior, en donde se busca que los puntajes se asocien a los rendimientos futuros de la población estudiantil.
- Concurrentes: se refieren a las asociaciones que deben presentan entre sí pruebas que intentan medir el mismo constructo.

Otro de los aportes de Messick (1989a, 1989b) se refiere a su reflexión en torno a que la validez no es una propiedad intrínseca de los instrumentos, sino que se define de acuerdo con el propósito de la medición, la población a la que va dirigida y el contexto específico de aplicación. Así, un instrumento puede exhibir un grado aceptable de validez para un propósito específico y para una población particular, pero no para otros. De manera que, para este autor la validez no es un rasgo dicotómico, al contrario, se trata de una cuestión de grado en que la prueba exhibe un grado aceptable de validez para ciertos usos específicos y con ciertas poblaciones.

Desde esta perspectiva, la validez psicométrica de un instrumento es solo una parte de la sistemática y rigurosa recolección de evidencia empírica, desde diferentes dimensiones, que debe emprenderse cuando se hace la pregunta: ¿Qué tan apropiadas son las inferencias generadas a partir de los puntajes de la prueba?

En primer lugar, las evidencias deben mostrar en qué medida el instrumento, como un todo, y los ítems que lo componen, representan adecuadamente al constructo teórico que se pretende medir y sus componentes.

Para esto es indispensable el manejo de los referentes teóricos y su correcta operacionalización.

Al respecto, Messick hace recapacitar a la comunidad de medición educativa cuando afirma que la persona constructora del instrumento no solo debe cumplir con lo científico-técnico sino también con lo ético: debe preocuparse por el uso que se da a los instrumentos y por las consecuencias derivadas de la aplicación de los mismos (Messick, 1989a y 1989b; Padilla et al, 2006). En torno a este punto, Messick (1989a; 1989b) analiza tres conceptos muy útiles:

- Variancia relevante al constructo: el objetivo es que las puntuaciones reflejen, precisamente, los diferentes niveles que toma el constructo de interés en los sujetos examinados. En otras palabras, se desea que la variabilidad observada entre las puntuaciones sea variabilidad verdadera, debida a las diferencias en el constructo que presentan los sujetos examinados.
- Variancia irrelevante al constructo: es generada por variaciones en los puntajes del instrumento que no representan diferencias reales en el constructo de interés, sino que son debidas a otros factores, por ejemplo, sesgos y errores de medición. Un ejemplo de variancia irrelevante sería un instrumento para medir habilidad cuantitativa en la resolución de problemas, en donde los enunciados de los ítems están cargados de vocabulario poco común y complejidad verbal. Es probable entonces que los puntajes de la prueba no solo reflejen habilidad cuantitativa, sino también conocimiento de vocabulario y comprensión verbal, introduciendo así un sesgo y una fuente de invalidez en la interpretación de los puntajes.
- Sub-representación del constructo: se presenta cuando un instrumento mide solamente un componente o dimensión de un constructo que es más complejo. Un ejemplo sería las pruebas de inteligencia tradicionales, las cuales miden ciertas dimensiones específicas del constructo, quedando otras sub-representadas. Otro caso serían aquellas pruebas de certificación que miden objetivos o habilidades establecidos en una tabla de especificaciones, pero no todas las áreas de los contenidos curriculares.

Considerando este planteamiento teórico las personas constructoras de pruebas deben documentar y demostrar el grado de propiedad de las inferencias que se generan a partir de los puntajes de los instrumentos. De acuerdo con Montero-Rojas (2013) este proceso de recolección de evidencias empíricas implica, normalmente, como primer paso la consulta a jueces expertos, aunque esta técnica no es suficiente para generar evidencia de validez sólida y suficientemente creíble. También se requiere una aplicación piloto del instrumento y un análisis psicométrico básico del instrumento y de los ítems que lo componen. Entre los métodos y modelos de análisis utilizados en este proceso la autora señala los siguientes:

- Análisis de factores exploratorio y confirmatorio
- Teoría Clásica de los Tests (TCT)
- Teoría de Respuesta a los Ítems (TRI)
- Modelo de Rasch

- Teoría G (Generalizabilidad)
- Análisis del funcionamiento diferencial del ítem (DIF por su acrónimo en inglés)
- Equiparación de puntajes

Para Montero-Rojas (2013) en la actualidad se cuenta con un vasto recurso de software computacional para realizar estos análisis psicométricos. Tanto la TCT como el análisis de factores exploratorio se encuentran disponibles en la gran mayoría de paquetes estadísticos de propósito general, incluyendo SPSS, Stata y SAS. El análisis factorial confirmatorio, al ser un caso particular de los modelos de ecuaciones estructurales (SEM, por sus siglas en inglés) están implementados en diversos paquetes de software especializado, siendo los más comunes, LISREL, EQS, Mplus y Amos, entre otros. En cuanto a software para modelos de TRI se pueden mencionar el BILOG y el IRTPRO. Mientras que para los modelos de Rasch existen dos paquetes especializados, el Winsteps y el ConQuest, que son los más utilizados en las ciencias sociales. Igualmente, la mayoría de estos procedimientos están implementados, con diversos niveles de sofisticación, en librerías de R, lenguaje y ambiente gratuito de programación para análisis estadístico, que se ha hecho muy popular en los últimos años.

También existe software disponible para DIF y equiparación de puntajes en los paquetes de softwares especializados. Ahondar más en la fundamentación, teoría y aplicación de estos modelos de análisis escapa al propósito de este documento, basta decir que se trata de temas de cierta complejidad técnica que exigen dedicación para su estudio y cabal comprensión, al igual que sólidas bases estadísticas y matemáticas (Montero-Rojas, 2013).

Confiabilidad

El concepto de confiabilidad se refiere a la precisión, consistencia y estabilidad en las repeticiones de una medida. Una definición más ilustrativa indica que un instrumento es confiable si aplicado en las mismas condiciones a los mismos sujetos produce los mismos resultados (Nunnally & Bernstein, 1995).

La confiabilidad es una condición necesaria, pero no suficiente para la validez, como lo explica Montero-Rojas (2013), un instrumento que exhibe un grado aceptable de validez, también debe poseer un grado aceptable de consistencia y estabilidad de sus resultados. No obstante, lo opuesto no es cierto, si un instrumento exhibe un alto nivel de confiabilidad no necesariamente es válido, esto porque puede estar midiendo con alta precisión y consistencia, pero sin garantía de que lo medido sea el constructo de interés (Babbie, 2010). La evidencia de confiabilidad es entonces un requisito necesario, pero no suficiente para la validez (Babbie, 2010).

Entre los indicadores de confiabilidad usados con más frecuencia en psicometría se incluyen el Alfa de Cronbach que es el resultado más importante de la Teoría Clásica de los Tests (TCT), el índice de discriminación, calculado en la TCT como la correlación ítem-total, así como la cantidad de error de medición y el tamaño de la función de información en Teoría de Respuesta a los Ítems (TRI) y el modelo de Rasch (Martínez et al, 2006; Muñiz, 2003; Prieto & Delgado, 2003).

Referencias

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *The Standards for Educational and Psychological Testing*. American Educational Research Association.

Babbie, E. (2010). *The Practice of Social Research*. Wadsworth.

Martínez, M. R., Hernández M.J. & Hernández, M.V. (2006). *Psicometría*. Alianza Editorial.

Messick, S. (1989a). *Meaning and values in test validation: The science and ethics of assessment*. Educational Researcher, 18 (2), 5-11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1989b). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). American Council on education and Macmillan.

Muñiz. J. (2003). *Teoría Clásica de los Tests*. Ediciones Pirámide.

Montero, E. (2013). *Referentes conceptuales y metodológicos sobre la noción moderna de validez de instrumentos de medición: implicaciones para el caso de personas con necesidades educativas especiales*. Actualidades en Psicología, 27 (114), 113-128. http://pepsic.bvsalud.org/pdf/apsi/v27n114/a08.pdf

Nunnally, J.C. & Bernstein, I.J. (1995). *Teoría psicométrica* (3ª ed). Editorial McGrawHill Latinoamericana.

Padilla J.P., Gómez, J., Hidalgo, M.D, & Muñiz, J. (2006). *La evaluación de las consecuencias del uso de los tests en la teoría de validez*. Psicothema, 18 (2), 307-312. http://www.psicothema.com/pdf/3215.pdf

Prieto, G. & Delgado A.R. (2003). *Análisis de un test mediante el modelo de Rasch*. Psicothema, vol. 15, nº 1, pp. 94-100. http://www.psicothema.com/psicothema.asp?id=1029

Sección 1.3

Evidencias de validez basadas en relaciones con otras variables

Luis Rojas-Torres

Una de las evidencias más relevantes asociadas a que las puntuaciones de una prueba realmente miden el constructo pretendido es que estas presenten las mismas redes de asociaciones con las mediciones de otros constructos u otras variables, que son establecidas en la teoría sobre el constructo de interés. Encontrar evidencias que apoyen las asociaciones teóricas de un constructo con otros proporciona un conjunto de evidencias de validez que se denominan basadas en relaciones con otras variables (AERA, APA y NCME, 2014). Estas indican que las puntuaciones de la prueba se comportan de manera similar al constructo pretendido.

Es esperable que una medición de un constructo determinado presente fuertes asociaciones con otras mediciones del mismo constructo o constructos similares y, contrariamente, muestre asociaciones débiles con constructos teóricamente distintos o lejanos; estos patrones de asociaciones se suelen denominar evidencia convergente y evidencia discriminante, respectivamente. Por ejemplo, es esperable que dos pruebas de conocimiento matemático del mismo nivel educativo presenten altas asociaciones; en cambio, se espera que estas pruebas se asocien en menor medida con las mediciones de conocimientos históricos (aunque no se espera una correlación nula, ya que ambos tipos de conocimientos pueden ser promovidos por elementos en común como acceso a un sistema educativo de calidad).

Por otro lado, los usuarios de las pruebas pretenden que las puntuaciones de la prueba estén correlacionadas con variables asociadas a los propósitos de la prueba, las cuales se denominarán variables criterio. Las evidencias de validez basadas en la relación con una variable criterio son muy relevantes en pruebas de altas consecuencias, debido a que brindan una justificación evidente del uso de la prueba. Por ejemplo, en las pruebas de selección de estudiantes de un programa de estudios o en las pruebas de selección de empleados de una empresa es esperable que sus puntuaciones se asocien con características buscadas en los seleccionados, como notas de los cursos, en un caso, o niveles de producción, en el otro. Para las pruebas estandarizadas en el contexto costarricense se recomiendan l os siguientes estándares.

1. Las puntuaciones de la prueba deben estar altamente asociadas con pruebas consolidadas que evalúen en el mismo constructo. El cumplimiento de este estándar indica de primera entrada que las puntuaciones presentan la dirección esperada, es decir, a mayor nivel en el constructo, mayor puntuación en la prueba.

Para evaluar este estándar se puede recurrir al uso de la técnica estadística de las correlaciones. Por ejemplo, en Cassady y Johnson (2002) se desarrolló una nueva escala de medición de ansiedad ante los exámenes y se evaluó su correlación con otra muy consolidada que había sido utilizado desde los años sesenta.

2. Se debe contar con un abordaje teórico actualizado sobre las redes de asociaciones que involucran al constructo evaluado. Un aspecto básico del enfoque científico asociado a la medición de constructos es el carácter asociacionista, es decir, la mayoría de los constructos se definen con base en la relación con otros. Por tanto, un aspecto que indica el conocimiento del constructo evaluado es la claridad sobre los constructos con los que se relaciona y la forma en que se relaciona.

Por otro lado, la propuesta de evidencias discriminantes se vuelve relevante cuando se quiere evidenciar que la prueba mide un constructo distinto de otro con el que se suele confundir, por ejemplo, conocimiento matemático y razonamiento matemático.

3. Se debe generar evidencias de que la medición realizada reproduce las asociaciones establecidas teóricamente. Estas asociaciones son una evidencia fundamental de validez, ya que indican si la medición realmente reproduce el comportamiento que teóricamente tiene el constructo. Esta es una de las principales formas de determinar si realmente se está evaluando el constructo pretendido.

Por ejemplo, en Reilly y Neumann (2013) se presentan varias teorías que explican por qué los hombres y las mujeres presentan diferentes niveles en la habilidad espacial, por lo que la obtención de un resultado contrario en un test de habilidad espacial constituye un indicador de que la prueba no está midiendo el constructo pretendido. Al igual que en los estándares previos, el estudio de este puede ser desarrollado mediante técnicas correlacionales.

Por otro lado, la institución usuaria de las pruebas no puede desentender de la calidad de la medición de las variables, puesto que un descuido en las mediciones externas puede llevar a conclusiones erróneas.

4. El uso de una prueba con altas consecuencias debe estar claramente justificado. Las instituciones están en la obligación ética de informarle a los examinados por qué utilizan esa prueba. Por lo general, la razón se asocia a la utilidad de sus resultados para los propósitos de la institución usuaria o de otras instituciones. Esta razón debe ser explicada teóricamente y apoyada empíricamente (lo cual se profundiza en el siguiente estándar).

Por ejemplo, una razón por la que un programa de estudios en matemática utiliza una prueba de razonamiento matemático para la selección de sus estudiantes es debido a que se espera desde la teoría que, a mayor rendimiento en la prueba, mejor desempeño en la carrera.

5. Se deben generar evidencias empíricas sobre la justificación del uso de una prueba de altas consecuencias para demostrar que está cumpliendo su propósito. Para generar estas evidencias, primeramente, hay que construir las variables asociadas a las consecuencias del test. Por ejemplo, una variable criterio para una prueba de selección puede ser el desempeño de los admitidos, mientras que para una prueba de certificación, una variable criterio sería el dominio de la habilidad o conocimiento certificado en los contextos que se requiera. Luego, la asociación de las puntuaciones de la prueba con las variables criterio se puede estudiar mediante modelos correlacionales, que van desde las correlaciones simples hasta modelos donde se controlan los efectos de otras variables que también se asocian con la variable criterio.

Por ejemplo, en Jiménez y Morales (2010) y Rojas (2013) se expone que la prueba de ingreso a la Universidad de Costa Rica se asocia positivamente con el rendimiento académico universitario, mientras que en Rojas (2014) se muestra que a mayores puntuaciones en una prueba para ingreso a carreras universitarias se presentan mayores probabilidades de aprobación en los cursos de estas.

6. Se deben realizar mediciones frecuentes de las variables criterio, que cumplan con los criterios de calidad de las mediciones. Como se indicó en el estándar 3, es fundamental garantizar la calidad de la medición de las variables criterio. Por otro lado, se agrega el elemento de la frecuencia, ya que el proceso de la validación es un proceso constante y, además, el ente usuario de las pruebas debe garantizar a sus examinados la pertinencia que tiene el uso de dichas pruebas.

7. El uso de una prueba en la toma de decisiones está limitado a aquellas en las que se posee un cúmulo significativo de evidencias de validez. Los entes usuarios son responsables de garantizar a los examinados que los usos son apropiados. El tomar decisiones cuando no se tiene certeza de la pertinencia de las puntuaciones de la prueba, conlleva un agravante ético que puede llevar a decisiones equivocadas.

Referencias

Cassady, J. & Johnson, R. (2002). *Cognitive Test Anxiety and Academic Performance.* Contemporary Educational Psychology 27, 270-295. https://www.sciencedirect.com/science/article/abs/pii/S0361476X0191094X

Jiménez, K. & Morales, E. (2009-2010). *Validez predictiva del Promedio de Admisión de la Universidad de Costa Rica y sus componentes.* Actualidades en Psicología, 23, 21-55. http://pepsic.bvsalud.org/pdf/apsi/v23-24n10-111/a02.pdf

Reilly, D. & Neumann, D. L. (2013). *Gender-Role Differences in Spatial Ability: A Me ta-Analytic Review.* Sex Roles, 68(9) 5-31. DOI: https://doi.org/10.1007/s11199-013-0269-0

Rojas, L. (2013). *Validez predictiva de los componentes del promedio de admisión a la Universidad de Costa Rica utilizando el género y el tipo de sexo como variables control.* Revista Actualidades Investigativas en Educación, 13 (1), 1- 24. https://www.scielo.sa.cr/pdf/aie/v13n1/a03v13n1.pdf

Rojas, L. (2014). *Predicción de la reprobación de cursos de matemática básicos en las carreras de Física, Meteorología, Matemática, Ciencias Actuariales y Farmacia.* EDUCARE, 18(3) 3-15. DOI: https://doi.org/10.15359/ree.18-3.1

Sección 1.4

Validez y Equidad

Lucrecia Alfaro-Rojas

En los apartados anteriores se ha señalado la importancia de la uniformidad de los materiales y los procedimientos, así como los análisis de resultados, para lograr una medición válida y precisa del constructo meta que fundamente la toma de decisiones en función de los objetivos de una prueba estandarizada. Al garantizar lo anterior, el equipo a cargo de realizar la medición se asegura de que cada persona examinada tenga las mismas oportunidades de mostrar su nivel real en el constructo meta y que así los resultados obtenidos por todas las personas sean comparables entre sí (Messick, 1995).

La uniformidad en las condiciones de aplicación de una prueba, es decir, la estandarización, apunta a la recopilación de información específica que permita hacer inferencias con respecto al nivel de cada persona en el constructo medido (Georgia Department of Education, 2008). La estandarización lleva, además, a que todas las personas evaluadas tengan igualdad de oportunidades para mostrar su nivel "real" en el constructo mediante la prueba. Este concepto es especialmente importante cuando se trata de pruebas de alto impacto, cuya aplicación afecta directamente el curso de la vida de las personas.

El uso de pruebas estandarizadas suele asociarse a una medición rigurosa y de calidad, pues se espera certificar la habilidad de las personas examinadas en un dominio específico o seleccionar a aquellas con ciertas características para un fin particular. Esta percepción de las pruebas como instrumentos que permiten ordenar y clasificar de manera objetiva a las personas, se puede cuestionar cuando no son adecuadamente aplicadas o las inferencias e interpretaciones no son apropiadas, ya que vulneran los derechos de las personas evaluadas. Un caso ilustrativo sería cuando se aplica una prueba validada con baremos españoles en una muestra costarricense y se califica conforme con los criterios establecidos en el contexto de origen. En este ejemplo, se estaría partiendo de que las características de ambas poblaciones son equivalentes, cuando no necesariamente lo son.

Al respecto, Madaus (1994) afirma que la medición es una práctica que lleva implícita el ejercicio del poder por parte de ciertos grupos que eligen o excluyen a personas de acuerdo con criterios establecidos; por ello se debe asegurar que los criterios relacionados con pertinencia, precisión, relevancia y calidad respondan a intereses objetivos y no a intereses de grupos específicos. Según este mismo autor, cuestionar si el diseño de un instrumento puede producir inequidades o ventajas para ciertos grupos (sesgos, valores y asunciones), es una tarea que siempre debe asumir toda persona que trabaje en el desarrollo de una prueba.

La igualdad y la equidad son conceptos distintos, aunque están intrínsecamente ligados. Las diferencias entre individuos son una característica de todos los grupos humanos, ya que no todos los miembros de la sociedad son iguales en términos simbólicos ni materiales (Malaise, Friant y Demeuse, 2014), lo que, en el contexto de la medición podría interpretarse como diferencias en cuanto a la valoración que se hace de las personas pertenecientes a ciertos subgrupos y a las oportunidades de fomentar y expresar sus habilidades. Son estas diferencias las que dan pie a la necesidad de garantizar un tratamiento equitativo y brindar oportunidades a los diferentes subgrupos que componen una población, de manera que se propicie la justicia.

La equidad en el proceso de evaluación implica medir correctamente el nivel de

cada persona en el constructo medido, sin que este se vea alterado por variables como origen étnico, condición de salud, discapacidad, estrato económico de procedencia o género, entre otros. Para lograr esto, Madaus (1994) propone que el contenido de una prueba, en tanto a que responde a valores propios de las personas responsables de su construcción, deben ser validados en el contexto de las diferentes personas examinadas, para asegurar que la prueba incorpore aquellos valores, conocimientos y destrezas que sean más generales y compartidos por las distintas poblaciones en el contexto de la medición por realizar.

La diversidad inherente a los subgrupos que componen una población puede llevar a que sus condiciones asociadas al origen étnico, las manifestaciones culturales o el manejo de una lengua distinta a la usada en el instrumento de medición, incidan en lo que se conoce como "varianza irrelevante al constructo" (Messick, 1989a; 1989b; 1995). Como se señaló en el apartado anterior, esta variancia irrelevante se presenta cuando características propias de una persona evaluada, como del contexto en el que ha crecido, por ejemplo, nivel educativo de los padres o nivel socioeconómico, que no tienen relación con el constructo por medir, llevan a interpretaciones erróneas sobre su nivel en el constructo (Cawthon, Winton, Garberoglio y Gobble, 2011; AERA, APA y NCME, 2014).

Cuando los puntajes obtenidos por un grupo de personas examinadas en una misma prueba sean producto de otros factores y no de una diferencia real en el constructo, se estaría atentando contra la validez de la interpretación de los puntajes obtenidos para ese grupo de personas examinadas. Un ejemplo de esto es la situación que pueden enfrentar las personas en condición de discapacidad, cuando en contextos de evaluación no se les garantiza la completa accesibilidad a la prueba. En este tipo de situaciones, si no se cuenta con un protocolo para la provisión de apoyos adecuados para la realización de la prueba, se podría generar una interpretación errónea de su nivel de logro. Por ello, siempre que se diseñe o utilice un instrumento de medición, dichos factores deben ser previamente identificados y tomarse todas las precauciones para que estos no afecten la medición.

El concepto de adecuaciones se torna relevante en el contexto de las pruebas estandarizadas, al ser un recurso que apoya la equidad en los procesos de medición cuando se trata de personas en condición de discapacidad que, de otro modo, no podrían alcanzar su máximo desempeño en una prueba. Las adecuaciones se han definido tradicionalmente como cambios en las instrucciones o en las prácticas de medición que reducen el impacto de la discapacidad de una persona en su interacción con el material de la prueba. Pueden incluir cambios en las instrucciones, el tiempo asignado para completar una tarea, el método de respuesta, los materiales o el equipo que apoya la interacción con el material relativo a la prueba (Ketterlin-Geller y Johnstone, 2006).

Para efectos de este documento, el concepto de adecuaciones será reemplazado por el de "apoyos", entendido como todos los cambios o recursos que garantizan la equidad a la totalidad de personas evaluadas. De esta manera, se garantiza una participación real y efectiva de la población en condición de discapacidad en todo el proceso de evaluación, eliminando las barreras presentes en el entorno que interfieren con la correcta

medición del constructo. Además, estos apoyos garantizan una medición intacta del constructo mientras que, por otra parte, existen prácticas que deben ser evitadas debido a que sí cambian, alteran o reducen las expectativas de lo que se mide, estas son las llamadas modificaciones (Stone, Cook, Laitusis y Cline, 2010).

Las modificaciones consisten en cambios en el proceso de aplicación o en los materiales que sí dan una ventaja efectiva a las personas que las reciben. Un ejemplo sería cuando se aplica una versión del instrumento con menos ítems o ítems de menor dificultad, o cuando la fórmula modificada tiene ítems de selección con menos opciones (Georgia Department of Education, 2008). Estas modificaciones sí tienen implicaciones para la validez de las inferencias realizadas a partir de los resultados de una prueba, ya que imposibilitan hacer comparaciones entre las personas evaluadas.

Ante la necesidad de brindar apoyos a una persona para la realización de una prueba en condiciones de equidad, debe siempre revisarse si estos colocan a la persona examinada en una condición de ventaja, o bien, si estos eliminan las barreras que obstaculizan una medición en igualdad de condiciones. Para lograr esto, se proponen los siguientes estándares que contribuyen a garantizar la calidad de la medición en la población de personas examinadas (American Educational Research Association, American Psychological Association y National Council on Measurement in Education, 2014).

1. Desde la fase de diseño del instrumento se deben eliminar aquellas características de la prueba que, sin ser relevantes para la medición del constructo meta, podrían constituir barreras al estar afectadas por particularidades de tipo lingüístico, comunicativo, cognitivo, cultural, físico o de otro tipo.
2. Las personas a cargo del desarrollo de una prueba deben incluir en la muestra utilizada para los estudios piloto, una representación de subgrupos considerados relevantes, de manera que se descarte el funcionamiento diferencial de los ítems para estos conglomerados.
3. La totalidad de personas examinadas deben recibir un tratamiento equivalente en el proceso de aplicación y calificación de la prueba, sin el otorgamiento de ventajas a ningún subgrupo. Este tratamiento equivalente no es discutible, pues permite una correcta medición de la totalidad de las personas y, por ende, una comparación del nivel de las diferentes personas en el constructo medido. En este sentido, los apoyos son considerados ajustes razonables al no imponer a los constructores o aplicadores de las pruebas cambios que alteren la correcta medición del constructo.
4. Las personas a cargo de la construcción y aplicación de una prueba son las responsables de desarrollar y proveer apoyos siempre que sean apropiados, factibles y reduzcan la varianza irrelevante al constructo.
5. Las personas a cargo de la construcción y aplicación de una prueba son los responsables de generar el protocolo de aplicación sin afectar la estandarización y de monitorear la correcta implementación de los apoyos.
6. Debe garantizarse un tratamiento correcto a la población para evitar improvisaciones que podrían afectar la medición del constructo, perjudicando a ciertos grupos o personas examinadas.

7. Debe darse un tratamiento científico a los resultados de la totalidad de la población, sin subestimar las capacidades o posibilidades de éxito de ciertos subgrupos en el contexto de la medición realizada.
8. Las personas a cargo de la construcción y adaptación de la prueba deben obtener las evidencias de la validez de las interpretaciones de los puntajes, siempre que los tamaños de muestra así lo permitan.
9. La construcción de pruebas estandarizadas debe apoyarse en los fundamentos técnicos y metodológicos más actualizados que garanticen la calidad de la medición, sin descuidar las especificaciones atinentes a las características de los distintos subgrupos de la población. Por ello, es necesaria una revisión continua de los procedimientos, en aras de incorporar a sus prácticas los más recientes avances tecnológicos y metodológicos en materia de accesibilidad.
10. Los apoyos deben permitir una participación plena en la evaluación para que las personas examinadas demuestren su nivel de conocimientos o habilidades; por lo tanto, solo se deben garantizar los estrictamente necesarios. De lo contrario, el proveer apoyos no requeridos puede interferir e impactar de forma negativa en el desempeño del sujeto.
11. Los apoyos deben responder a las necesidades particulares de cada persona para facilitar su independencia, por lo que deben estar justificadas y documentadas individualmente.
12. Los apoyos brindados a las personas examinadas no deben ser introducidos por primera vez en el contexto de la evaluación, deben ser congruentes con los implementados durante su proceso educativo.

Referencias

American Educational Research Association, American Psychological Association y National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Cawthon, S.W., Winton, S.M., Garberoglio, C.L. y Gobble, M.E. (2011). *The Effects of American Sign Language as an Assessment Accommodation for Students Who Are Deaf or Hard of Hearing*. Journal of Deaf Studies and Deaf Education, 16 (2), 198-211. DOI: 10.1093/deafed/enq053

Georgia Department of Education (2008). *Accommodations Manual: A Guide to Selecting, Administering, and Evaluating the Use of Test Administration Accommodations for Students with Disabilities*. Georgia Department of Education.

Ketterlin-Geller, L.R., y Johnstone, C. (2006). *Accommodations and Universal Design: Supporting Access to Assessments in Higher Education*. Journal of Postsecondary Education and Disability, 19 (2), 163-172. https://eric.ed.gov/?id=EJ844632

Madaus, G. (1994). *A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy*. Harvard Educational Review, 64 (1), 76-96. https://doi.org/10.17763/haer.64.1.4q87663r0j76rwv1

Malaise, S., Friant, N., & Demeuse, M. (2014). *The equity aspect within the framework of the assessment of the quality of Higher Education: Developing indicators to identify students with a higher risk of failure at university with a view to improving equality of chances of success*. Revista de Docencia Universitaria, 12 (2), 119-141. https://dialnet.unirioja.es/servlet/articulo?codigo=4845366

Messick, S. (1989a). *Meaning and values in test validation: The science and ethics of assessment*. Educational Researcher, 18 (2), 5-11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1989b). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). American Council on education and Macmillan.

Messick, S. (1995). *Standards of validity and the validity of standards in performance assessment*. Educational Measurement: Issues and practice, 14 (4), 5-8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Stone, E.; Cook, L. Laitusis, C.C. y Cline, F. (2010). *Using Differential Item Functioning to Investigate the Impact of Testing Accommodations on an English-Language Arts Assessment for Students who are Blind or Visually Impaired*. Applied Measurement in Education, 23, 132-152. https://doi.org/10.1080/08957341003673773

Sección 1.5

La ética en pruebas estandarizadas en Costa Rica

Pablo Mena-Castillo

La palabra ética proviene del griego ethos, que significa costumbres, conducta habitual, usos, comportamiento y carácter. Lejos de toda imposición o visión legalista, la ética se fundamenta en la autonomía y la responsabilidad de cada persona, en su deber de autorregularse. La ética está presente en la persona autónoma, merecedora de respeto y trato digno, inmersa en un particular entorno social. Esta se ocupa justamente de formar y preparar al ser humano como persona con valores para que pueda construir y ejercer su condición humana en el mundo.

En toda evaluación la ética debe ser un principio rector que guíe cada una de las fases del proceso evaluativo, porque en ella se trabaja con personas, quienes deben ser tratadas con respeto, dignidad, integridad, honradez y salvaguardando sus derechos humanos en todo momento. En la evaluación se deben respetar las diferencias culturales, costumbres, creencias, género, origen étnico, capacidad, edad, orientación sexual, lengua y prácticas religiosas de los evaluados.

Por su parte, las personas evaluadas deben tener plena libertad para decidir voluntariamente su participación en la evaluación; quienes tienen derecho a la información oportuna y fiable acerca de lo que se espera de ellos y además de saber qué pueden esperar del evaluador y de los resultados de la evaluación. Los principios éticos como el derecho a la privacidad y la confidencialidad de la persona evaluada, deben estar inmersos en toda evaluación. A continuación, se presentan algunos estándares éticos que deben privar en las pruebas estandarizadas de Costa Rica, los cuales se han fundamentado en los aportes de Cohen y Swerdlik (2001), Santos (1999), Moreno (2011), Muñiz (2003) y Ormart (2004).

Del equipo evaluador

1. Tener la formación profesional que los acredite como evaluadores y ser conscientes de la necesidad de la actualización permanente, conforme con los avances científicos y profesionales en evaluación.
2. Respetar la dignidad de las personas evaluadas en todas sus dimensiones, por lo que, las pruebas estandarizadas no deben causarles daño físico, moral, ético, social o psicológico.
3. Evitar las etiquetas a las personas evaluadas en términos despectivos basándose en los resultados de las pruebas.
4. Garantizar la eliminación de sesgos o discriminaciones en las pruebas debidas a edad, género, etnia, nacionalidad, hábitat, religión, orientación sexual, discapacidad, idioma, nivel socioeconómico, clase social o cualquier otra característica sociodemográfica de las personas examinadas.
5. Proporcionar evidencias del uso apropiado y de las interpretaciones de los resultados de la prueba.
6. Considerar para la toma de decisiones, conclusiones o recomendaciones las evidencias aceptables de validez, de acuerdo con los propósitos de la prueba.

7. Velar por el resguardo y seguridad de las pruebas, así como por los resultados y datos personales de las personas examinadas para evitar que audiencias no cualificadas usen incorrectamente esa información.
8. Facilitar los resultados de las pruebas a las personas evaluadas o a otras audiencias cualificadas que los requieran.
9. Garantizar que el diseño de la prueba, su aplicación y calificación sean acordes con los procedimientos técnicos para garantizar una medición objetiva y así no privilegiar los intereses propios, estatus social o político de grupos particulares sobre los intereses de todos los implicados en la evaluación.
10. Evitar un tratamiento desfavorable a las personas evaluadas en la toma de decisiones.
11. Rechazar gratificaciones, regalos o favores que puedan perjudicar o influir en los juicios o actividades de las pruebas.
12. Estar dispuesto a suspender en cualquier momento una prueba, si se observa que su continuación puede afectar la evaluación, por ejemplo, la pérdida o robo de un ejemplar de la prueba.
13. Comunicar a los superiores institucionales y al resto de su equipo, cuando se presenten posibles conflictos de intereses entre ellos y las pruebas, por ejemplo, cuando familiares de primer grado son examinados.
14. Aplicar la prueba bajo las mismas condiciones para que todas las personas examinadas puedan resolver la prueba en un ambiente adecuado y sin presiones coercitivas.

De la persona evaluada

1. Participar de manera voluntaria en las pruebas estandarizadas.
2. Tener derecho a saber por qué se le está aplicando una prueba, cómo se usarán los resultados y qué información, si es que existe alguna, será revelada y a quién.
3. Tener derecho a conocer anticipadamente los contenidos y los procesos que serán medidos.
4. Tener derecho a ser informado, en un lenguaje que pueda comprender, de los resultados, las razones de la calificación obtenida, las conclusiones y recomendaciones de la prueba realizada.
5. Tener la garantía de que los resultados serán confidenciales y solo serán divulgados a terceros que harán un uso apropiado de esos registros.
6. Dar voluntariamente su consentimiento informado para el uso la información de los resultados de la prueba.
7. Recibir la devolución de la calificación de la prueba en los plazos definidos.

Referencias

Cohen, R. J. y Swerdlik, M. (2001). *Pruebas y evaluación psicológicas.* Introducción a las pruebas y a la medición. McGRAW-HILL.

Moreno, T. (2011). *Consideraciones Éticas en la Evaluación Educativa.* Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 9 (2), 130 - 144. https://revistas.uam.es/reice/article/view/4712/5146

Muñiz, J. (2003). *Teoría Clásica de los Tests.* Ediciones Pirámide.

Ormart, E. (2004). *La ética en la evaluación educativa.* Etic@net, II (3), 103 - 112. https://dialnet.unirioja.es/servlet/articulo?codigo=6871774

Santos, M.A. (1999). *Metaevaluación: rigor, mejora, ética y aprendizaje.* En B. Jiménez (ed.) Evaluación de programas, centros y profesores (pp. 265 - 298). Editorial Síntesis.

PARTE 2

Etapas de la construcción, validación y aplicación de pruebas estandarizadas de alto impacto

Como se ha descrito en el primer apartado, por los principios que dictan la psicometría, la evaluación educativa y la ética, las pruebas de alto impacto podrán ser utilizadas para la toma de decisiones, sólo cuando se cuente con suficientes evidencias de su calidad técnica. Además de los aspectos técnicos-científicos, los equipos responsables y usuarios de pruebas estandarizadas deben velar por el adecuado resguardo, uso y aplicación, lo que incluye consideraciones sobre la seguridad y confidencialidad de estas pruebas y los bancos de ítems que los componen; así como garantizar la transparencia e imparcialidad en su manejo. Las etapas críticas para la construcción, aplicación y validación de las pruebas de alto impacto se resumen en:

1. La definición de los conceptos, constructos o contenidos a medir y su respectiva operacionalización.
2. Propósitos y objetivos de las pruebas, especificaciones y restricciones de su uso.
3. El diseño de la prueba.
4. La generación de la tabla de especificaciones.
5. Construcción y juzgamiento de ítems para pruebas estandarizadas.
6. La prueba piloto y el análisis psicométrico.
7. El desarrollo de un banco de ítems.
8. El ensamblaje de la prueba.
9. Entrega de calificaciones y reporte de resultados.

Posteriormente se presenta la última sección denominada: Recomendaciones para la aplicación de pruebas estandarizadas, compuesta de dos apartados. El primero llamado: *Buenas prácticas en la aplicación de pruebas estandarizadas y el segundo: Aplicación de pruebas psicológicas en contextos de certificación y selección. Con ello, se realiza el cierre de este documento.*

Sección 2.1

Definición de los constructos y su operacionalización: Etapa 1

Vanessa Smith-Castro

Muchos fenómenos evaluados mediante pruebas estandarizadas como el desempeño, la habilidad, los conocimientos, las destrezas, la idoneidad mental, la depresión, la ansiedad y la inteligencia, no tienen una existencia concreta similar a las entidades físicas que se ofrecen a nuestros sentidos. Son conceptos que sobrepasan la observación empírica y muchas veces expresan entidades teóricas. A tales conceptos se les denomina constructos, conceptos no observacionales, variables latentes o rasgos latentes (Bollen, 2002, Montero, 2008).

Para saber en qué medida un rasgo latente está presente en las personas examinadas, el equipo de constructores debe establecer las definiciones operativas para determinar su presencia o ausencia (o su magnitud), lo que les permitirá desarrollar pruebas adecuadas para tomar las decisiones pertinentes. Esto se conoce como operacionalización de variables.

Por lo general, estas operaciones son mediciones que involucran registros numéricos (como los minutos que invierten las personas en resolver una tarea o sus puntuaciones en un examen de matemáticas). De hecho, las escalas o pruebas psicométricas se definen, comúnmente, como un conjunto de indicadores operacionales de un rasgo o dominio de comportamiento (Tornimbeni, Pérez & Olaz, 2008).

En la medida en que operacionalizar implica definir la presencia, ausencia o magnitud de los fenómenos en términos numéricos, es preciso que el equipo constructor posea tanto la idoneidad profesional y académica como el dominio teórico y metodológico de los principios de la medición y psicometría, especialmente sobre los conceptos de validez y confiabilidad.

También resulta central informarse e informar claramente sobre el concepto o constructo que se pretende medir, la forma de medirlo y justificar su pertinencia para los fines de los procesos de evaluación. Esto significa que el equipo constructor de pruebas es responsable de presentar la documentación teórica y técnica necesaria para justificar la representatividad de los constructos medidos y la relevancia de los mismos para las poblaciones en las que se va a utilizar la prueba.

Normalmente, la pertinencia y utilidad de los constructos se hace visible en la fundamentación teórica de las pruebas. Este marco teórico describe un conjunto de proposiciones o constructos (conceptos) interrelacionados que representan una visión sistemática de las relaciones entre las variables con el propósito de explicar y predecir fenómenos y puede presentarse en forma de taxonomías, tipologías o modelos.

Un sustento teórico sólido define los constructos de manera sustantiva y operativa, explica lógicamente relaciones entre variables, posibilita realizar inferencias causales, genera predicciones precisas, es parsimonioso, pero aplicable a fenómenos complejos y es generalizable a universos claramente definidos.

El fin último de esta primera etapa en la construcción de pruebas estandarizadas de alto impacto es presentar los argumentos teóricos, sustentados en investigaciones empíricas previas, que informan sobre la forma elegida para medir los contenidos o constructos de interés. Al respecto, Villarreal, Alfaro-Rojas y Brizuela (2015, 23) recomiendan tres prácticas específicas en esta etapa particular:

a) Describir el tipo de conocimientos o habilidades que la prueba pretende medir.
b) Documentar el grado en que la prueba representa el constructo que se intenta medir.
c) Recopilar y sistematizar la evidencia empírica que permita estimar si la evaluación puede cumplir el propósito previsto.

Para estos fines el equipo constructor puede nutrirse de alianzas estratégicas con la academia, para que las pruebas sean objeto de investigación científica, para cumplir con el estándar de proporcionar la teoría y la evidencia empírica necesaria del constructo medido en la prueba.

Referencias

Bollen, K.A. (2002). *Latent Variables in Psychology and the Social Sciences.* Annual Review of Psychology, 53, 605-634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Montero, E. (2008). *Escalas o Índices para la medición de constructos: El dilema del analista de datos.* Avances en Medición, 6, 15-24. https://dialnet.unirioja.es/servlet/articulo?codigo=2981180

Tornimbeni, S., Pérez, E. & Olaz, F. (2008). *Introducción a la Psicometría.* Paidós.

Villarreal, M.P., Alfaro-Rojas, L. & Brizuela, A. (2015). *Construcción de pruebas estandarizadas en el ámbito de la medición educativa y psicológica.* Serie Cuadernos Metodológicos del Instituto de Investigaciones Psicológicas. Universidad de Costa Rica. http://iip.ucr.ac.cr/sites/default/files/contenido/cuaderno.pdf

Sección 2.2

Propósitos de las pruebas, especificaciones y restricciones de su uso: Etapa 2

Vanessa Smith-Castro

Una vez definido el constructo a medir, es necesario que tanto los equipos responsables de la construcción y aplicación como las personas evaluadas identifiquen claramente la finalidad del instrumento y estimen si resulta apropiado para medir el constructo, de acuerdo con su marco de referencia teórico. Para ello deben informarse e informar sobre el alcance y capacidad del instrumento para cubrir el dominio comportamental por medir y, principalmente, especificar las restricciones de su uso.

Los instrumentos son operacionalizaciones de constructos que sólo tienen sentido en relación con otros constructos en el marco de las redes conceptuales que proporcionan los modelos teóricos, y que pueden existir variaciones conceptuales importantes de teoría en teoría (Martínez, Hernández, y Hernández, 2006). Por ello, un primer paso es que todos los involucrados en el proceso de evaluación comprueben la idoneidad del modelo teórico en función de los propósitos de la prueba.

En un segundo paso todos los involucrados deben informarse e informar sobre el uso y sus limitaciones, los propósitos y, sobre todo, de los grupos poblacionales para los cuales han sido desarrollados y validados los instrumentos. Como se puede observar, el punto clave en esta etapa es describir claramente el para qué de las pruebas, qué es lo que miden, cuáles decisiones pueden tomarse con estas pruebas y quiénes pueden ser evaluados con ellas.

En lo que respecta al para qué de las pruebas es importante mencionar que las personas usuarias de la prueba que intenten utilizarlas para un propósito distinto del original o en una población en la que no ha sido utilizada con anterioridad, deben proporcionar evidencias empíricas que apoyen el uso de la prueba para los nuevos propósitos y en las nuevas poblaciones (AERA, APA & NCME, 2014).

Un tercer paso concierne a la explicación de los posibles efectos negativos en el uso de los resultados de una prueba, por ello, el equipo evaluador y las personas examinadas deben comprometerse a explicar con evidencia científica estos posibles efectos a los responsables políticos para que puedan tomar las decisiones pertinentes. En general se recomienda advertir sobre el uso inapropiado de las interpretaciones y estar vigilante ante las consecuencias negativas de dichas interpretaciones para grupos poblacionales específicos e individuos particulares (AERA, APA & NCME, 2014).

En lo que respecta a los grupos poblacionales involucrados es importante mencionar que el equipo evaluador debe describir claramente el público meta (stakeholders) general de sus pruebas. Esto implica una clara descripción de las personas examinadas y usuarias de la prueba y de las autoridades o instituciones que podrían emplear los resultados para tomar decisiones (AERA, APA & NCME, 2014).

Debido a que las decisiones afectan las oportunidades de vida individuales es importante ser particularmente cuidadosos con las personas examinadas, se proponen las siguientes recomendaciones para el equipo evaluador, de acuerdo con los estándares propuestos por AERA, APA & NCME (2014).

1. Cumplir rigurosamente con las reglas explícitas sobre la determinación de las personas examinadas que pueden ser evaluadas con las pruebas y quienes pueden estar exentos. Estas políticas deben ser aplicadas de manera uniforme para asegurar una interpretación

válida y apropiada de las puntuaciones y de las comparaciones entre subgrupos.

2. Prestar la atención adecuada a las personas examinadas que hablan diferentes idiomas. Si un estudiante carece de dominio del idioma empleado en la prueba, entonces esta puede convertirse en un instrumento de dominio del idioma, sin ser este el objetivo. En tales casos, no deben ser empleadas porque las personas examinadas no pueden entender las instrucciones o la prueba en general.

3. Tomar medidas para asegurar que las inferencias de las calificación de las pruebas reflejan con precisión el constructo, independientemente de las características individuales asociadas a la discapacidad que son ajenas a la intención de la medición, por lo que es necesario desarrollar todo un programa para proporcionar los apoyos necesarios que permitan remover las barreras que impidan la medición del constructo, sin que estos apoyos comprometan la validez en las inferencias de los resultados.

Referencias

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

Martínez, M.R, Hernández, M.J. & Hernández, M.V. (2006). *Psicometría*. Alianza Editorial.

Sección 2.3

Diseño de la prueba: Etapa 3

Guaner Rojas-Rojas

El diseño de una prueba podría definirse dentro del marco de acción e interacción de los subprocesos: insumos, modelos, espacio de resultados. El proceso de diseño de la prueba inicia con la delimitación completamente documentada y detallada de los insumos y culmina con los resultados esperados a partir de la aplicación de la prueba, como se representa en la Figura 1.

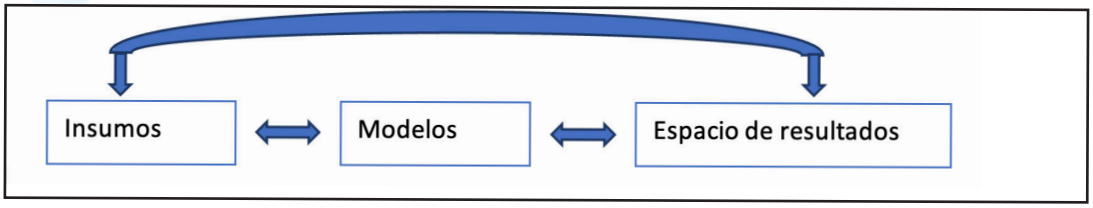


Figura 1. Marco de interacción del diseño de la prueba.

El reconocimiento del punto de partida y el de finalización del diseño de la prueba permitirá a los equipos de investigadores y evaluadores la identificación de los posibles aciertos o desaciertos del diseño. Para esto se seguirán dos tareas esenciales para mantener la calidad requerida en el diseño: verificación y aprobación. La verificación se realiza al final de uno o varios subprocesos del diseño y consiste en una serie de revisiones e inspecciones para evaluar los requerimientos y especificaciones. La aprobación se realizará una vez terminada la verificación y consiste en la revisión de la presencia de desviaciones de los resultados esperados. Si hay un resultado inesperado o desviado, la aprobación permitirá revisar nuevamente los requerimientos y especificaciones y evidentemente confirmar si los resultados vuelven a lo esperado.

El primer subproceso para el diseño de una prueba se enfoca en la especificación del constructo, el propósito de la prueba y el formato y tipo de ítems, como se ha mencionado en los apartados anteriores; mientras que el segundo corresponde al establecimiento del modelo de medición. Una vez definidos estos subprocesos, se verifica la correspondencia entre el constructo y el propósito de la prueba con el modelo de medición y el formato y tipo de ítems. La delimitación del constructo condicionará la escogencia del modelo y este condicionará el formato y tipo de ítems, como se explica en los siguientes apartados. Si se aprueba la correspondencia entre estos tres elementos, el siguiente subproceso consistirá en el establecimiento de las características técnicas de la dificultad promedio y la confiabilidad, restricción del tiempo de administración y accesibilidad a la prueba. Finalmente, se seleccionará el modelo de interpretación de las puntuaciones derivadas de la aplicación de la prueba.

Las siguientes recomendaciones que podrían guiar el diseño de la prueba se fundamentan en el aporte de Downing & Haladyna (2006), Rojas et al. (2020) y Wilson (2005).

1. Circunscribir el propósito de la prueba entre los más comunes tales como: nivel de instrucción, orientación o instrucción diagnóstica, selección, asignación, clasificación, certificación.

2. Adecuar los ítems a algunos de los formatos más comunes, por ejemplo, selección única, selección múltiple, sí o no, verdadero o falso, escalas tipo Likert, completar o desarrollar.
3. Seleccionar un modelo entre los marcos más reconocidos: teoría clásica de los test, teoría de respuesta al ítem, modelos psicométricos de clasificación diagnóstica, teoría de la generalizabilidad, modelos de Rasch o modelos logísticos de rasgos latentes.
4. Establecer el tiempo de administración, lo que permitirá evaluar costos y medios de aplicación en función de los recursos humanos y la capacidad de acción.
5. Determinar la longitud del instrumento en función del número de ítems necesarios para obtener puntuaciones objetivas y adecuadas al propósito de la prueba.
6. Implementar las condiciones de accesibilidad para todas las personas examinadas en términos de las adecuaciones sin afectar la medición del constructo, por ejemplo, tiempo adicional, formularios o cuadernillos diferentes, interpretación de los ítems en otros idiomas, tamaño de letra, aplicaciones individuales, cuadernillos en braille, entre otros.
7. Seleccionar el modelo de interpretación de las puntuaciones, sea referidos a normas o referidos a criterios, para decidir si a cada persona examinadas se le dará una o varias puntuaciones.
8. Construir y actualizar la tabla de especificaciones para evidenciar la representatividad de los ítems en los distintos dominios medidos en la prueba.
9. Generar y mantener un banco de ítems de la prueba para garantizar la calidad psicométrica de los ítems y de la prueba en general.

Referencias

Downing, S., & Haladyna, T. (2006). *Handbook of test development*. Lawrence Erlbaum Associates

Rojas, G., Jiménez, K., Calvo, K., Martínez, D., Ordóñez, K., Pérez, N., Rojas, L., Solano, S., Cerdas, D., Solano, L., Arias, M., Montero, E., & Valverde, M. (2020). *Manual de procesos de una prueba estandarizada: la PAA de la UCR*. Programa Permanente de la Prueba de Aptitud Académica de la Universidad de Costa Rica.

Wilson, M. (2005). *Constructing Measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Sección 2.4

La generación de la tabla de especificaciones: Etapa 4

Tania Elena Moreira-Mora

A partir de la definición del constructo, los propósitos y usos de la prueba, se puede iniciar con su diseño mediante la generación de la tabla de especificaciones. Para Tristán y Molgado (2007) es una matriz que combina un rasgo de dimensión cognitiva (o también afectiva, psicomotriz, etc.) en términos de una variable definida por medio de contenidos, objetivos o componentes de competencia (filas) con los niveles de complejidad o niveles taxonómicos (en las columnas). En cada celda se detalla la proporción en la cual interviene un contenido de un nivel de complejidad dado dentro de la prueba. Según estos autores, la tabla de especificaciones es similar a una tabla de validez de contenido, pero incluye información de los objetivos educativos o componentes de competencia (conocimientos, habilidades, destrezas motrices) asociados con la prueba y puede incluir una descripción detallada de los ítems que integran la prueba. Con esta herramienta se proporciona las primeras evidencias asociadas al contenido de la prueba, puesto que la medición del constructo se materializa mediante la construcción de ítems.

La validez de contenido se basa en la medición de determinados objetivos y dominios o áreas de contenido, con el propósito de determinar en qué medida los ítems de la prueba constituyen una muestra representativa del dominio cognitivo que se pretende medir. Según American Education Research Association (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCME) (2014), estas evidencias son claves para la alineación entre los estándares de aprendizaje del estudiante y los contenidos de la prueba, por tal razón, las especificaciones deben ser descritas cuidadosamente, como lo apuntan en estos dos estándares:

- Estándar 4.1: Las especificaciones de una prueba deberían describir el objetivo de la prueba, la definición del constructo o dominio medido, la población meta y la interpretación de los resultados.
- Estándar 4.2: Las especificaciones deberían definir los contenidos, el tamaño de la prueba, el formato de los ítems, las propiedades psicométricas de los ítems y de la prueba, el tiempo de aplicación, instrucciones para las personas examinadas, administración de la prueba, el procedimiento de calificación y los sistemas tecnológicos de apoyo.

En el ámbito educativo, Hurtado (2009) recomienda los siguientes pasos para el diseño y construcción de una prueba.

1. Determinar las áreas, sub-áreas y contenidos temáticos que serán medidos en la prueba.
2. La tabla de especificaciones debe diseñarse en forma de matriz con los contenidos en las filas y los niveles de complejidad en las columnas, como se muestra en la Tabla 2
3. Asignar la proporción (%) de preguntas que deberá tener cada contenido: Algunos criterios empleados para la asignación de proporciones son:
   - Pertinencia del contenido dentro del programa del área y del perfil del alumno de la institución educativa.
   - Tiempo empleado en su enseñanza y en el aprendizaje.
   - Importancia, aplicación o utilidad dentro de la misma área u otras áreas.

- Predictividad con respecto al criterio evaluado.

4. La muestra de preguntas en la prueba debe estar equilibrada con los contenidos y niveles asignados, de modo que estas representen el 100% a evaluar.
5. Se sugiere que todas las preguntas tengan el mismo peso. Los contenidos con mayor peso deberán contener más preguntas, en lugar de una sola pregunta por tema.
6. Una vez elaborada la tabla de especificaciones se procede con la construcción de la prueba, cuyo número de preguntas dependerá del tiempo disponible.

Tabla 2

Modelo de una tabla de especificaciones de una prueba de matemáticas.

		Niveles de complejidad			
	Contenido	Básico	Intermedio	Avanzado	Total
Álgebra	Ecuaciones de primer grado		4	2	6
	Ecuaciones de segundo grado	2	3		5
	Ecuaciones exponenciales	1	1		2
	Ecuaciones logarítmicas	1	2	2	5
Geometría	Área de regiones triangulares	1	2	3	6
	Área de regiones poligonales	1	2		3
	Área de regiones circulares	1	1	1	3
Trigonometría	RT de ángulos notables 1 2 3	1	2		3
	RT de ángulos agudos 3 2 5		3	2	5
	RT ángulos de cualquier magnitud	1	1		2
	Totales:	9	21	10	40

Fuente: Hurtado (2009).

Para iniciar la construcción de los ítems, destacan De los Santos-Roig y Pérez-Meléndez (2013), es necesario la especificación de sus características para que el equipo constructor los elabore de manera adecuada, entre las que mencionan: clarificar qué objetivo tiene el instrumento, qué tipo de ítems se formularán, qué formato tendrán, qué forma de puntuación, qué tipo de instrucciones podrían desarrollarse, como se muestra en el siguiente ejemplo de Tabla 3, que guio el proceso de elaboración de los ítems de una escala cognitiva.

Tabla 3

Modelo de una tabla de especificaciones de los ítems de una escala cognitiva.

Constructo a evaluar: Representación cognitiva de la enfermedad (R.C.E)

Instrumento: Escala sobre la representación cognitiva de la enfermedad, ERCE

Objetivo general de la medición: Evaluar las creencias sobre la enfermedad y sus cinco componentes: identidad, causas, control/cura, consecuencias y curso.

Aplicación: Investigación.

Áreas de contenido a evaluar

Componente 1: Identidad

Objetivo: Evaluar los síntomas experimentados por el paciente que este relaciona/ identifica como parte de la enfermedad.

Tareas:

- El evaluado deberá indicar los síntomas que padece o que están presentes en su vida cotidiana.
- Deberá distinguir cuáles de ellos son representativos o forman parte de la propia identidad de la enfermedad, indicando su relación con esta.

Atributos del ítem

Formato: Listado y pregunta directa

Contenido: Síntomas y pregunta en cada síntoma sobre su relación con la enfermedad

---

Atributos de las respuestas:

Formato de la escala de respuesta: Dos escalas dicotómicas; una para marcar una casilla en caso de "presencia" y otra con formato Sí/No (Marcar con una cruz la presencia del síntoma y en caso de presencia, señalar si se enlaza o no con la enfermedad).

Proporción de ítems dentro de la escala: 50% (síntomas) 50% (preguntas sobre la relación del síntoma con la enfermedad).

Ejemplo de instrucción:

"A continuación aparece una lista con una serie de síntomas. Queremos que nos indiques cuáles de ellos padeces ACTUALMENTE, EN TU DÍA A DÍA. Para ello, marca con una cruz la casilla que aparece junto a cada uno. Posteriormente, sobre cada síntoma marcado, queremos que nos indiques si crees que está relacionado o no con tu enfermedad. Si tienes alguna duda, levanta la mano y te atenderé de inmediato."

Ejemplo de ítem:

1. Náuseas ¿Crees que este síntoma tiene algo que ver con tu enfermedad? Sí __ No __

Fuente: De los Santos-Roig y Pérez-Meléndez (2013).

Para Sireci y Faulkner-Bond (2014) el método más común para la obtención de evidencias de validez de contenido de la prueba se basa en el criterio de expertos, quienes se pueden enfocar en (a) emparejar los ítems con su respectivo objetivo, (b) evaluar el grado en que los ítems representan el contenido y especificación cognitiva, o (c) evaluar el grado en que los ítems son relevantes para el dominio de los contenidos. Para estos estudios usualmente se usa la técnica "matching task" con una escala tipo Likert para medir la congruencia entre cada ítem y cada uno de los aspectos evaluados, como se puede ver en la Tabla 4.

Tabla 4

Ejemplo para el emparejamiento entre ítems y una hipotética prueba de conocimientos en Matemática

	Cognitive level								
Content area	Knowledge, Comprehension			Application Analysis, Synthesis			Evaluation		
	NR	PFA	GM	NR	PFA	GM	NR	PFA	GM
Nº Item									
1									
2									
3									
4									
5									
...									

Directions: Please review each item and indicate (a) the content area, and (b) Cognitive level you think the item is measuring. Please be sure to make two selections for each item—one for content area, and one for cognitive level.

Content area: Number relations (NR), Patterns, functions &Algebra (PFA) y Geometry & measurement (GM) & measurement (GM)

Fuente: Adaptado de Sirecci y Faulkner-Bond (2014)

En la tabla 5 se muestra otro ejemplo para determinar el grado en que cada ítem representa un dominio. Según Sirecci y Faulkner-Bond (2014), la representación se refiere al grado en que una prueba representa y mide adecuadamente el dominio definido en las especificaciones de la prueba.

Tabla 5
Ejemplo para evaluar la congruencia entre objetivos e ítems.

		How well does the item measure its objective? (circle one)					
Item	Objective	1	2	3	4	5	6
		Not at all				Very well	
1	Convert units of measure in the same systems						
2	Read values on a bar, line, or circle graph						
3	Find the average (mean) and range for a data set						
4	Find the perimeter of rectangles						
5	Infer meaning from gaps, clusters and comparisons of data.						
6	Directly measure and compare the radius, diameter, and circumference of a circle.						
8	Read and understand positive and negative numbers as showing direction and change.						
...							
100	Use a number line to represent the counting numbers						

Directions: Please read each item and its associated benchmark. Rate how well the item measures its objective using the rating scale provided.
Be sure to circle one rating for each item.
Fuente: Sirecci y Faulkner-Bond (2014)

En el dominio cognoscitivo se ha utilizado la taxonomía de Benjamín Bloom, propuesta desde 1956, para categorizar los distintos procesos del conocimiento en: conocimiento, comprensión, aplicación, análisis, síntesis y evaluación. Posteriormente fue revisada en el 2001 por David Krathwohl y Lorin Anderson y se conoce como Taxonomías de Anderson y Krathwohl 2001, como se muestra en la Tabla 6. Como destaca Owen (2016), las principales diferencias radican en los diferentes tipos y niveles de conocimiento del hecho, conceptual, procedimental y metacognitivo, específicamente:

- Del hecho: Los estudiantes deben conocer los elementos básicos para tener conocimiento de una disciplina o resolver problemas.
- Conceptual: Las interrelaciones entre los elementos básicos dentro de una estructura más grande que les permiten funcionar juntos.
- Procedimental: Cómo llevar a cabo el procedimiento y los criterios de ejecución de las habilidades, algoritmos, técnicas y métodos.
- Metacognitivo: El conocimiento de la cognición en general, así como la sensibilización y el conocimiento de la propia cognición de uno.

Tabla 6

Comparación de las taxonomías de los procesos cognitivos

Taxonomías de Bloom 1956	Taxonomías de Anderson y Krathwohl (2001)
1. Conocimiento: Al recordar o recuperar material previamente aprendido. Los ejemplos de verbos que se relacionan con esta función son: conocer, identificar, relacionar, listar -definir, recordar, memorizar, repetir -registrar, nombrar, reconocer y adquirir.	1. Recordar: reconocer o recordar conocimientos de la memoria. Recordar es cuando la memoria se utiliza para producir definiciones, hechos, o listas, o recitar o recuperar material.
2. Comprensión: La capacidad de captar o construir significado a partir de material. Los ejemplos de verbos que se relacionan con esta función son: -replantear, localizar, informar, reconocer, explicar, expresar -identificar, analizar, describir, analizar, revisar, inferir	2. Descripción: Construir significado a partir de diferentes tipos de funciones, escritos o gráficos de actividades como interpretar los mensajes, proponiendo clasificaciones, resúmenes, inferir, comparar, y explicar.

-ilustrar, interpretar, dibujar, representar, diferenciar y concluir.	
3. Aplicación: La capacidad de utilizar el material aprendido, o para aplicar el material en situaciones nuevas y concretas. Los ejemplos de verbos que se relacionan con esta función son: -aplicar, relacionar, desarrollar, traducir, usar, operar -organizar, emplear, reestructurar, interpretar, mostrar, ilustrar -practicar, calcular, demostrar, exponer y dramatizar.	3. Aplicación: Llevar a cabo o mediante el procedimiento de ejecución, o de poner en práctica. Aplicar se refiere a las situaciones en que se utilicen materiales adquiridos a través de productos como modelos, presentaciones, entrevistas o simulaciones.
4. Análisis: La capacidad de dividir o distinguir las partes de material en sus componentes de forma que su estructura organizativa pueda entenderse mejor. Ejemplos de verbos que se relacionan con esta función son: -analizar, comparar, sondar, investigar, examinar, contrastar, categorizar -diferenciar, contrastar, investigar, detectar, encuestar, clasificar, deducir -experimentar, escudriñar, descubrir, inspeccionar, analizar, discriminar y separar.	4. Análisis: descomponer el material o conceptos en partes, la determinación de cómo las partes se relacionan o se interrelacionan entre sí o a la estructura general o propósito. Acciones mentales incluidas en esta función se diferencian, organizan y se atribuyen, así como ser capaz de distinguir entre los componentes o piezas. Cuando uno está analizando él/ella puede ilustrar esta función mental para crear hojas de cálculo, encuestas, diagramas o esquemas, o representaciones gráficas.

5. Síntesis: La capacidad de poner las piezas juntas para formar un nuevo todo coherente o único. Los ejemplos de verbos que se relacionan con esta función son: -componer, producir, diseñar, montar, crear, preparar, predecir, modificar, decir -planear, inventar, formular, recopilar, crear, generalizar, documentar, combinar, relacionar -proponer, desarrollar, organizar, construir, organizar, originar, derivar, escribir y proponer.	5. Evaluación: hacer juicios en función de criterios y normas de control y crítica. Las críticas, recomendaciones, y los informes son algunos de los productos que se pueden crear para demostrar los procesos de evaluación. En la nueva taxonomía está la evaluación antes de crear ya que es a menudo una parte necesaria de la conducta previa antes de crear algo.
6. Evaluación: Es la capacidad de juicio, revisión y crítica del valor del material dado. Ejemplos de verbos relacionados con esta función son: -juzgar, comparar, evaluar, concluir, medir, deducir -argumentar, decidir, seleccionar, estimar, calificar -validar, considerar, apreciar, valorar, criticar e inferir.	6. Crear: reunir elementos para formar un todo coherente y funcional, reorganizar elementos en un nuevo modelo o estructura a través de la generación, planificación o producido. Crear requiere que los usuarios pongan las piezas juntas de un modo nuevo o sintetizar las piezas en algo nuevo y diferente con un nuevo formulario o producto. Este proceso es el más difícil como función mental en la nueva taxonomía.

Fuente: Adaptado de Owen (2016).

En suma, la tabla de especificaciones es una herramienta muy útil para la construcción de una prueba de alto impacto en cualquier contexto académico y profesional y para la obtención de evidencias asociadas a la validez de contenido, las cuales no son suficientes para justificar los usos de los resultados de una prueba. No obstante, como lo propone Messick (1991), en contextos en que la evidencia sobre la validez de constructo es menos específica, estos tipos de validez pueden respaldar (o no) las implicaciones que el significado atribuido a los puntajes tiene para la acción, lo que en última instancia justifica un determinado uso de las pruebas.

Referencias

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

De los Santos-Roig, M. y Pérez-Meléndez, C. (2013) *De la definición semántica a los ítems a través de las tablas de especificaciones: elaboración de la Escala sobre Representación de la Enfermedad, ERCE*. Anales de Psicología, 29 (2), 345-359. http://www.redalyc.org/pdf/167/16726244004.pdf

Hurtado, L. (2009). *La tabla de especificaciones de una prueba*. http://miprofesordematematicas.blogspot.com/2009/07/tabla-de-especificaciones-de-una-prueba.html

Messick, S. (1991). *Validez de la interpretación de una prueba y su uso*. Programa de Promoción de la Reforma Educativa en América Latina y el Caribe. https://www.thedialogue.org/wp-content/uploads/2016/07/2000-Validez-de-la-interpretacion-de-una-prueba-y-su-uso-Messick.pdf

Sireci, S. &Faulkner-Bond, M. (2014). *Validity evidence based on test content*. Psicothema, 26, (1), 100-107. doi: 10.7334/psicothema2013.256

Tristán, A. y Molgado, D. (2007). *Tablas de validez de contenido*. https://www.itsescarcega.edu.mx/documentos/desacad/Curso%20Taller/TABLAS~1.PDF

Owen, L. (2016). Anderson and Krathwohl Bloom's Taxonomy Revised. *Understanding the New Version of Bloom's Taxonomy*. https://quincycollege.edu/wp-content/uploads/Anderson-and-Krathwohl_Revised-Blooms-Taxonomy.pdf

Sección 2.5

Construcción y juzgamiento de ítems para pruebas estandarizadas: Etapa 5

Armel Brizuela- Rodríguez

La construcción de pruebas de alto impacto se sustenta en las definiciones del constructo o los contenidos que se pretenden medir, en sus propósitos y usos, como se ha explicado con sumo detalle en los apartados anteriores. Este marco también será la base para las etapas subsiguientes de la construcción, juzgamiento y aplicación de la prueba. Al respecto, es pertinente retomar las palabras de Messick (1994, p. 16) sobre las implicaciones de construir pruebas estandarizadas:

> "A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of constructed-based scoring criteria and rubrics."

Cualquier instrumento de medición está compuesto por un conjunto de ítems, entre los más frecuentes, de selección única y de respuesta construida. La estructura usual de los ítems de selección única es de tres partes: (1) uno o varios textos en los que se describe, explica, narra o argumenta una situación o tema; (2) una pregunta relacionada con el texto; y (3) dos o más opciones de respuesta. En el caso de las pruebas diseñadas para la medición de capacidades (destrezas, competencias, aptitudes, habilidades, etc.) y conocimientos, una respuesta se considera como correcta, mientras que el resto son incorrectas. Para las pruebas cuyo propósito es la evaluación de actitudes, rasgos de personalidad, opiniones y estados emocionales, ninguna de las opciones de respuesta se considera como correcta.

Por otra parte, los ítems de respuesta construida no presentan un único formato usual o típico, ya que este depende de la naturaleza específica de lo que se pretenda medir. Generalmente, las pruebas construidas con este tipo de ítems deben proporcionar una serie de instrucciones, tareas o preguntas para que las personas examinadas desarrollen una respuesta (p.ej., redacción de un ensayo argumentativo o resolución de un problema matemático) o demuestren su desempeño en una tarea motora (p.ej., una prueba práctica para obtener la licencia de conducir o una audición de teatro). Posteriormente, los productos derivados de las actividades desarrolladas por las personas examinadas son calificadas por uno o varios expertos con base en un conjunto de criterios establecidos a priori en una "rúbrica de calificación".

Tanto los ítems de selección única como los de respuesta construida deben ser construidos tomando en consideración los mismos criterios de validez, confiabilidad y equidad (Messick, 1994). En este sentido, Messick (1994, p. 15) plantea que "el contraste entre ítems de selección múltiple y las tareas de desempeño de respuesta abierta no es una dicotomía sino un continuum que representa diferentes grados de estructuración de la respuesta", por lo cual no es apropiado plantear criterios de calidad psicométrica diferentes para ambos tipos de ítems.

Por lo tanto, las recomendaciones para la construcción y juzgamiento de ítems deben sustentarse en los criterios actuales sobre validez, confiabilidad y equidad en la medición psicológica y educativa (AERA, APA y NCME, 2014) para ambos tipos de formato.

Las siguientes recomendaciones tienen el propósito de eliminar las posibles fuentes de variancia irrelevante y la sub-representación en el constructo medido por los ítems de la prueba. Estas recomendaciones se fundamentan en el aporte de diversos autores (Darling-Hammond, 2014; Haladyna y Rodriguez, 2013; Johnson, Penny y Gordon, 2009; Lane y Iwatani, 2016; Rodriguez, 2016) y se brindan para ítems de selección única, de respuesta construida y, en caso de que no especificarse, debe asumirse que será para ambos tipos.

1. La resolución de cada ítem solo debe basarse en un contenido y en una estrategia cuando se desee construir una escala unidimensional, de modo que siempre quede claro qué se pretende medir con cada uno de los ítems.
2. En la redacción de los ítems se debe utilizar información desconocida por las personas examinadas cuando se desea promover el uso de estrategias de pensamiento más complejas. De esta manera, se reducen las posibilidades de que las personas evaluadas acudan a conocimientos memorizados previamente para contestar los ítems.
3. La respuesta de un ítem no debe depender de las respuestas a otro ítem, sino únicamente del nivel en el constructo de la persona examinada. Se debe recordar que, una vez aplicada la prueba, será necesario analizar las respuestas de acuerdo con algún modelo de medición apropiado a la naturaleza de estos; generalmente, dichos modelos parten del supuesto de que las respuestas a los ítems son independientes entre sí.
4. El grado de especificidad del contenido del ítem debe ser adecuado al marco de referencia, la tabla de especificaciones y el propósito de la prueba, por lo que debe evitarse la inclusión de contenidos demasiado generales o específicos.
5. No es recomendable incluir opiniones y juicios de valor en los ítems, a menos que estos estén debidamente cualificados.
6. Los ítems no deben estar redactados de manera truculenta o engañosa, ya que, la dificultad de un ítem no debe verse disminuida ni aumentada por aspectos de redacción que no se relacionan con el constructo que se pretende medir.
7. En el caso de los ítems de selección única, las opciones de respuesta deben aparecer ordenadas verticalmente de acuerdo con algún criterio (longitud, alfabético, magnitud, etc.).
8. El número y la complejidad de palabras, frases y oraciones deben reducirse todo lo posible para evitar que problemas de comprensión lectora interfieran en la medición del constructo de interés. Por ejemplo, es importante evitar el uso de negaciones y de dobles negaciones.
9. La idea o mensaje central de un ítem de selección única debe desarrollarse en el texto principal, no en las opciones.

10. Para cada ítem de selección única deben redactarse opciones plausibles; generalmente, tres o cuatro alternativas son suficientes (Rodriguez, 2005). En otras palabras, no se deben incluir opciones de respuesta incorrectas absurdas, humorísticas u obviamente falsas.
11. Ningún ítem de selección única debe aparecer con más de una opción correcta, excepto que la prueba haya sido desarrollada con opciones parcialmente correctas. De ser así, será necesario construir los ítems con base en un modelo de medición apropiado en el que las respuestas se califiquen no en términos de correctas o incorrectas, sino de acuerdo con el grado de corrección de cada respuesta (Nering y Ostini, 2010).
12. Las opciones de respuesta de los ítems de selección única deben ser mutuamente excluyentes, de manera que la escogencia de una alternativa no descarte o implique automáticamente otra. Las opciones incorrectas deben ser diferentes entre sí, de modo que al rechazar una de ellas no se descarten automáticamente las que sean equivalentes.
13. En los ítems de selección única se debe evitar el uso de opciones como "Ninguna de las anteriores", "Todas las anteriores", "No sé" y similares.
14. La extensión de todas las opciones con ítems de selección única debe ser similar para evitar que las personas examinadas utilicen la longitud de estas como un criterio para descartarlas o seleccionarlas.
15. Las opciones de respuesta de los ítems de selección única no deben incluir términos absolutos como "siempre", "nunca", "completamente", "absolutamente", entre otros; ya que estos suelen ser descartados fácilmente por ser poco plausibles.
16. La respuesta correcta no debe incluir palabras del texto principal de un ítem de selección única, con el objetivo de evitar que quienes posean un bajo nivel en el constructo acierten el ítem basándose meramente en esta repetición de palabras.
17. La estructura gramatical y el estilo de las opciones en un ítem de selección única debe ser homogénea y congruente con respecto a la instrucción.
18. Cuando se elaboren ítems de respuesta construida, la calificación es de manera independiente y, al menos, con dos personas calificadoras, quienes son entrenados para esta labor por el equipo evaluador. El equipo calificador se conforma con expertos en el constructo y debe entrenarse específicamente en el uso de la rúbrica hasta que muestren un nivel aceptable de concordancia entre sí.
19. Las rúbricas y otros materiales empleados por el equipo calificador deben medir una muestra de tareas representativas del universo de capacidades y conocimientos que interesa evaluar. Lo anterior implica que no se deben incluir rubros irrelevantes al constructo, ni se deben excluir componentes o facetas importantes de este
20. El equipo constructor de los ítems, conformado por expertos en el constructo de interés, deben basarse en la tabla de especificaciones en la que se indiquen las capacidades y conocimientos que se desee evaluar.
21. Una vez construidos los ítems, estos deben ser evaluados o juzgados por expertos que no hayan participado en la construcción. .

También es conveniente realizar un proceso de juzgamiento por parte del mismo equipo constructor para garantizar que los jueces externos reciban los ítems sin errores fácilmente detectables.

22. Los aspectos más importantes por evaluar en el proceso de juzgamiento son aquellos establecidos en la tabla de especificaciones y en los relacionados con: plausibilidad de las opciones de respuesta en los ítems de selección única y normas de redacción (ortografía, gramática, coherencia, entre otros). También el análisis de sensibilidad de los contenidos para evitar la inclusión de temas que favorezcan a una determinada población, o que sean controversiales, según características de género, étnicas, culturales o religiosas.

23. Es recomendable confeccionar una guía de juzgamiento para orientar apropiadamente la labor del equipo de jueces, como los ejemplos de las Tablas 4 y 5. También es necesario que reciban un entrenamiento adecuado en el uso de la tabla de juzgamiento.

Referencias

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

Darling-Hammond, L. (Ed.) (2014). *Next Generation Assessment: Moving Beyond the Bubble Test to Support 21st Century Learning*. Jossey-Bass-Wiley.

Haladyna, T. & Rodriguez, M. (2013). *Developing and Validating Test Items*. Routledge.

Johnson, R., Penny, J. & Gordon, B. (2009). Assessing Performance: Designing, Scoring, and Validating Performance Tasks. The Guilford Press.

Lane, S. & Iwatani, E. (2016). *Design of Perfomance Assessments in Education*. En S. Lane, M. Raymond & T. Haladyna (Eds.), Handbook of Test Development (pp. 274 - 292). Routledge.

Messick, S. (1994). *The Interplay of Evidence and Consequences in the Validation of Performance Assessments*. Educational Researcher, 23(2), 13-23. DOI: 10.2307/1176219

Nering, M. & Ostini, R. (2010). *Handbook of Polytomous Item Response Theory Models*. Routledge.

Rodriguez, M. (2005). *Three Options Are Optimal for Multiple-Choice Items: A Meta Analysis of 80 Years of Research*. Educational Measurement: Issues and Practice, 24(2), 3-13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rodriguez, M. (2016). *Selected-response Item Development*. En S. Lane, M. Raymond & T. Haladyna (Eds.), Handbook of Test Development (pp. 259 - 273). Routledge.

Sección 2.6

Aplicación piloto y análisis psicométrico: Etapa 6

Eiliana Montero-Rojas

Una fase esencial del proceso de validación de las pruebas estandarizadas, a diferencia de las pruebas de aula, es la aplicación piloto a una muestra representativa de la población de interés y la generación de diversas medidas empíricas de la calidad del instrumento, considerando modelos de medición proveídos por la psicometría, como disciplina que brinda el marco de referencia teórico y metodológico para generar indicadores objetivos de la validez y confiabilidad de las mediciones.

Especialmente si la prueba va a utilizarse con propósitos de toma de decisiones sobre las personas examinadas, es fundamental que antes de su aplicación operacional sea sometida a una aplicación piloto y análisis de calidad técnica bajo modelos psicométricos. Estas técnicas y métodos se basan en enfoques cuantitativos y utilizan conceptos, procedimientos y medidas de la estadística y la matemática (Nunnally & Bernstein, 1995; Bond & Fox, 2001; Martínez, Hernández & Hernández, 2006).

Por supuesto que a partir de estas notas no será posible para el lector poco familiarizado, obtener el suficiente conocimiento en el tema del análisis psicométrico, por lo que se recomienda contar siempre en el equipo evaluador con una persona especialista en psicometría, medición educativa o métodos cuantitativos y conocedora de modelos de medición para la construcción de instrumentos.

Si bien, tal como se dijo anteriormente, se deben recolectar evidencias empíricas de los atributos de los ítems por medio de la consulta a jueces expertos, usualmente esta evidencia no es suficiente para garantizar de manera adecuada la calidad del instrumento y los reactivos que lo componen, según el propósito y la población a la que va dirigido, de ahí que la aplicación piloto sea un requisito fundamental como parte del proceso de validación.

Se debe recordar que la aplicación piloto tiene como principal finalidad generar medidas empíricas de la calidad técnica de los ítems, y, en algunos casos, brindar información sobre otras condiciones del proceso de administración de las pruebas, tales como su duración y comprensión de las instrucciones, entre otras.

Como regla general, la muestra piloto debería ser representativa de la población a la que va dirigida en la prueba, por tanto muestras de conveniencia no son recomendadas.
El tamaño de muestra debe ser tal que permita una correcta estimación de los parámetros, según los modelos de medición que se van a utilizar para el análisis. Por ejemplo, en el caso de la Teoría Clásica de los Tests (TCT) la muestra debería ser al menos de 75 sujetos, mientras que para el Modelo de Rasch se ocuparía un mínimo de 200.

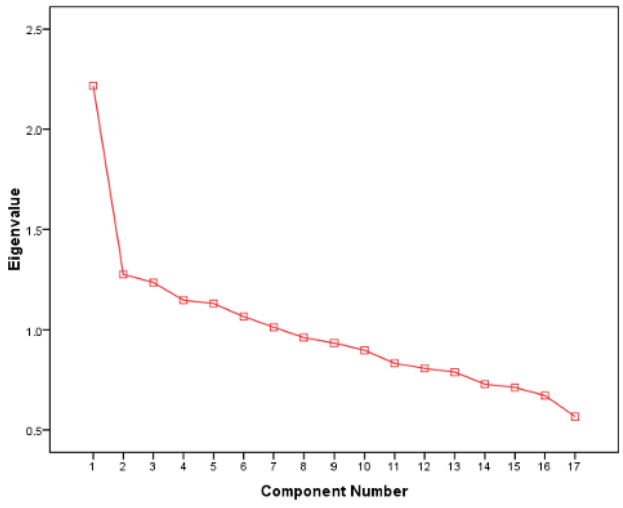
Como se puede apreciar de las cifras anteriores, este último modelo es más exigente que la TCT en cuanto a tamaño de muestra, sin embargo al ofrecer mejores herramientas técnicas para la generación de pruebas referidas a criterios, calibradas según sus propósitos, Rasch y sus diversas extensiones representan una opción más adecuada para la construcción y validación de pruebas de alto impacto en contextos de certificación o selección, especialmente si se desean generar descripciones sustantivas del desempeño de las personas examinadas, de acuerdo con la calificación o puntaje obtenido en la prueba.

Cuando la prueba incluye un conjunto de ítems o reactivos de selección única, el primer paso del análisis psicométrico implica la realización de un análisis de factores exploratorio (AFE) para determinar si el instrumento es unidimensional. Esta propiedad se refiere a que los ítems deben estar midiendo fundamentalmente midiendo un solo rasgo, conocimiento o habilidad (Nunnally & Bernstein, 1995; Martínez, Hernández & Hernández, 2006). El análisis de factores exploratorio es una técnica de la estadística multivariada cuyo precursor fue Charles Spearman, en el siglo XIX, a partir de sus estudios iniciales para aproximarse empíricamente a la medición de la inteligencia.

A continuación, se presentan dos gráficos de sedimentación, los cuales son generados como parte de los resultados del AFE. El primero ilustra la existencia de un solo factor rasgo o constructo fundamental, mientras que el segundo ilustra la existencia de tres. La explicación técnica sobre la que se basan estas interpretaciones escapa a la naturaleza de este documento y puede encontrarse en las referencias bibliográficas de este apartado.

En el caso de más de un factor, como en el segundo gráfico, se debe generar una solución factorial para el número sugerido de factores y verificar que la distribución de los ítems en estos tenga sentido teórico. En general cada agrupación de ítems de un mismo factor se debe analizar separadamente, ya que la mayoría de los diversos modelos psicométricos, asumen unidimensionalidad.

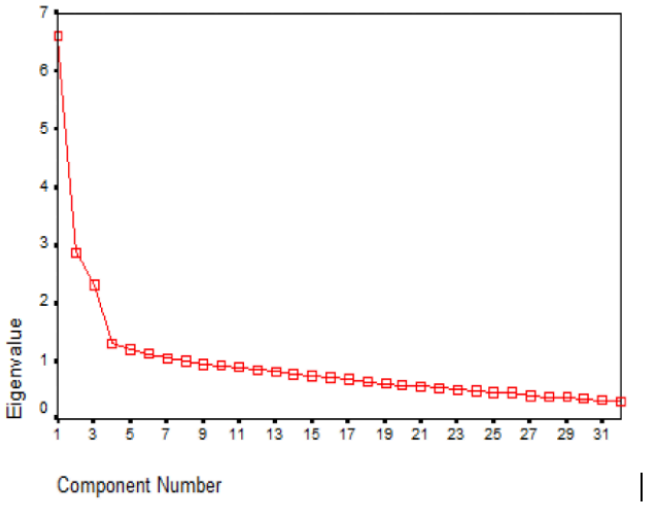
Figura 2
Gráfico de sedimentación representando un solo factor principal.



Fuente: Elaboración propia

Figura 3

Gráfico de sedimentación representando tres factores principales.



Fuente: Elaboración propia

Una vez que se ha logrado el requisito de unidimensional se debe aplicar algún modelo de medición psicométrico para aproximarse empíricamente a los índices de calidad técnica, tanto de los ítems como del instrumento como un todo.

En cuanto a la TCT, quizás el modelo de medición más conocido, sus desarrollos iniciaron en el siglo XIX y dominaron el campo de la medición psicológica y educativa hasta la primera mitad del siglo XX. Este enfoque propone métodos empíricos para el análisis de calidad de una prueba y de los ítems que la componen (Muñiz, 2003). Uno de sus resultados más importantes es el coeficiente de confiabilidad de consistencia interna Alfa de Cronbach, que se define para la totalidad del instrumento constituido por un conjunto de ítems y cuya fórmula es la siguiente:

$$A = (k / k-1)\ (1 - Ss_i^2 / S_y^2)$$

donde

A: coeficiente Alfa de Cronbach
K: es el número de ítems
Ss_i^2: es la sumatoria de las variancias individuales de los ítems
S_y^2: es la variancia de la suma total de los puntajes

Además, la TCT provee indicadores de calidad para los ítems o reactivos individuales que conforman la prueba, dos de los más relevantes son los parámetros de dificultad y discriminación del ítem. La dificultad se define como la proporción de respuestas correctas en el caso de ítems de calificación dicotómica (0-1) o como el promedio de las respuestas de las personas examinadas si se trata de ítems de calificación graduada de la respuesta (por ejemplo, de 1 a 5 ó de 0 a 10).

Dependiendo del propósito de la prueba se deberán utilizar ítems más o menos difíciles, conociendo que entre mayor sea el promedio del ítem o su proporción de correctas más fácil es el ítem. Por ejemplo, si se trata de una prueba muy competitiva se deberán seleccionar ítems con niveles altos de dificultad, por el contrario, si se desea medir con precisión los examinados de menor rendimiento para ubicar en un programa remedial, por ejemplo, deberán seleccionarse ítems fáciles.

Por otro lado, se define el índice de discriminación en la TCT como la correlación (r de Pearson) entre la puntuación obtenida en un ítem particular, y el puntaje total en el instrumento, para la muestra de examinados. La discriminación es una medida de la capacidad que tiene el ítem para distinguir entre examinados con puntajes altos y con puntajes bajos, y, usualmente se considera óptima cuando su valor es igual o superior a 0,30.

Esta medida se puede interpretar como una aproximación al porcentaje de variabilidad verdadera que representan los puntajes observados en la prueba. Es decir, indica qué porcentaje de la variabilidad en los puntajes observados de la prueba representa la variancia verdadera de un universo hipotético de ítems diseñados para medir el constructo de interés.

A partir de la fórmula para calcular el Alfa de Cronbach se observa que está afectada positivamente por el número de ítems, es decir, en general para un mismo nivel de calidad técnica, al aumentar el número de ítems aumentará la confiabilidad. Sin embargo, no se recomienda sencillamente incrementar el número de ítems para lograr valores mayores de Alfa, dado que esa acción no sustituye una buena elaboración técnica de reactivos que sean representativos del constructo de interés. De igual forma, al aumentar la longitud de la prueba se presentan con mayor probabilidad, efectos de cansancio o fatiga que pueden invalidar la interpretación de los puntajes. Asimismo, el Alfa de Cronbach será mayor cuanto más alta sea la correlación bivariada entre los ítems, y más alta sea la variabilidad del puntaje total en la prueba. En general para propósitos de toma de decisiones se recomienda que el valor de Alfa no sea menor a 0,8.

A continuación, se presenta como ejemplo un extracto de resultados arrojados bajo la TCT. Se nota que el ítem con el nombre VAR24 debería ser eliminado del instrumento, puesto que su índice de discriminación es muy bajo y al eliminarlo el valor de Alfa de Cronbach sube a un valor mayor al que ostenta originalmente. Si se trata de una aplicación piloto la decisión debe ser no utilizar este ítem en la aplicación principal u operativa, pues no cumple con los requisitos de calidad técnica bajo la TCT.

Tabla 7
Extracto de un de análisis de ítems con la TCT

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected item Total Correlation	Alpha if Item Deleted
Var01	12,0231	22,4310	,3106	,8131
Var02	11,9793	22,3449	,3505	,8113
Var03	12,0489	22,0271	,3960	,8093
Var04	12,1041	22,1040	,3626	,8108
Var05	12,0591	21,9641	,4071	,8087
Var06	12,0796	21,6060	,4838	,8051
Var07	11,9418	22,5459	,3214	,8125
Var08	12,1494	21,7942	,4247	,8078
Var09	12,2139	21,2740	,5384	,8022
Var10	12,3097	21,8284	,4231	,8079
Var11	12,1202	21,8230	,4230	,8079
Var12	12,1745	21,4889	,4910	,8046
Var13	12,1243	21,7838	,4311	,8075
Var14	11,9686	22,3220	,3621	,8109
Var15	12,2032	21,8421	,4099	,8085
Var16	12,3097	22,4314	,2875	,8143
Var17	12,3036	22,5761	,2547	,8158
Var18	12,3440	22,1834	,3503	,8114
Var19	12,3968	22,6298	,2629	,8152
Var20	12,4934	22,8307	,2547	,8152
Var21	12,5547	22,9927	,2527	,8150
Var22	11,9766	22,6084	,2871	,8140
Var23	11,9783	22,0718	,4189	,8084
Var24	12,5170	23,8618	,-0029	,8245

Nota:
N of Cases = 210,0
N of Items = 24
Cronbach's Alpha = 0,8192
Fuente: Elaboración propia

En cuanto a los modelos de TRI (Teoría de Respuesta a los Ítems) los parámetros de los ítems y de las personas examinadas son constructos o variables latentes que deben estimarse a partir del ajuste estadístico de un modelo matemático a los datos empíricos generados por la prueba. Estos modelos inician su desarrollo en el siglo XX y a partir de la segunda mitad de este siglo comienzan a dominar el campo de la medición educativa.

En general, los modelos de TRI requieren de muestras más grandes que la TCT para la aplicación piloto (usualmente al menos 200 sujetos) y el conocimiento básico de estadística y matemática para su correcta interpretación es más sofisticado comparado con la TCT, pues se utilizan diversas herramientas de la estadística inferencial como modelos de probabilidad, estimación y prueba de hipótesis a nivel univariado y multivariado (Hambleton, Swaminathan & Rogers, 1991; Montero, 2001).

Entre los modelos de TRI más conocidos y utilizados se distinguen los modelos logísticos de uno, dos y tres parámetros. En particular el modelo logístico de un parámetro se denomina Modelo de Rasch, en honor al matemático danés Georg Rasch, quien en la década de los años 60 y 70 descubrió una serie de propiedades muy atractivas que poseía este modelo a nivel práctico, y que le conferían muchas ventajas para la interpretación sustantiva de los puntajes en educación y psicología (Bond & Fox, 2001; Wilson, 2005).

Estas propiedades resultan especialmente importantes cuando se desea inferir, a partir de los resultados de la prueba, qué puede o no realizar la persona examinada en términos del constructo de interés bajo medición. A este tipo de descripciones se les llama interpretaciones referidas a criterios. Suele ser el caso de las pruebas de selección de personal en el contexto laboral y de las pruebas de certificación en el ámbito educativo y laboral. En estas situaciones poco interesa el puntaje numérico obtenido, sino más bien lo que indica en torno a la competencia del individuo, según el constructo bajo evaluación.

El modelo de Rasch permite de manera clara y eficiente responder a esta pregunta, a diferencia de los otros modelos de TRI. Según esta autora esta es la principal razón para su actual popularidad. El modelo clásico de Rasch se presenta en dos expresiones que son equivalentes a nivel matemático:

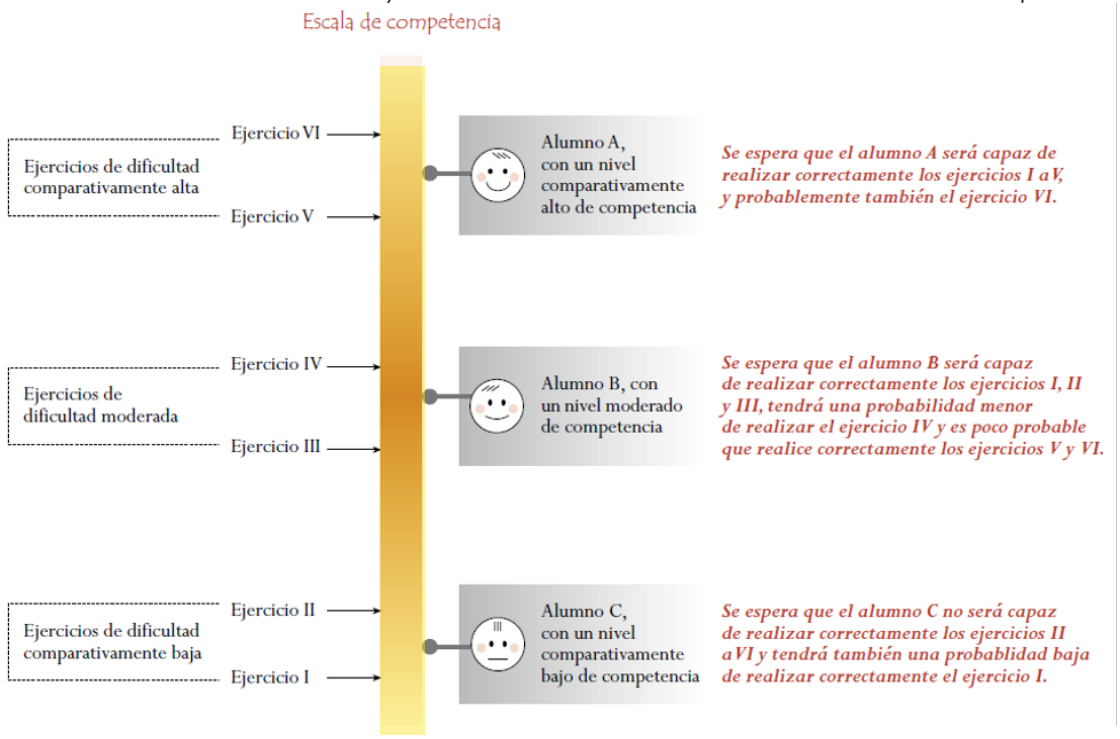
$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

$$\ln[P_i\,(\theta)/1 - P_i(\theta)] = \theta - bi$$

A partir de su definición matemática se puede observar que la probabilidad de acertar (o fallar el ítem) solo depende de la distancia entre la habilidad del sujeto y la dificultad del ítem. Además, la estimación de los puntajes del examinado y la dificultad de los ítems están en la misma unidad de medida, a esta propiedad se le denomina medición conjunta y permite ubicar en una misma escala de medición la estimación del nivel del constructo en el examinado y la dificultad de los ítems. Gracias a esta característica Rasch es especialmente útil para generar interpretaciones sustantivas de desempeño, tal como se ilustra seguidamente.

Figura 4:

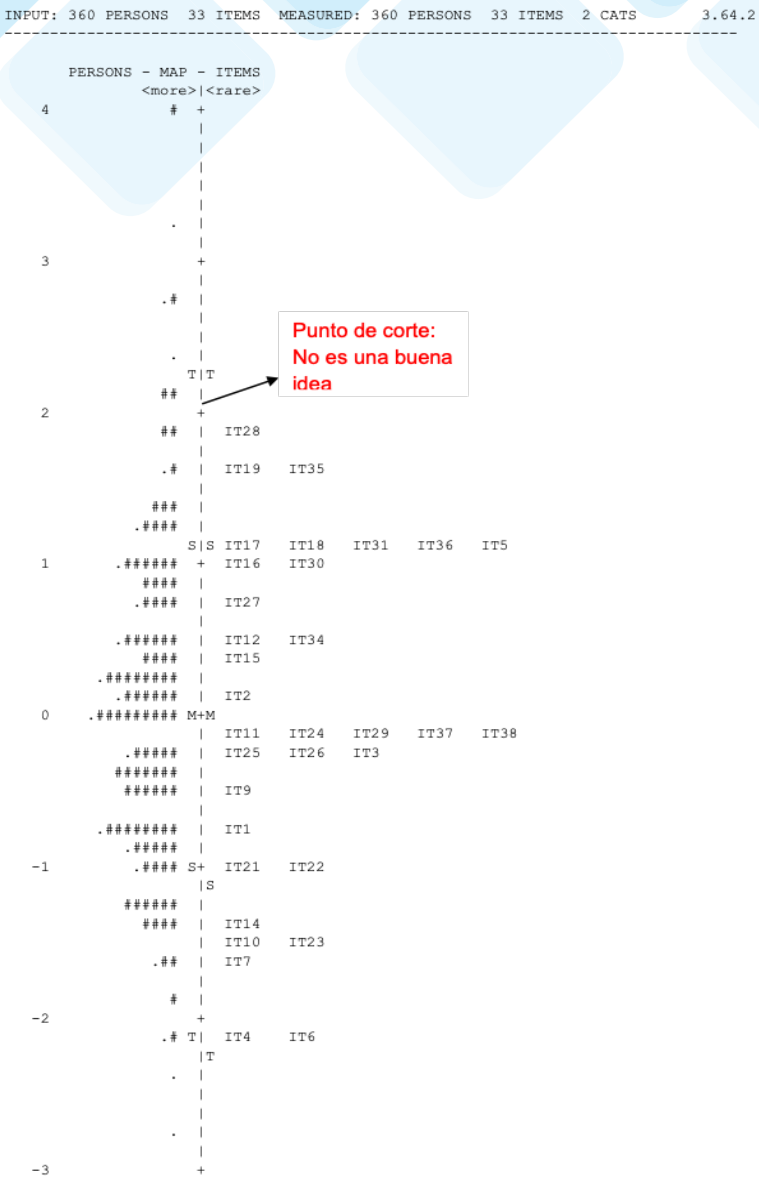
Relación entre los ítems del test y los niveles del estudiantado en una escala de competencia



Fuente: Organización para la Cooperación y el Desarrollo Económicos (2004).

A este tipo de interpretaciones se llega por medio del "mapa de personas versus ítems", un resultado de gran utilidad que es posible generar con el modelo de Rasch, a partir de la propiedad de medición conjunta, como se muestra en el siguiente ejemplo.

Figura 5
Mapa de personas versus ítems en el modelo de Rasch

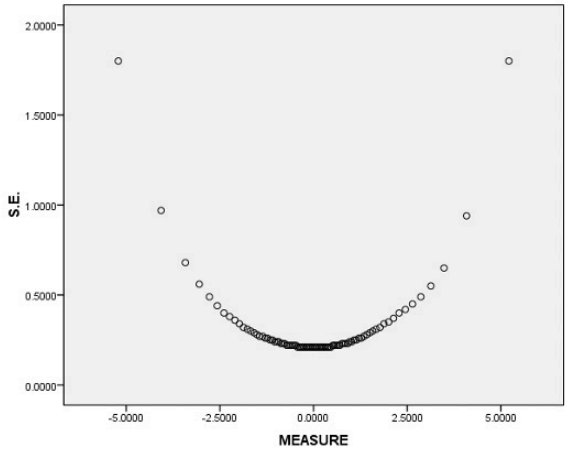


Fuente: Elaboración propia

Igualmente es necesario analizar la precisión con que se está midiendo la estimación del constructo en cada examinado, para lo cual se puede utilizar el diagrama de dispersión que relaciona cada estimación del constructo con su respectivo error de medición. Las personas responsables de las pruebas deberán analizar si el comportamiento de los errores de medición es el adecuado de acuerdo con el propósito del instrumento. En el ejemplo ilustrativo que se presenta seguidamente la estimación es mucho más precisa alrededor del promedio de dificultad de los ítems que usualmente se centra en 0.

Figura 6

Diagrama de dispersión para las variables estimación del constructo y su correspondiente error de medición en el modelo de Rasch



Fuente: Elaboración propia

Finalmente, aunque el propósito principal de estas notas no es profundizar sobre el tema de construcción y validación de ítems de respuesta construida o abierta, o pruebas de ejecución o desempeño, es relevante indicar que en estos casos la calificación de cada producto debe ser llevada a cabo por al menos dos jueces que hayan recibido la necesaria capacitación, y que se debe aplicar algún modelo de medición para generar indicadores de concordancia entre ellos, y otros indicadores de calidad técnica. Para estas situaciones es recomendable utilizar una extensión del modelo clásico del Rasch, el modelo de Rasch de Facetas Múltiples, o MFRM por su acrónimo en inglés (Many Facet Rasch Model).

Por último, una palabra de caución sobre el uso de pruebas comerciales. Es común en nuestro país que diferentes instancias compren los servicios de compañías que producen pruebas a nivel internacional, y que estas se utilicen para la toma de decisiones de alto impacto, con frecuencia, para la selección de personal en puestos laborales, por ejemplo.

En este contexto, es necesario indicar que, si se adquiere una prueba comercial de venta libre, hay una gran amenaza a la validez de las interpretaciones generadas a partir de su aplicación, pues existe la posibilidad, muy real, de que la prueba sea conocida por los examinados con anterioridad, dado que cualquiera puede comprarla (o bien alguna persona titulada en psicología), por ejemplo. A esta situación se le denomina sobre-exposición de la prueba y es una circunstancia que invalida cualquier resultado psicométrico generado originalmente. Esto porque las respuestas de los examinados solo reflejarán el haberse aprendido de memoria los ítems y no su verdadero nivel en el constructo de interés. La única forma de solventar esta situación es que exista un contrato específico con la compañía que elabora la prueba de tal forma que se firme un compromiso de confidencialidad para mantener cautivo el banco de ítems y no compartirlo con ningún

otro posible cliente, lo cual suele representar un costo prohibitivo para la mayoría de las instancias en nuestro país. Entonces, lo recomendable es no utilizar pruebas comerciales de venta libre para la evaluación de alto impacto en contextos educativos o laborales.

A partir de la discusión anterior, se derivan los siguientes estándares de calidad:

1. Contar en el equipo evaluador con especialistas en medición o psicometría, o en su defecto, estadística con conocimientos de modelos de medición
2. Realizar al menos una aplicación piloto con una muestra representativa de la población meta
3. Verificar la unidimensionalidad de la prueba por medio del Análisis Factorial Exploratorio (AFE) y tomar las decisiones pertinentes si la aplicación piloto refleja la existencia de más de un factor
4. Utilizar un modelo de medición psicométrico que sea apropiado para generar las interpretaciones de interés, según el propósito de la prueba (TCT no es apropiado, en general, para interpretaciones criteriales).
5. Utilizar tamaños de muestra en las aplicaciones piloto que permitan generar estimaciones precisas de acuerdo con el modelo de medición a utilizar.
6. Emplear niveles de dificultad de los ítems que sean apropiados al propósito de la prueba y a la población meta.
7. Según el modelo de medición, generar estimaciones de la precisión con que se está midiendo, para las personas si se usa la TCT y para personas e ítems si se usan modelos de TRI, incluyendo Rasch.
8. Utilizar el modelo de Rasch si se desea realizar interpretaciones criteriales (descripciones del desempeño del examinado de acuerdo con los resultados obtenidos en las pruebas).
9. Revisar y evaluar el ajuste de los examinados si se usa el modelo de Rasch.
10. Revisar y evaluar el ajuste de los ítems si se usa el modelo de Rasch.
11. Si se ha usado la TCT, revisar si los ítems presentan índices de discriminación adecuados.
12. Si se ha usado la TCT, verificar que el Alfa de Cronbach sea al menos 0,8.
13. Si se ha utilizado el modelo de Rasch, verificar que la confiabilidad de personas y de ítems sea al menos 0,8.
14. Para la aplicación definitiva u operativa, con la cual se tomarán las decisiones sobre los examinados, verificar que se haya ensamblado a partir de ítems con parámetros psicométricos conocidos y aceptables.
15. Abstenerse de usar pruebas comerciales por los efectos de sobreexposición, a menos que exista un contrato de exclusividad para mantener los ítems cautivos.
16. En pruebas con preguntas de respuesta abierta, construida, de ejecución o desempeño, contar con al menos dos jueces calificadores para cada pregunta y que hayan recibido la necesaria capacitación.
17. Utilizar el modelo de Rasch de Facetas Múltiples, o MFRM por su acrónimo en inglés (Many Facet Rasch Model) en pruebas de respuesta abierta, respuesta construida, ejecución o desempeño, estimar y valorar la confiabilidad entre jueces calificadores.

Referencias

Bond, T. & Fox, C. (2001). *Applying the Rasch model: Fundamental Measurement in the Human Sciences*. Laurent Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications.

Martínez, M. R., Hernández M.J. & Hernández, M.V. (2006). *Psicometría*. Alianza Editorial.

Montero, E. (2001). *La teoría de respuesta a los ítems: una moderna alternativa para el análisis psicométrico de instrumentos de medición*. Revista de Matemática: teoría y aplicaciones, 7 (1-2), 217-228.
DOI 10.15517/RMTA.V7I1-2.191

Muñiz. J. (2003). *Teoría Clásica de los Tests*. Madrid: Ediciones Pirámide.

Nunnally, J.C. & Bernstein, I.J. (1995). Teoría psicométrica (3ª ed). Editorial McGrawHill Latinoamericana.

Organización para la Cooperación y el Desarrollo Económicos (2004). *Informe PISA 2003: Aprender para el mundo del mañana*. Santillana Educación S.L.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. New Jersey, Estados Unidos de America: Lawrence Erlbaum Associates.

Sección 2.7

Desarrollo de un banco de ítems: Etapa 7

Karol Jiménez-Alfaro

En los apartados anteriores se ha indicado que las pruebas de alto impacto en el contexto académico y profesional se enmarcan en el referente teórico del constructo, el propósito y usos de sus resultados y las propiedades psicométricas de las pruebas estandarizadas. Es en este último referente donde el banco de ítems es una condición fundamental para garantizar la calidad psicométrica de cualquier instrumento de medición.

Un banco de ítems es un conjunto de ítems que miden el mismo dominio de conocimiento, o constructo, los cuales son almacenados en una computadora junto con las estimaciones empíricas de sus parámetros (Van der Linden & Eggen, 1986). Este banco incluye tanto los ítems como la información de su historial psicométrico. Específicamente, las estimaciones empíricas de parámetros como la dificultad y la discriminación de cada aplicación en el que se utilizó, categorías de clasificación, expectativas de solución, si son apropiados para diversas poblaciones (vg. personas en condición de discapacidad) y cualquier otro elemento que se considere relevante. El almacenamiento de los ítems y su información adicional puede hacerse en forma física con tarjetas, pero gracias a los avances tecnológicos y, en especial, por la necesidad de ensamblar pruebas informatizadas y por el desarrollo de pruebas adaptativas computarizadas (Reckase, M. D., 2010), lo usual es que se almacenen en equipos de cómputo.

La función fundamental de un banco de ítems es garantizar que las pruebas cumplan con las características psicométricas establecidas a priori por el equipo evaluador.

Usualmente, entre estas metas, están los niveles de dificultad y confiabilidad esperado de las pruebas. Así, no habrá sorpresas al aplicar la prueba si las estadísticas de los ítems son adecuadas y actualizadas con cada aplicación. Por ejemplo, en el caso de pruebas de certificación, se podría predecir qué porcentaje de la población quedará por encima del punto de corte, puntaje mínimo necesario para certificarse, si se cumplen con los parámetros de dificultad establecidos a priori.

Otra función del banco es facilitar el ensamblaje de fórmulas paralelas, las cuales deben tener las mismas características psicométricas para ser aplicadas en un mismo año o en años consecutivos. Con estas fórmulas se puede hacer mediciones según el nivel de habilidad de las personas examinadas; construir formularios con diferentes cantidades de ítems para ampliar o reducir el tiempo de aplicación según el propósito de la prueba y, en especial, orientar el trabajo de futuras investigaciones para su propia actualización.

El aplicar año con año una prueba con un mismo conjunto de ítems, y más aún si se realizan varias aplicaciones en un mismo año, podría traer la sobreexposición de los ítems. Por ejemplo, en las personas examinadas que no la aprueban en una primera vez, es posible que memoricen ítems y utilicen esta información en su próxima aplicación, o la compartan con otras personas que realizarán la prueba pronto, lo que afectaría la validez de los resultados (AERA, APA y NCME, 2014), en especial la medición apropiada del contenido de la misma, lo que afectaría la toma de decisiones, más cuando se trata de selección, promoción o acreditación de candidatos.

A continuación, se presentan algunos estándares para la construcción y mantenimiento de un banco de ítems de una prueba de alto impacto, no necesariamente computarizadas.

Para mayor información de los bancos informatizados para pruebas adaptativas se puede consultar Reckase (2010), Van der Linden & Veldkamp (2006), Van der Linden & Eggen (1986), entre otros. Los documentos o tarjetas a las que se hace referencia podrían ser en formato de papel o electrónico.

1. Elaborar un documento o manual con la descripción de los parámetros o características técnicas que deben cumplir los ítems para pertenecer al banco, los propósitos de la prueba, el marco teórico y el modelo psicométrico elegido para el análisis de los mismos, tales como, Teoría de Respuesta del Ítem (TRI), Teoría Clásica de los ítems (TCT), modelo de Rasch, entre otros.
2. Contar con una cantidad suficiente de ítems construidos según los fundamentos teóricos de la prueba y de probada calidad con su respectiva calibración. Los ítems que formarán parte del banco deben haberse experimentado previamente, ya sea a través de aplicaciones piloto en la población meta, principalmente, para la construcción de un nuevo banco; o bien, para aumentar los ítems de uno existente. Otra opción para experimentar nuevos ítems en pruebas diseñadas con un banco es incluirlos en una aplicación, pero en una cantidad mínima, entre un 10% y 15%, para no afectar la confiabilidad y validez de la de las inferencias de los resultados, además de no ser tomados en cuenta en la calificación de las personas examinadas.
3. Sistematizar en una tarjeta o documento digital la información del ítem: el enunciado y las opciones de respuestas, la clave y otros datos, acorde con los lineamientos de digitación y diagramación establecidos en protocolos, por ejemplo:
    - Un identificador del ítem, es decir, una numeración que podría ser alfanumérica para representar la categoría a la que pertenece el ítem. Por ejemplo, si se trata de una prueba de conocimientos matemáticos y el ítem es de la categoría de geometría, podría clasificarse como 135-MA-GEO.
    - Contar con un elemento identificador del tipo de población en que es apropiado su uso. Por ejemplo, si un ítem requiere una imagen en su solución, probablemente, no será adecuado para personas en condición de discapacidad visual, por lo que, es necesario indicar en la tarjeta o documento no usar en esa población.
4. Incluir un resumen con las estadísticas de los parámetros estimados en cada una de las aplicaciones. En el documento o tarjeta se debe incluir:
    - La numeración alfanumérica asociada a la tarjeta del enunciado y el número de folleto, si fuera el caso de que en una misma aplicación se utilicen dos o más folletos de la misma prueba.
    - La posición del ítem en el folleto.
    - Las estadísticas de los parámetros, gráficas o curvas, obtenidas en las aplicaciones previas, acorde con la teoría establecida para los análisis de los ítems, y que brinden información para los próximos ensamblajes.
5. Actualizar el banco de ítems, preferiblemente, después de cada aplicación. Las personas responsables de su mantenimiento tendrán entre sus tareas:

- Incorporar al banco los nuevos ítems que cumplen todas las características psicométricas.
- Actualizar la información estadística de aquellos ítems que conforman el banco, con los resultados obtenidos del último análisis.
- Eliminar aquellos ítems que, a partir los análisis psicométricos, no mantienen las características deseables para mantenerse en el banco.

6. Al tratarse de material confidencial, por seguridad, es importante crear protocolos para establecer las personas responsables de la actualización del banco y de otras tareas relacionadas con el material del banco.

Referencias

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *The Standards for Educational and Psychological Testing*. American Educational Research Association.

Barbero, M. I. (1996). Bancos de ítems. En J. Muñiz (Ed.): *Psicometría*. Editorial Universitas.

Reckase, M. D. (2010). *Designing item pools to optimize the functioning of a computerized adaptive test*. Psychological Test and Assessment Modeling, 52(2), 127-141. http://www.psychologie-aktuell.com/fileadmin/download/ptam/2-2010/01_Reckase.pdf

Van der Linden, W., Ariel, A., & Veldkamp, B. (2006). *Assembling a Computerized Adaptive Testing Item Pool as a Set of Linear Tests*. Journal of Educational and Behavioral Statistics, 31(1), 81-99. http://www.jstor.org/stable/3701289.

Van der Linden, W. & Eggen, T. (1986). *An Empirical Bayesian Approach to Item Banking*. Applied Psychological Measurement, 10(4), 345-354. https://doi.org/10.1177/014662168601000403

Sección 2.8

Ensamblaje de la prueba: Etapa 8

Carolina Chacón-Vega

Las pruebas de alto impacto son diseñadas a partir de un procedimiento técnico riguroso para garantizar su calidad, equidad y objetivad; como se ha detallado en el apartado anterior; por ello, también su ensamblaje debe ser producto de una aplicación meditada y sólida de principios (Cohen y Swerdlik, 2003). Esto implica que debe ser un proceso normado y riguroso, de acuerdo con el contexto de la evaluación, para definir los correspondientes lineamientos específicos.

Así las cosas, es un hecho que el ensamblaje debe ser tan riguroso como sea posible y que permita recoger información válida sobre la medición del comportamiento o ejecución del individuo, ya que, podría tener usos variados como diagnosticar una situación o determinar el nivel de logro, entre otros. Por ello, lo recomendable es que la prueba sea ensamblada siguiendo los criterios establecidos en un instructivo, elaborado por especialistas, los cuales deben ser claros y públicos para garantizar puntuaciones normalizadas y objetivas. En este apartado se abordan algunos estándares básicos de calidad para el ensamblaje de pruebas de alto impacto.

Validez en el ensamblaje de una prueba

Los resultados de cualquier instrumento de medición deben evidenciar un grado alto de validez con el fin de utilizar sus interpretaciones con toda confianza. Así lo mencionan Martínez, Hernández y Hernández (2006, pág. 14) como el aspecto más importante de una prueba que debe ser considerado desde muchos puntos de vista, dependiendo del uso que se pretende dar a los datos. Igualmente, Muñiz (2003) considera que "la pertinencia de las inferencias obedece al seguimiento de procedimientos científicos para la recogida de datos, nunca exhaustivos" (pág.151). De esta manera, el ensamblaje de una prueba debe:

- Ser un insumo para las interpretaciones realizadas a posteriori de las puntuaciones.
- Considerar la conceptualización y el contexto de aplicación de la prueba, ya que a partir de ahí se forja el juicio de validez de las inferencias.
- Responder a un contexto en particular, ya que, la respuesta a la pregunta de la existencia o no, de evidencias de validez, no es dicotómica, sino que se explica a través del proceso continuo de recolección de validez que inicia desde el ensamblaje.

En el primer apartado se profundiza en las evidencias asociadas a la validez de las inferencias y usos de las puntuaciones.

Confiabilidad en el ensamblaje de una prueba

Cualquier medición de variables tiene un error asociado como característica inherente. Este error debe ser minimizado en el mayor grado y se logra a través de la estimación de ese error para que las puntuaciones sean verdaderas, este es el objetivo de la confiabilidad (Muñiz, 1996, pág.4). Por lo tanto, cada vez que se aplique la prueba a poblaciones similares, sus resultados deben ser consistentes. En ese sentido, la consistencia de la medición no hace alusión a la estabilidad del constructo medido, ya que se parte del principio que la variable medida es estable; sino al cumplimiento de los parámetros psicométricos establecidos para el ensamblaje de una prueba de alto impacto, por ello debe:

- Mostrar resultados consistentes cuando es aplicada en diferentes oportunidades a diferentes poblaciones.
- Lograr un valor aceptable en el coeficiente de confiabilidad, de acuerdo con los propósitos y usos de la prueba, mediante la utilización de ítems con estadísticas psicométricas aceptables.

Para una mayor comprensión de esta propiedad psicométrica revisar el primer apartado de este documento.

Tabla de especificaciones en el ensamblaje de una prueba

Considerando que la tabla de especificaciones es una herramienta que legitima la prueba como una muestra representativa de los contenidos, a través de la cual se hace valida el contenido del instrumento (García, 2018) se reconoce que el ensamblaje de la prueba debe:

- Estar en función de la tabla de especificaciones, la cual tiene el reto de plasmar de manera exacta la complejidad cognitiva del constructo por medir.
- Seleccionar el tipo de ítem para cada uno de los objetivos por evaluar, por lo que, es necesario revisar la tabla de especificaciones antes del ensamblaje.
- Ser una representación fiel de la correspondencia absoluta que existe entre todos los elementos de la tabla de especificaciones.

En la sección 2.4 se explica con mayor detalle el proceso de construcción y validación de una tabla de especificaciones.

Intención meta de una prueba estandarizada

Como se destacó en el primer apartado, según el tipo de prueba, de los propósitos y de las características de la población se establece la intención meta. González (2001) puntualiza que "la identificación de una meta o propósito fundamental de la actividad evaluativa facilita establecer sus peculiaridades, y distinguirla de otras actividades que comparten muchos de sus procesos y fases". De esta manera, el nivel de dificultad de la medición de la conducta o del contenido es controlado a priori en el ensamblaje, por lo que, se debe:

- Establecer el nivel de dificultad de la prueba acorde con los propósitos de la evaluación.
- Considerar las características de las personas examinadas: rango de edad, nivel de lectura y factores culturales, entre otros, que pueden interferir en la resolución de la prueba.

Tabla de ensamblaje de la prueba

Es conveniente contar con un instrumento que facilite la concatenación de ítems en la prueba. Este instrumento constituye una tabla de ensamblaje, en la cual se consigna la distribución visualmente accesible de la ubicación de los ítems en la prueba y sus

correspondientes datos estadísticos y otra información relevante para que el equipo especialista ensamble la prueba, por lo que, se debe:

- Seleccionar los ítems según la proporción asignada para los de anclaje, experimentación y de banco, conforme con la distribución de la tabla de ensamblaje
- Permitir la graduación de la dificultad meta de la prueba al incluir o quitar ítems con diferentes índices de dificultad.

Tipo de ítems para el ensamblaje de la prueba

En la sexta etapa de este documento se describe con suficiente detalle los estándares para la redacción y validación de los ítems; por lo que en este punto se resalta aquellos aspectos relevantes para el embalaje de una prueba estandarizada, tales como:

- Seleccionar a personas con alto dominio de los criterios teóricos y técnicos para la revisión de la prueba durante el ensamblaje.
- Utilizar en el ensamblaje solo los ítems adecuados para interpretar la puntuación como una medida del constructo (Martínez, 2005). Esto es, que posean las calidades técnicas requeridas y parámetros estadísticos fundamentados, que no hayan sido concedidos anteriormente por error técnico y que cumplan con alguna normativa idiomática.
- Utilizar ítems de banco puesto que su calidad técnica se ha comprobado a partir de procesos rigurosos de construcción y juzgamiento en cuanto a la discriminación, dificultad y análisis de sensibilidad y otros análisis psicométricos de los ítems.

Revisión de la prueba

En la etapa de la revisión de la prueba es aconsejable tomar en cuenta los siguientes requerimientos.

- Planificar varias revisiones en diferentes momentos del ensamblaje con objetivos específicos, que contribuyan a una depuración continua del instrumento.
- Realizar una primera revisión interna, luego del ensamblaje, para que el equipo especialista evalúe la pertinencia, el formato de los textos o el estímulo utilizado, la redacción, la ortografía, claridad de los enunciados, los distractores y precisión de la clave en apego a la tabla de especificaciones.
- Seleccionar expertos en el constructo para la revisión del contenido y la pertinencia de los ítems de acuerdo con los propósitos del instrumento.
- Corroborar el cumplimiento de la distribución de ítems en la prueba, según la tabla de ensamblaje.
- Registrar en actas las sugerencias o comentarios para que el equipo evaluador realice los ajustes necesarios.
- Considerar una revisión filológica y de formato por un especialista.

- Realizar continuamente reuniones entre jueces y personas especialistas encargadas del ensamblaje para evacuar dudas y generar realimentación.

Diseño de pruebas

En el apartado anterior se explica con detalle el proceso técnico del diseño de pruebas, por lo que, en este punto interesa retomar algunos aspectos técnicos para el ensamblaje fórmulas paralelas o equivalentes, en las cuales se debe:

- Considerar el uso de diversas fórmulas cuando no se pueda aplicar todos los ítems existentes a toda la población.
- Valorar su uso cuando se pretenda analizar y comparar el desempeño general de los estudiantes en un área de contenido particular y en cada uno de los ítems.
- Permitir un análisis de costo-beneficio de la utilización de pruebas paralelas por medio de la obtención de conclusiones estadísticas robustas que se realicen a posteriori.

Formato administrativo-informativo en la prueba

En cuanto al formato de la prueba deben contemplarse las siguientes consideraciones:

- Tener una portada que contemple:

Logo de la institución encargada de la prueba.
Nombre de la institución encargada de la prueba.
Nombre de la prueba.
Día, mes y año de aplicación.
Tipo de prueba o número de fórmula de prueba.
La información básica y necesaria para la identificación de la persona examinada.
Espacio para la identificación y firma del delegado aplicador.

- Insertar una hoja informativa con aclaraciones pertinentes sobre la aplicación de la prueba: fechas de entrega de resultados y requisitos para realizar recursos de revocatoria.
- Contener un apartado de instrucciones generales para resolver la prueba: cantidad total de ítems, duración de la prueba y materiales necesarios.
- Presentación de la prueba: espacios adecuados entre un ítem y otro
- Contemplar las condiciones reales para su aplicación: el tiempo disponible para su administración y los diferentes tipos de ítems.
- Insertar anexos si se considera necesario.

Referencias

Cohen, R. & Swerdlik, M. (2002). *Pruebas y Evaluación Psicológicas.* Introducción a las pruebas y a la medición. https://www.scribd.com/document/461678157/Cohen-Swerdlik-Pruebas-y-evaluacion-psicologicas-Introduccion-a-las-pruebas-y-a-la-medicion-pdf

García, R. (2018). *Diseño y construcción de un instrumento de evaluación de la competencia matemática: aplicabilidad práctica de un juicio de expertos.* Ensaio: Avaliação e Políticas Públicas em Educação, 26 (99), 347-372. https://doi.org/10.1590/s0104-40362018002601263

González, M. (2001). *La evaluación del aprendizaje: tendencias y reflexión crítica.* Educación Médica Superior, 15(1), 85-96. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S08642141200100010&lng=es&tlng=es

Martínez, R. (2005) *Psicometría: Teoría de los tests psicológicos y educativos.* Editorial Síntesis.

Martínez, M.R, Hernández, M.J. & Hernández, M.V. (2006). *Psicometría.* Alianza Editorial.

Muñiz, J. (1996). *Psicometría.* Universitas.

Muñiz. J. (2003). *Teoría Clásica de los Tests.* Ediciones Pirámide.

Sección 2.9

Entrega de calificaciones y reporte de resultados a las personas examinadas: Etapa 9

María Paula Villarreal-Galera y
María Julieta Solórzano-Salas

Para efectuar el proceso de calificación y reporte de los resultados se recomienda complementar este apartado con el segundo capítulo del Cuaderno Metodológico (Villarreal, Alfaro y Brizuela, 2015), donde se detallan las distintas etapas de desarrollo de una prueba estandarizada. También se sugiere revisar los estándares propuestos por Educational Testing Service (2014). En este apartado se destacan aquellos procedimientos que se pueden adaptar al contexto académico y profesional costarricense, considerando las condiciones de calificación, las necesidades de información de las distintas audiencias y los usos de las pruebas.

De la generación de calificaciones

Cuando se realizan las estimaciones de los puntajes de las personas examinadas, algunas entidades poseen un generador automático de calificación que permite calcular los puntajes de manera mecanizada, lo cual es rápido, eficiente y minimiza los errores.

Para este procedimiento se utiliza una base de datos con las respuestas de cada una de las preguntas resueltas por las personas examinadas y una matriz conocida como plantilla de claves (archivo de las respuestas correctas). La precaución es siempre contar con la variable identificadora (como un código de barras) de las personas examinadas para reportar la información correcta; así como contar con la plantilla de clave revisada por, al menos, dos profesionales que verifiquen las respuestas correctas.

Por otro lado, no todos los programas de pruebas tienen acceso a sistemas informatizados diseñados a la medida, pero sí utilizan paquetes informáticos para estimar las calificaciones. A esto se le llama sistema manual de calificación, puesto que emplea computadoras y diversos softwares; sin embargo, no es completamente automatizado, por lo que existen fuentes de error que deben mantenerse controladas mediante diversas estrategias, por ejemplo, realizar la calificación al menos dos veces (de ser posible con dos profesionales diferentes) y comparar ambos resultados. Además, de contar con un manual de calificación que especifique la forma de obtener los puntajes, en el cual se recomienda:

1. Especificar los procedimientos y criterios para calificar la prueba, con el nivel de detalle y claridad que sea necesario.
2. Elaborar un protocolo para la lectura de hojas de respuestas que contemple dos procesos de lectura independientes y compare ambos resultados para comprobar que no existen errores en la lectura de las respuestas registradas por las personas examinadas. Además, se debe revisar una muestra aleatoria de las hojas leídas para detectar posibles errores. En aquellos casos en los que se utilice un mecanismo de digitación manual de las respuestas, se deberá realizar una revisión de cada una de las hojas para tener certeza de que estas fueron leídas correctamente, es decir, para asegurar que la veracidad de las respuestas de cada persona examinada en papel corresponde a las respuestas digitalizadas.

3. Elaborar la plantilla de claves para que el programa informatizado proceda a calificar las respuestas de las personas examinadas en cada versión de la prueba.
4. Revisar las posibles inconsistencias en la calificación de las pruebas que permitan reducir, tanto como sea posible, los errores en el procesamiento de las respuestas de las personas examinadas.
5. Documentar los procesos de escalamiento o equiparación entre las distintas versiones, en el caso de aplicar varias versiones o fórmulas de una misma prueba; así como la información relacionada con los ítems empleados como anclaje.
6. Explicitar los modelos psicométricos utilizados (Teoría Clásica de los Test, Teoría de Respuesta al Ítem, Modelo de Rasch, Teoría de la Generalizabilidad, Modelos de Diagnóstico Cognitivo, entre otros) para evaluar las propiedades psicométricas de los ítems. Según el modelo empleado, deben registrarse los parámetros pertinentes y considerados como aceptables de los ítems (ajuste del ítem, dificultad, discriminación, entre otros). A partir de estos análisis psicométricos previos, conceder a todas las personas examinadas los ítems que se identifiquen con baja calidad técnica o eliminarlos de la calificación.
7. Construir un protocolo para el análisis de los ítems con los correspondientes parámetros estadísticos para responder en términos de excelencia y equidad en la prueba. Para ello es necesario analizar los ítems con el fin de encontrar información relevante, tal como diferencias entre categorías de población (impacto), o bien, detectar DIF (diferente probabilidad de responder correctamente un ítem entre las personas examinadas con las mismas competencias, pero que pertenecen a distintas categorías poblacionales). Con esta información tomar decisiones, por ejemplo, si el ítem presentara DIF debería ser excluido.
8. Documentar la justificación y los procedimientos empleados para establecer determinados puntos de corte (como por ejemplo las notas mínimas de aprobación de un examen), cuando las interpretaciones hechas a partir de los puntajes de una prueba así lo demanden.
9. Establecer puntos de corte para definir categorías, considerando grupos de equiparación equivalente cuando el propósito de la prueba lo requiera, y realizar interpretaciones sustantivas. Esto se debe generar con base en datos empíricos adecuados sobre la relación entre el desempeño en la prueba y otros criterios relevantes. Este proceso debe permitir que las personas especialistas en los contenidos o constructo medidos por la prueba utilicen su conocimiento y su experiencia para establecer esos puntos de corte.
10. Documentar las estadísticas que sean necesarias para describir cómo se comportó la población en la aplicación de la prueba: comparaciones de calificación según variables relevantes, variancia explicada por los factores, indicadores e índices que permitan contribuir con las aclaraciones necesarias en términos de fiabilidad y validez de los resultados.

Del reporte de resultados

En relación con la entrega de resultados, tal como se indica en Villarreal, Alfaro y Brizuela (2015), es necesario que se reporten de la manera más comprensible y exacta, es importante indicar que se deben reportar las calificaciones a todas las partes interesadas (audiencias), con el objetivo de velar por los principios de transparencia y accesibilidad a la información. En un ejemplo puntual como una prueba estandarizada de selección de aspirantes que desean ingresar a una institución académica universitaria, se debe valorar las distintas audiencias a quienes reportar los resultados.

Una primera y evidentemente la más clara de ellas es cada persona examinada, a quien se le brinda un reporte detallado de su calificación. Una segunda audiencia serían las autoridades y una tercera el público en general. En estas dos últimas poblaciones no se presentarán datos individuales para mantener el resguardo de la confidencialidad de las personas examinadas.

En pruebas de certificación profesional también es importante rendir un informe de los resultados a las autoridades, responsables de la entrega de estas certificaciones. Normalmente, las autoridades no requieren conocer los resultados particulares de cada persona examinada, sino el análisis global de los resultados para orientar sus políticas, líneas de acción, planificación y decisiones entre otros.

En el III Congreso Latinoamericano de medición y evaluación educativa (COLMEE, 2018) se reflexionó sobre las pruebas estandarizadas en el campo de la evaluación de los aprendizajes en el contexto latinoamericano, donde los desarrolladores de pruebas estandarizadas tienen una deuda en la comunicación efectiva de los resultados que incidan directamente en la mejora y la validez de la toma de decisiones, sea en la calidad educativa, o bien, en que las personas evaluadas realmente puedan comprender el alcance de los resultados. En este marco del COLMEE se concluye que el coste de las pruebas estandarizadas es muy alto y el impacto de sus resultados en la toma de decisiones es bajo, en función de buscar respuestas efectivas a las debilidades recurrentes que cada instancia evaluadora conoce.

Del reporte de resultados a las personas examinadas

En esta etapa es imprescindible resguardar y mantener la confidencialidad de los resultados, de forma que se entreguen únicamente a la persona evaluada. El reporte de resultados será un documento privado, se debe advertir a la persona examinada que cuando sus datos se compartan, serán en forma conjunta con los resultados de las demás personas examinadas, sin que exista manera de reconocer de quién es cada nota, sino para efectos meramente de análisis estadístico e investigación, sin que estos puedan asociarse a la persona examinada.

Para el reporte de resultados se puede considerar las siguientes recomendaciones:

1. Reportar a las personas examinadas los resultados en un periodo razonable y de manera comprensible.
2. Informar los resultados conforme lo establecido por la legislación vigente sobre la protección de datos sensibles de las personas evaluadas.
3. Garantizar la accesibilidad de la información a poblaciones en condición de vulnerabilidad debido a condiciones de discapacidad o que requieran apoyos.
4. Responder a las apelaciones de los resultados (ante las instancias correspondientes indicadas por los desarrolladores de las pruebas), que presenten las personas examinadas en un periodo razonable.
5. Respetar el derecho de las personas examinadas a recibir respuestas tan explícitas como sea necesario para comprender el desacierto o el acierto de sus respuestas.
6. Corregir interpretaciones erróneas de los puntajes de la prueba, así como lidiar con consecuencias inesperadas del uso de la prueba (ver apartado 1.1.2.5 del Cuaderno Metodológico, Villarreal et al. 2015).
7. El equipo evaluador responsable debe evitar la identificación individual de las personas examinadas.
8. Reportar información acerca del tamaño de la muestra o población y la forma o dispersión de la distribución de los puntajes, cuando se divulguen en contextos educativos.
9. Tomar en cuenta las siguientes recomendaciones para elaborar el documento oficial que comunica de manera personal el desempeño del examinado en la prueba:

- Incluir datos generales: nombre completo e identificación de la persona examinada, así como la fecha de aplicación de la prueba.
- Dar un resumen sobre el constructo medido e indicar la cantidad de ítems incluidos en la prueba.
- Describir el propósito de la prueba.
- Presentar el puntaje total obtenido en la prueba y desagregarlo, cuando corresponda, en cada uno de los constructos o dimensiones en el caso de pruebas multi dimensionales. En pruebas de selección única es necesario presentar el desglose de respuestas en orden, indicando correctas e incorrectas. En el caso de rúbricas se debe desglosar cada uno de los rubros y su respectivo puntaje.
- Detallar el cálculo de la nota con las fórmulas matemáticas empleadas, de una manera clara para que puedan ser fácilmente entendidas.
- Describir el tipo de interpretación de los puntajes, si es una prueba referida a criterios o a normas. Es importante que en ambos casos se detalle con claridad cómo interpretar la nota obtenida.
- Indicar el periodo de vigencia que tendrá el resultado. Hay casos en los que la vigencia de la calificación es de 1 año o 2 años.
- Realizar aclaraciones cuando el equipo evaluador lo considere necesario. Por ejemplo, cuando no es posible comparar los resultados de la prueba de un año con otro, o bien, la prueba incluye ítems experimentales y otros detalles que deban conocer las personas examinadas.

- Indicar correos electrónicos, teléfonos y aquellos medios de contacto que sean convenientes para que la persona examinada solicite información adicional o realice apelaciones en caso de que la requiera.

Del reporte de resultados en informes ejecutivos para otras audiencias

A lo largo de este documento se han descrito las normas para el diseño y la construcción de una prueba estandarizada con puntuaciones válidas y confiables que permita la toma de decisiones a las instancias vinculadas, sin este proceso, las mediciones no tendrían razón de ser, en esta línea Rodríguez et al. (2016, p. 15) indican que: "La evaluación debe difundir sus resultados en un reporte o informe de acceso público, manifestando una especial preocupación por la devolución y entrega de información a los actores involucrados con el objeto de la evaluación."

Ahora bien, la devolución de resultados para otras audiencias interesadas en las pruebas estandarizadas y sus resultados se da generalmente mediante informes escritos, a modo de resumen ejecutivo, que de acuerdo con Ferrer (2006, p.35) pueden contemplar los siguientes estándares.

1. Incluir el significado de los diferentes puntajes y las habilidades asociadas a estos para la interpretación oportuna de los resultados, el tipo de prueba, sea referida a normas y criterios y las interpretaciones específicas que se están buscando a partir de los resultados.
2. Describir las evidencias de validez y confiabilidad, incluyendo los análisis psicométricos, que se explican en este documento.
3. Demostrar que los resultados obtenidos corresponden al conjunto de conocimientos, atributos o habilidades establecidos en la tabla de especificaciones de la prueba.
4. Tomar en consideración los aspectos técnicos para la comparación entre grupos de examinados, tales como la equiparación de puntajes para fórmulas diferentes de examen.
5. Hacer uso de los resultados conforme con los propósitos de la prueba.
6. Comunicar en un lenguaje accesible los aspectos de diseño, análisis y resultados de la prueba.
7. Detallar la información del contexto en el que se realizó la prueba: recinto, fecha, duración, dificultades en la aplicación, imprevistos, condiciones de los examinados, caracterización de las poblaciones evaluadas y apoyos utilizados para personas con necesidades educativas o con condiciones asociadas a discapacidad.
8. Sistematizar un protocolo de entrega de resultados, que a su vez permita a la instancia evaluadora certificar la entrega oportuna de la información a cada persona evaluada que incluya: cronograma, vías de comunicación y entrega, comprobante de acuse de recibido sea físico o digital, entre otras.
9. Identificar la utilidad de la información recopilada para la toma de decisiones según el propósito de la prueba. Por ejemplo, en pruebas para la evaluación de aprendizajes se debe informar a las autoridades y responsables para que atiendan las deficiencias detectadas en las personas examinadas, o bien, planteen ajustes al diseño de la prueba como tal.

10. La divulgación de los resultados de las pruebas en espacios de comunicación masiva debe realizarse cuidando la rigurosidad técnica.
11. Realizar periódicamente actividades de educación y divulgación de la naturaleza de la prueba y su uso para diferentes audiencias meta.

Referencias

Educational Testing Service (2014). *ETS standards for quality and fairness*. Author. https://www.ets.org/s/about/pdf/standards.pdf

III Congreso Latinoamericano de medición y evaluación educativa COLMEE (10 al 12 de mayo del 2018). *El desarrollo de la cultura de la evaluación en América Latina*. Uruguay. http://colmee.uy/es/Pages/program/presentaciones

Ferrer, G. (2006). *Sistemas de evaluación de aprendizajes en América Latina balance y desafíos*. http://200.6.99.248/~bru487cl/files/libros/Evaluacion_AL/Ferrer_eval.pdf

Rodríguez, P., Martinic, S., Soberón, L., Desirée, S., Guzmán, A., & Tapella, E. (2016). *Estándares de Evaluación para América Latina y el Caribe*. Akian Gráfica Editora. https://www.researchgate.net/publication/308143662_Estandares_de_Evaluacion_para_America_Latina_y_el_Caribe

Villarreal, M.P., Alfaro-Rojas, L. & Brizuela, A. (2015). *Construcción de pruebas estandarizadas en el ámbito de la medición educativa y psicológica*. Serie Cuadernos Metodológicos del Instituto de Investigaciones Psicológicas. Universidad de Costa Rica. http://iip.ucr.ac.cr/sites/default/files/contenido/cuaderno.pdf

Sección 2.10

Recomendaciones para la aplicación de pruebas estandarizadas

Esta sección se compone de dos apartados, uno sobre buenas prácticas en la aplicación de pruebas estandarizadas y el otro relacionado con la aplicación de pruebas psicológicas en contextos de certificación y selección.

Buenas prácticas en la aplicación de pruebas estandarizadas

Eulalia Gómez-González y
Helen Picado-Barrantes

El procedimiento en la aplicación de una prueba depende tanto de las características del instrumento (individual o colectivo, contra tiempo o no, lectura o ejecución) como de las características de las personas (edad, nivel educativo, condición física y mental, entre otros), por lo cual, se requiere tener presente estas condiciones para disminuir los factores que pueden afectar el desempeño en la prueba (Aiken, 1996). Partiendo de esta aseveración, el presente apartado pretende servir de guía para la implementación de buenas prácticas dirigidas a prevenir la afectación del desempeño durante la aplicación de las pruebas de alto impacto en nuestro contexto académico y profesional.

Por ello es fundamental que el equipo evaluador a cargo establezca un protocolo o una guía de buenas prácticas que describa las acciones y etapas a ejecutar en el proceso, así como aspectos relativos a responsables y materiales. Las siguientes recomendaciones podrían ser tomadas en cuenta para distintas aplicaciones: individual, colectiva a un solo grupo y colectiva simultánea a varios grupos. Además, de las expuestas en el apartado de validez y equidad, cuando la persona evaluada requiera de algún tipo de apoyo.

De los materiales

1. Asegurar el uso de pruebas no fotocopiadas y materiales en buen estado para la aplicación.
2. Realizar oportunamente la convocatoria y la confirmación de las personas examinadas para lograr una aplicación exitosa. Se sugiere la elaboración de un protocolo en este proceso.
3. Enviar por medios electrónicos o documento impreso las citas de las pruebas. Este comunicado debe indicar el nombre de la persona, el número de identificación, la fecha, la hora y el lugar de la aplicación, así como cualquier otro dato que se requiera y sea pertinente.
4. Comunicar el procedimiento a seguir a las personas que requieran de algún apoyo educativo, así como contar con un protocolo que oriente la atención a estos casos, según lo establecido en el apartado de validez y equidad.
5. La persona coordinadora, en aplicaciones colectivas, es la responsable del retiro y la entrega del material de la sede a su cargo.

De la preparación del equipo aplicador

1. Realizar un entrenamiento previo a la persona aplicadora de la prueba (Hossiep y Paschen, 1998). En los casos de aplicaciones colectivas, la persona coordinadora asume esa responsabilidad con todo el personal a su cargo.
2. Entrenar al personal de apoyo para que sea capaz de aplicar pruebas cuando sustituye a una persona aplicadora en el salón, asistir a la coordinación en la entrega y recepción de materiales y mantener el orden en las instalaciones. En estos casos de aplicaciones colectivas, la persona coordinadora es quien asume esta capacitación.

3. Contar con el mayor número de personas aplicadoras para garantizar las condiciones adecuadas de aplicación, según la cantidad de personas presentes en un solo grupo, para disminuir la tentativa de fraude (AERA, APA y NCME, 2014; Aiken, 1996, 2003).
4. No participar en la aplicación en caso de ser un futuro postulante a dicha prueba o ser un familiar hasta segundo grado de alguna persona evaluada, con el fin de evitar conflictos de intereses.
5. Bajo ninguna circunstancia podría aplicar pruebas, quien se encuentre en un estado o condición de salud que afecte su capacidad física o mental para la aplicación adecuada de las pruebas.
6. Conocer tanto las características de la población como de la prueba, esto con el fin de asegurar las condiciones adecuadas para la aplicación.
7. Conocer el instrumento y las instrucciones de uso, previo a la aplicación de la prueba estandarizada (Aiken, 1996).
8. Firmar un compromiso de confidencialidad.
9. Conformar un equipo de trabajo en los casos de una aplicación colectiva de las pruebas, con la asignación de las correspondientes responsabilidades de las personas coordinadoras, aplicadoras y de apoyo.

De la preparación del salón

1. Asegurar que las condiciones de higiene, iluminación, ruido, espacio, ventilación, temperatura (Aiken, 1996, 2003), seguridad y acceso al lugar sean las adecuadas para la aplicación (AERA, APA y NCME, 2014).
2. Ordenar el salón con suficiente tiempo y de acuerdo con lo requerido para la prueba, la cantidad y condiciones de las personas evaluadas. Se debe asegurar que el lugar cumpla las condiciones mínimas de distracción para evitar varianza irrelevante al constructo (AERA, APA y NCME, 2014).
3. Colocar todos los materiales a utilizar según la guía o protocolo indicado (AERA, APA y NCME, 2014).
4. Colocar el mobiliario según el protocolo para reducir la posibilidad de tentativa de fraude (AERA, APA y NCME, 2014). También se pueden usar diferentes fórmulas (reactivos distintos o en orden diferente) de la prueba y distribuirlas de manera alterna para que queden distanciadas entre sí (Aiken, 1996, 2003).
5. Garantizar las mismas condiciones en todos los salones, en caso de aplicaciones colectivas, de acuerdo con los respectivos instructivos.
6. Rotular cada salón junto con la lista de las personas examinadas para indicar quienes están autorizados a ingresar.

De la aplicación de la prueba

1. Permitir el ingreso a las instalaciones, en caso de una aplicación colectiva, en un tiempo prudencial y previamente establecido, con el fin de que las personas evaluadas puedan ubicar el salón asignado.

2. Autorizar el ingreso al salón cuando la persona coordinadora, en casos de aplicación colectiva, lo indique al equipo aplicador.
3. Autorizar el ingreso de las personas evaluadas al salón de aplicación luego de verificar su identificación con el documento oficial y se puede hacer uso de un acta de firma.
4. Ubicar ordenadamente a las personas evaluadas, una a una, en el lugar correspondiente.
5. Solicitar a las personas evaluadas apagar sus dispositivos móviles o cualquier dispositivo electrónico, antes de iniciar la prueba.
6. Leer el consentimiento informado a las personas examinadas, cuando corresponda. Asimismo, aclarar cualquier duda al respecto y solicitar por escrito la firma de la persona adulta a evaluar. En el caso de personas menores de edad se hará según la capacidad progresiva de la persona lo cual deberá estar establecido en un protocolo.
7. Se recomienda entregar una copia del consentimiento informado a la persona evaluada, el cual debe explicar la finalidad de la prueba, el manejo de los datos, los derechos y deberes de las personas evaluadas, así como el número de contacto de la entidad o la persona responsable de la aplicación.
8. Brindar las instrucciones necesarias para asegurar que las personas examinadas realicen la prueba de manera correcta, así como la entrega del material. Se recomienda elaborar un instructivo. Una vez iniciada la aplicación no se deben atender consultas (Hossiep y Paschen, 1998).
9. Facilitar una copia de las instrucciones a cada persona evaluada en aplicaciones colectivas.
10. Asegurar la comprensión de las instrucciones, por lo cual, la persona aplicadora puede comprobar el seguimiento de dichas instrucciones, sin que ello implique una fiscalización de las respuestas (Fernández, Seisdedos y Mielgo, 2001).
11. Indicar la hora de inicio y finalización de la prueba, según lo establecido en el manual (Hossiep y Paschen, 1998).
12. No ausentarse del salón sin dejar a otra persona aplicadora a cargo.
13. Seguir lo establecido en el protocolo para el retiro de personas evaluadas o interrupción de la prueba por situaciones de emergencia o desastre.
14. Seguir las instrucciones de manera estricta referente a la aplicación de la prueba prestando atención a los límites de tiempo, el completar los datos y las hojas de respuesta por parte de las personas examinadas (Aiken, 1996).
15. Crear un ambiente adecuado (rapport) para disminuir la tensión y propiciar el mejor desempeño de las personas examinadas (Aiken, 1996).
16. Velar por el resguardo de los materiales bajo su responsabilidad (AERA, APA y NCME, 2014).

17. Utilizar una vestimenta que no constituya un distractor para las personas evaluadas.
18. Portar los instrumentos adecuados para contabilizar el tiempo de aplicación. Asimismo, la persona aplicadora debe controlar el tiempo de acuerdo con la hora oficial.
19. Prestar atención a lo que sucede en el salón y en las condiciones del entorno, a fin de poder responder de manera oportuna a cualquier situación que se presente. En las aplicaciones colectivas, la persona coordinadora es quien resuelve cualquier situación atípica que se presente en la sede durante el proceso de aplicación.
20. Recoger los materiales entregados a la persona examinada cuando esta concluye la prueba. En caso de aplicaciones colectivas la entrega será de manera individual, en orden y de la forma más silenciosa para evitar interrumpir a quienes aún no han finalizado.
21. Realizar en una sola fecha la aplicación y en un mismo lugar, pero en casos en que la cantidad de personas examinadas supere la capacidad instalada se pueden programar dos o más fechas y lugares.

De la finalización de la prueba

1. Establecer un protocolo con los pasos a seguir cuando concluye la aplicación de la prueba para garantizar el orden y la recolección completa de los materiales.
2. Definir un medio que garantice a la persona evaluada que estuvo presente durante la aplicación de la prueba y que culminó o no la misma. Algunos medios sugeridos son: la entrega de comprobantes y el uso de actas de firma a la salida.
3. Guardar los materiales y proteger su integridad, en caso de que así esté establecido, conforme con el protocolo dado por la organización responsable.
4. En aplicaciones colectivas cada persona aplicadora entrega las pruebas y el material a la coordinación o al personal de apoyo para su conteo (Aiken, 1996, 2003) y revisión, con el objetivo de asegurar que no falte nada.

Referencias

Aiken, Lewis R. (1996). *Test Psicológicos y Evaluación (8ª. Edición)*. Prentice Hall Inc.

Aiken, Lewis R. (2003). *Test Psicológicos y Evaluación (11ª edición)*. Pearson Educación.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. American Psychological Association.

Fernández, J. L., Seisdedos, N. & Mielgo, M. (2001). *CPS: cuestionario de personalidad situacional: manual (2ª edición, revisada)*. Publicaciones de Psicología Aplicada, 243. TEA Ediciones.

Hossiep, R. & Pachen, M. (1998). *Manual del Inventario Bochum de Personalidad y Competencias* (Trad. Arribas, David; Corral, Sara y Pereña, Jaime). Hogrefe-Verlag GMBH & Co. KG.

Aplicación de pruebas psicológicas en contextos de certificación y selección

Fiscalía del Colegio de
Profesionales en Psicología de Costa Rica

En relación con las pruebas psicológicas estandarizadas de aplicación exclusiva por parte de profesionales en Psicología, existen lineamientos establecidos en lo que respecta a su construcción, aplicación, interpretación, protección, entre otros aspectos. Una aplicación apropiada de las pruebas psicológicas se establece en los mismos manuales de las pruebas psicológicas, y los principios deontológicos que se consignan en capítulo III del Código de Ética y Deontológico del Colegio de Profesionales en Psicología de Costa Rica.

En este apartado se presentan algunos de los lineamientos existentes para la aplicación de pruebas psicológicas estandarizadas en contextos académicos y profesionales de selección y certificación, tomando como base el Código de Ética y Deontológico del Colegio de Profesionales en Psicología de Costa Rica (2019), así como los Principios Éticos y Código de Conducta (Ethical Principles of Psychologists and Code of Conduct) de la Asociación Americana de Psicología (APA por sus siglas en inglés, 2010), y de la Sociedad Australiana de Psicología (2010), relacionados con el uso de pruebas psicológicas.

Otros tipos de pruebas para uso clínico, diagnóstico, peritaje forense y otros específicos, deberá consultarse el Código de Ética y Deontológico del Colegio de Profesionales en Psicología de Costa Rica, particularmente en los artículos 28, 29, 30 y 31 de dicha normativa.

Del material de las pruebas

En primera instancia, es necesario considerar los siguientes aspectos referentes al material de pruebas psicológicas:

1. De acuerdo con APA (2010), el término "materiales de prueba" se refiere a manuales, instrumentos, protocolos y preguntas de las pruebas o estímulos, y no incluye la información de la prueba. La persona profesional en Psicología debe realizar esfuerzos razonables para mantener la integridad y seguridad de los materiales de las pruebas y otras técnicas de evaluación, en forma consistente con la ley y obligaciones contractuales, y de una manera que permita la adherencia a los Principios Éticos y Código de Conducta de la APA (2010).
2. Los profesionales en Psicología, deben evitar el uso de materiales de pruebas que estén desactualizadas, o bien, que se encuentren obsoletas o que sean no útiles para efectuar evaluaciones o decisiones de intervención o recomendaciones (APA, 2010).
3. Los profesionales en Psicología deben contar con suficientes evidencias de validez para los propósitos de la prueba psicológica que pretenden utilizar, especialmente, para selección y certificación en contextos académicos y profesionales.

De los usos de Pruebas Psicológicas

Según la APA (2010, p. 12) las personas profesionales en Psicología son quienes deben administrar, adaptar, calificar, interpretar o emplear técnicas de evaluación, como las pruebas (tests), a la luz de la investigación o en evidencia de la utilidad y aplicación apropiada de las mismas.

Además de lo anterior, la persona profesional en Psicología debe utilizar instrumentos de evaluación cuya validez y confiabilidad se ha establecido para el uso de los miembros de una población evaluada; en caso de que ambas características no han sido establecidas, deberá describir las fortalezas y limitaciones de los resultados e interpretaciones de las pruebas utilizadas. Respetando el derecho de las personas evaluadas a conocer la naturaleza, propósitos, usos y consecuencias de las mismas, así como conocer la relación entre el objetivo que se persigue y los contenidos medidos por la prueba.

De la aplicación de las pruebas psicológicas

Las personas profesionales en psicología deben:

1. Respetar los derechos de las personas evaluadas relacionados con la voluntariedad, confidencialidad y devolución de la información.
2. Ser personas profesionales competentes, incorporadas y activas en el Colegio de Profesionales en Psicología de Costa Rica.
3. Demostrar un nivel de capacitación y actualización en medición, evaluación y psicometría, según se establece en la normativa, específicamente en los artículos 5, 23, 28, 29, 30 y 31 del Código de Ética y Deontológico del Colegio de Profesionales en Psicología de Costa Rica, para el uso y aplicación de pruebas psicológicas.
4. Prohibir a personas no calificadas el uso de técnicas de evaluación psicológicas de aplicación exclusiva por profesionales en Psicología, excepto cuando dicho uso sea conducido para propósitos de entrenamiento, con la supervisión adecuada (APA, 2010). Al respecto, el Código de Ética y Deontológico del Colegio de Profesionales en Psicología de Costa Rica establece en su artículo 20 que:

> "La persona colegiada no deberá suministrar técnicas, instrumentos o materiales de medición psicológica, específicos de la disciplina, ni instruir sobre su uso a quienes no tengan habilitación para ello, con excepción de los fines de formación profesional y académica en el área de la psicología, la cual deberá ser impartida por una persona profesional debidamente incorporada al Colegio; lo anterior, salvo las excepciones derivadas, legalmente, de la autonomía universitaria" (2019, p.11)

Por otra parte, según la APA (2000) y la Comisión Internacional para el uso de los Test (2013), las pruebas psicológicas se categorizan en tres niveles: A, B, y C. Siendo de aplicación exclusiva por personas profesionales en Psicología, aquellas categorizadas como B y C, de acuerdo con las normativas que rigen el ejercicio profesional en nuestro país, potestad delegada al Colegio de Profesionales en Psicología de Costa Rica.

5. Brindar los apoyos requeridos a las personas en condición de discapacidad para garantizar un acceso equitativo al contexto de aplicación, equivalente a las personas evaluadas regulares; de manera que, los apoyos no generen ventajas ni desventajas con respecto al resto de la población, sino una equivalencia en términos de oportunidades para mostrar el desempeño. Así, la comparación de los puntajes de todas las personas evaluadas resulta más legítima en cuanto a las interpretaciones sobre el nivel de

cada uno, en el constructo o atributo que se mide.

6. Evitar la sobreexposición de las personas evaluadas, a una prueba psicológica, debido a que la familiarización con el contexto de las pruebas puede afectar las inferencias de sus resultados; por ejemplo, en algunas de las pruebas psicológicas según refieren sus manuales de uso, deben ser administradas hasta al menos un año después.
7. Establecer condiciones de temperatura apropiada, adecuada ventilación que permita el recambio de oxígeno y el mobiliario del lugar debe ser cómodo.
8. Adecuar la iluminación de forma tal, que la persona evaluada pueda leer sin dificultad, ya sea de día o noche.
9. Velar por que la persona evaluada esté tranquila, siendo de vital importancia atender las recomendaciones que disminuyan la ansiedad o inquietud, tales como: explicar el objetivo y aspectos relevantes del instrumento de medición.
10. Proveer el material requerido a la persona evaluada para la realización de la prueba (e.g., lápiz, lápices de color, borrador, hojas, entre otros).
11. En el caso de pruebas automatizadas la persona profesional deberá apegarse a los manuales de los instrumentos, así como a los lineamientos generados por el Colegio de Profesionales en Psicología de Costa Rica.
12. Que la persona evaluada posea entrenamiento en lo referente a la relación gestual y no verbal, dado que estas pueden influir en el estado emocional del evaluado.
13. Tener presente el nivel de lecto-escritura de la persona evaluada, siendo que el mismo puede afectar el estado emocional e interpretación de los reactivos.
14. Antes de iniciar la aplicación, es importante cerciorarse de que la persona evaluada comprenda a cabalidad las consignas de ejecución.
15. La persona evaluada deberá estar en un estado psicológico funcional, que le permita responder los reactivos, por lo cual se deben descartar estados psicóticos, influencia de drogas o alcohol.
16. La persona profesional debe de seguir las instrucciones, recomendaciones y contraindicaciones del protocolo o manual de aplicación, siendo de importancia que este familiarizado con los instrumentos y sus características.

De la calificación e interpretación de resultados

Las personas profesionales en Psicología deben:

1. Asumir la responsabilidad de la aplicación, interpretación y uso apropiado de instrumentos de evaluación. (APA, 2010).
2. Tomar en cuenta el propósito de la evaluación, las habilidades de resolución y las características de la persona evaluada, como diferencias culturales, lingüísticas, personales y situacionales, que podrían afectar el juicio o reducir la precisión de las interpretaciones, incluyendo las automatizadas. La persona profesional en Psicología debe indicar cualquier limitación significativa en sus interpretaciones (APA, 2010).
3. Interpretar los resultados de pruebas psicológicas de respuesta típica (rasgos, valores, actitudes, creencias, etc.) con base en el modelo referido a normas; esto es, en la

posición relativa del evaluado con base en un grupo de referencia, por lo que el uso de puntos de corte invalida cualquiera interpretación sustantiva de los resultados. Los rasgos psicológicos propios de las personas evaluadas deben interpretarse a partir de los perfiles meta. No es posible indicar una puntuación mínima máxima o aceptable de logros, como si se justifica en el caso de pruebas educativas, en donde si existen respuestas correctas e incorrectas (APA, 2010).

Finalmente, resulta relevante indicar que el uso inapropiado de las pruebas psicológicas estandarizadas acarrea importantes riesgos para las personas evaluadas, la persona profesional en Psicología y la disciplina, entre los que se destacan: el estrés personal; la percepción personal inadecuada a lo largo de la vida a partir de etiquetas o información incorrecta que se suministre; decisiones de vida y de carrera pobremente informadas; amenazas a oportunidades de vida y consecuencias en la autopercepción, amenazas al desarrollo personal y profesional debido al uso incorrecto de resultados de funcionamiento cognitivo e intelectual.

Referencias

American Psychological Association. (2000). *Report of the Task Force on Test User Qualifications.* https://www.apa.org/science/programs/testing/qualifications.pdf

American Psychological Association (2010). *Ethical Principles of Psychologists and Code of Conduct*. http://www.apa.org/ethics/code/principles.pdf

Colegio de Profesionales en Psicología de Costa Rica (2019). *Código de Ética y Deontológico del Colegio Profesionales en Psicología de Costa Rica.* https://psicologiacr.com/sdm_downloads/codigo-de-etica-y-deontologico-del-cppcr-reforma-2019

International Test Commission. (2013). *ITC Guidelines on Test Use.* https://www.intestcom.org/files/guideline_test_use.pdf

The Australian Psychological Society (2010). *Submission to the Psychology Board of Australia Consultation Paper*. Options for the protection of the public posed by the inappropriate use of psychological testing. http://www.psychologyboard.gov.au/documents/default.

UNIVERSIDAD DE COSTA RICA

IIP Instituto de Investigaciones Psicológicas