doi: https://doi.org/10.15446/rcp.v31n2.93018

Desarrollo del Instrumento para Evaluar la Calidad Técnica de Pruebas Psicológicas

AURA NIDIA HERRERA ROJAS
Universidad Nacional de Colombia, Colombia

FLOR ANGELA LEÓN GRISALES
Universidad Nacional de Colombia, Colombia

Como citar este artículo: Herrera Rojas, A. N. & León Grisales, F. A. (2022). Desarrollo del instrumento para evaluar la calidad técnica de pruebas psicológicas. *Revista Colombiana de Psicología,* 31(2), *65-76.* https://doi.org/10.15446/rcp.v31n2.93018

Correspondence concerning this article should be addressed to anherrerar@unal.edu.co

SCIENTIFIC RESEARCH ARTICLE

RECEIVED: JANUARY 24, 2021 – ACCEPTED: APRIL 04, 2022

Resumen

Una de las estrategias utilizadas para garantizar procesos de evaluación psicológica de acuerdo con altos estándares éticos y técnicos consiste en evaluar la calidad de las pruebas que se utilizan en tales procesos y divulgar ampliamente la información, para que los profesionales usuarios de pruebas tengan suficientes elementos de juicio a la hora de decidir sobre la selección y uso de las mismas. Este trabajo tuvo como objetivo desarrollar un instrumento para valorar la calidad técnica de las pruebas publicadas en español, utilizadas en Colombia. Con base en la revisión de seis modelos de evaluación de pruebas (alemán, español, holandés, británico, brasileño y el de la Federación Europea de Asociaciones de Psicólogos, EFPA) y la identificación de las pruebas más usadas en Colombia (Herrera, & León, 2015), se adelantó un proyecto colaborativo dentro del cual se conformó un grupo de expertos con participación de instituciones de todo el país para desarrollar la primera versión del instrumento; ésta fue sometida a revisión por jueces y se adelantó una aplicación piloto con seis de las pruebas más usadas. Este documento presenta en detalle la estructura, los indicadores de calidad y los criterios de evaluación que conforman el instrumento que se constituye en el primer modelo para Colombia..

Palabras clave: calidad de pruebas psicológicas, modelo de evaluación de calidad, revisión de pruebas psicológicas.

Development of a Model to Evaluate the Technical Quality of The Psychological Tests

Abstract

One of the strategies frequently used to sure that the psychological evaluation processes is carried out with high ethical and technical standards, consists of evaluating the quality of the tests used in such processes and widely spreading the information, so, the professionals would have useful information to make decisions about its selection and use. The main aim of this work was to develop an instrument to review the technical quality of the psychological tests. Based on the review of six test evaluation models (German, Spanish, Dutch, British, Brazilian and that of the European Federation of Psychologists Associations, EFPA) and the identification of the most used tests in Colombia (Herrera & León, 2015), a collaborative project was carried out. A group of experts was formed with the participation of institutions from all over the country to develop a first version of the instrument; this version was subjected to reviews by judges and a panel of experts, and a pilot application was carried out with six of the most used tests. This document presents in detail the structure, the quality indicators and the evaluation criteria of the instrument that constitutes this model

Keywords: model to review quality, psychological tests review, technical quality evaluation.

Introducción

LAS PRUEBAS psicológicas son los instrumentos de medición por excelencia en psicología. Además de su utilidad en la investigación son las herramientas que proveen información para sustentar el criterio profesional en diagnósticos, identificación de riesgos, planeación de estrategias de intervención, seguimiento de procesos o tratamientos, etcétera. Los resultados de las pruebas se utilizan como información para sustentar evaluaciones que apoyan decisiones que afectan la vida de los evaluados, por lo tanto es crucial que el profesional seleccione las pruebas que tengan la calidad técnica que respalde su uso en la población de interés y para el propósito del proceso evaluativo.

Varios autores (Bartram, 1998; Muñiz et al., 2001; Evers et al., 2012, entre otros) han reconocido que el desconocimiento de aspectos técnicos referentes a la fundamentación teórica y estadística de las pruebas es uno de los principales determinantes del uso indiscriminado de las mismas y las inadecuadas prácticas profesionales, lo cual afecta la validez de las interpretaciones de los resultados y las decisiones que ellas sustentan.

Una de las estrategias usada para hacerle frente a esta situación es la elaboración y divulgación de guías o recomendaciones técnicas como *Standards for Psychological Testing* de la American Educational Research Association, la American Psychological Association y el National Council on Measurement in Education (1999, 2014), las *Guidelines* de la *Internacional Test Commission* (2001, 2014, 2016, 2018, 2019) o la norma ISO 10667 (Organización Internacional de Normalización, 2011); esta última, desarrollada por iniciativa de los representantes alemanes en la Organización Internacional de Estandarización, para el ámbito laboral.

Otra estrategia implementada por algunas asociaciones gremiales de psicología para promover buenas prácticas en el uso de pruebas es la evaluación de su calidad técnica y la amplia divulgación de sus resultados, de manera que los profesionales dispongan de información precisa para seleccionar las pruebas (Prieto & Muñiz, 2000; Evers et al., 2012).

Generalmente los sistemas de evaluación de la calidad de las pruebas se desarrollan y administran por parte de mesas de trabajo, comités, institutos anexos o divisiones de las organizaciones que regulan el ejercicio profesional, y requieren articular decisiones políticas, requerimientos técnicos y consideraciones éticas con recursos económicos, humanos e institucionales. Por tanto, el desarrollo de un sistema de evaluación conlleva discusiones, acuerdos y esfuerzos gremiales que se concretan en una infraestructura que permita la gestión del proceso por parte de organizaciones con la autoridad, el reconocimiento y la capacidad para adelantarlo y para divulgar los resultados de las evaluaciones.

El aspecto central en cualquier sistema de evaluación de las pruebas es la adopción de un modelo que defina los criterios técnicos para valorar la calidad de las mismas. Directrices como los estándares antes mencionados de la *American Educational Research Association* (AERA), la *American Psychological Association* (APA) y el *National Council on Measurement in Education* (NCME) o las guía de la *International Test Commission* (ITC) proponen lineamientos útiles para este propósito, pero deben ser traducidos en términos de indicadores de calidad adaptados a un contexto particular, generalmente nacional.

Aunque el primer comité de evaluaciones nació en el seno de la APA en 1895 (Evers, 2012); solo a partir de mediados de siglo XX se han desarrollado de algunos sistemas de evaluación de pruebas. El primero proceso de revisión de pruebas surgió en Estados Unidos gracias a la iniciativa del entonces estudiante de maestría de la Universidad de Columbia, Oscar Krisen Buros, quien convocó a un grupo profesional para revisar algunas pruebas, estas evaluaciones se publicaron en 1938 bajo el título *Mental Meassurement Yearbook* (Buros, 1938), dando origen a la publicación con mayor trayectoria de esta naturaleza. Hoy se publica de manera periódica cada tres años por *The Buros Center for Testing*, del Departamento de Educación de la Universidad de Nebraska, cuenta con 20 ediciones publicadas en inglés y más recientemente inició una serie sobre pruebas publicadas en español (Schlueter, Carlson,

Geisinger & Murphy, 2013; Schlueter, Anderson, Carlson & Geisinger, 2018).

En sus orígenes el sistema Buros se basaba en una metodología de foro con participación de diferentes expertos que valoraban los aspectos de calidad de las pruebas. Hoy se adelanta un procedimiento sistemático de 11 pasos que van desde la identificación de las pruebas a revisar hasta la revisión y aprobación de los conceptos técnicos por parte de los editores y la publicación de resultados. En la evaluación participan expertos de diferentes nacionalidades, quienes reciben guías que orientan el desarrollo de la revisión, tomando en consideración los estándares ya citados de la APA, la AERA y el NMEC. El concepto técnico sobre la prueba evaluada está compuesto por cinco apartes: descripción, resumen del desarrollo de la prueba, fundamentación técnica, comentarios sobre fortalezas y debilidades y conclusiones de la revisión (Carlson & Geisinger, 2012, Buros Center of testing, 2014).

A lo largo del siglo pasado y lo corrido del actual se han desarrollado algunos sistemas nacionales en diferentes países; una revisión de algunos de ellos se encuentra en León (2017). Aunque todos tienen elementos similares, cada uno adoptó un modelo de calidad que se concreta en un instrumento o protocolo para el desarrollo de la evaluación. La Tabla 1 muestra las principales categorías de contenidos de algunos de ellos, aunque en la actualidad varios países han adoptado el instrumento de la *European Federation of Psycchologists' Association* (EFPA).

En primer lugar, el sistema alemán cubre una serie de pasos que van desde selección de las pruebas hasta la publicación de la valoración, incluyendo una revisión inicial para verificar que la prueba cumple con unos criterios básicos para ser evaluada. Una vez superada esta fase, se adelanta el proceso de evaluación utilizando un modelo basado en la norma alemana DIN 33430 (ver https://www.din33430portal.de/din33430/portal) que formula criterios de calidad para los procesos de evaluación (DIN, 2005 citado en Kersting & Hornke, 2006). Este modelo organiza los criterios de evaluación de las pruebas en ocho categorías (Hagemeister, Kersting, & Stemmler, 2012), como se muestra en la Tabla 1.

El modelo español, inicialmente propuesto en el 2000 (Prieto & Muñiz, 2000), es un formulario muy exhaustivo compuesto por tres partes: Descripción de la prueba, valoración de las características de la prueba y valoración final. La primera parte contiene 31 ítems con información sobre nombre de la prueba, autor, constructo medido, ámbito de aplicación, etcétera. La segunda parte incluye indicadores sobre la calidad de los materiales y los ítems, la fundamentación teórica, los procesos de adaptación o traducción, la confiabilidad, validez y las normas de calificación. Finalmente, se incluye una valoración general que recoge los principales comentarios e impresiones de los evaluadores.

Tabla 1
Resumen de los principales contenidos de diferentes modelos de evaluación de pruebas

País/Entidad	Contenidos generales
Estados Unidos/ Instituto Buros (The Buros Center for Testing)	Descripción de la prueba
	Resumen del desarrollo de la prueba
	Fundamentación técnica: Confiabilidad, validez y estandarización
	Comentario sobre fortalezas y debilidades
	Conclusión
Alemania/ Comité de Evaluación de pruebas (Board of Assessment and Testing)	Información general sobre la prueba
	Fundamentos teóricos
	Objetividad
	Normas

País/Entidad	Contenidos generales
	Confiabilidad
	Validez
	Otros criterios de calidad Susceptibilidad a errores o falsificación, etc)
	Evaluación final
España/ Colegio Oficial de Psicólogos–COP	Descripción general del test: Nombre de la prueba, autor, editor, clasificación, número y tipo de ítems, variables medidas, población, sistema de puntuación, precio de la prueba, bibliografía
	Valoración de las características del test: Calidad de los materiales, las instrucciones, los ítems, análisis de ítems, fundamentación teórica, validez, confiabilidad, normas de calificación
	Valoración global del test
Holanda / Dutch Committee on Testing (COTAN)	Bases teóricas de la prueba
	La calidad de los materiales de la prueba
	Compresibilidad del manual
	Normas
	Confiabilidad
	Validez de constructo
	Validez de criterio
European Federation of Psychologists´ Association- Board of Assessment	Descripción del instrumento: Descripción general, clasificación, puntuación, reportes generados por computados, costos y condiciones de distribución
	Evaluación del instrumento: Calidad de la información, calidad de los materiales de prueba, normas, fiabilidad, validez, calidad de los informes generados por computador, evaluación final.
	Bibliografía
Brasil/ Consejo Federal de Psicólogos	Base teórica del instrumento
	Validez
	Sustento psicométrico
	Estandarización aplicación
	Manual técnico

Uno de los sistemas de referencia en Europa es el holandés, administrado por el Comité Holandés de Pruebas (*Dutch Committee on Testing*, COTAN) de la Asociación Holandesa de psicología (*Dutch Psychological Association*, NIP). Su modelo de evaluación ha tenido cinco actualizaciones desde la publicación de 2006. La más reciente organiza los criterios de evaluación en siete categorías que incluyen, además, de los criterios de confiabilidad, validez y normas, la calidad de los materiales de la prueba y la compresibilidad del manual (Evers, 1996; Evers, Sijtsma, Lucassen & Meijer, 2010).

El modelo de la Sociedad Británica de Psicología (*British Psychological Society*, BPS) ha puesto su énfasis en la cualificación de los usuarios de pruebas; sin embargo, en ese proceso de desarrollaron estrategias para brindar a los profesionales, información de calidad sobre las pruebas (Bartram, 1996, 1998). El primer modelo de 1990 consistía en una publicación de la revisión de varias pruebas en la que se evaluaba la calidad de los materiales, la confiabilidad, la validez y las normas de calificación; sin embargo, a partir de la publicación del modelo de la EFPA, la Sociedad Británica decidió adoptar este modelo (Lindley & Bartram, 2012).

Con base en la revisión de los modelos de evaluación de varios países europeos, la EFPA desarrolló un modelo de evaluación de pruebas (Bartram, 2002) que ha tenido varias revisiones durante las dos décadas de este siglo. La última actualización publicada en 2013 (Evers, et al, 2013) está compuesta por tres secciones, una descriptiva,

una valorativa y las referencias bibliográficas utilizadas para la evaluación, como aparece en la Tabla 1. En la primera parte se describe detalladamente la prueba desde su identificación hasta las condiciones y costos para acceder a ella. Incluye la información sobre sus contenidos, usos y población a la que se dirige; composición, número y tipo de ítems; forma de aplicación, calificación y transformación de puntajes; y formas y características de los reportes de resultados. En la segunda parte se incluyen indicadores que son calificados en una escala liket de 0 a 4 y evalúan la calidad de la información presentada, de los materiales de prueba y de los informes generados por computadora, así como los aspectos técnicos involucrados en la obtención de las normas de calificación y de las evidencias de fiabilidad y validez.

Actualmente el modelo ha sido acogido por varias asociaciones nacionales de psicología de países europeos, entre ellas las de los cuatro modelos presentados aquí: España, Alemania, Holanda y Reino Unido.

El único sistema suramericano es el del Consejo Federal de Psicología de Brasil (*Sistema de Avaliação de Testes Psicológicos*, SATEPSI). A diferencia de los sistemas europeos, SATEPSI tiene carácter restrictivo y no solamente informativo, ya que solo las pruebas que han pasado el proceso de evaluación técnica y han obtenido resultados satisfactorios pueden ser legalmente usadas en el país (*Conselho Federal De Psicologia*, 2003). El modelo de revisión de calidad sigue un protocolo de tres partes que examinan seis aspectos de calidad a partir de indicadores bien definidos. Según Porto Noronha (2012), este protocolo se basa en el modelo propuesto por Prieto y Muñiz (2000) en España y considera las directrices técnicas de la ITC (2001) y de la APA (1999). Una particularidad de este modelo es que el protocolo puede ser utilizado con pruebas proyectivas.

Aunque los modelos revisados difieren en su nivel de estructuración, todos evalúan contenidos similares sobre todo en lo referente a la justificación teórica y técnica de la prueba. Este trabajo tuvo como objetivo desarrollar y pilotear un instrumento para evaluar las calidades técnicas de las pruebas publicadas en español y en particular, las usadas en Colombia.

Método

El trabajo adoptó la estrategia metodológica y los perfiles de participantes propuestos por Herrera (2009) como plan de acción de la División de Medición, Evaluación y Estadística Aplicada del Colegio Colombiano de Psicólogos (COLPSIC). El procedimiento cubrió tres fases: Desarrollo de la primera versión del instrumento, evaluación por pares de la versión del instrumento y pilotaje y conformación de la versión definitiva del instrumento.

Desarrollo de la versión preliminar

Para la definición de los criterios e indicadores de calidad se buscó conformar un grupo de profesionales de las diferentes regiones del país. El equipo de trabajo se conformó mediante invitación directa a los profesionales que cumplieran con el perfil previamente establecido, se les convocó a actividades de capacitación y, una vez acordada la participación en el proyecto, se solicitó el respectivo aval institucional. 31 profesionales participaron en las sesiones de capacitación y, finalmente, se conformó un equipo de 18 psicólogos profesionales, además de las autoras, con experiencia en psicometría o evaluación psicológica, docentes e investigadores vinculados a 12 universidades de 6 regiones del país, o al Instituto Colombiano para la Evaluación de la Educación – ICFES.

Se desarrollaron dos talleres de capacitación con profesores internacionales, una sobre el modelo de construcción de pruebas con modelo análisis IRT con apoyo del mapa conceptual (Curso taller: *Introduction irt: the rasch model for measurement*) y otro específico sobre la experiencia brasileña en evaluación de calidad de las pruebas (Curso *Consideraciones sobre la calidad de las pruebas. Desarrollo de los procesos de evaluación de pruebas psicológicas*).

La primera versión del instrumento se desarrolló en trece sesiones presenciales de trabajo. En las primeras tres sesiones se presentó un panorama

general de los instrumentos utilizados en varios países, se definió la estructura general del instrumento colombiano y se conformaron cuatro mesas de trabajo para definir los criterios e indicadores de los cuatro últimos. A partir de la cuarta sesión de trabajo se presentaron y discutieron los avances de cada mesa de trabajo sobre los criterios de calidad, los indicadores, descriptores de los niveles de cumplimiento y la escala de calificación.

Revisión por pares

La primera versión del instrumento se sometió a revisión por parte de un grupo de cinco jueces, tres profesionales de amplio reconocimiento nacional y trayectoria en medición y evaluación, y dos internacionales, líderes de los procesos de evaluación de calidad de las pruebas, en sus respectivos países. A los jueces se les pidió que revisaran el contenido del instrumento, la pertinencia de los indicadores y la claridad y graduación de los niveles de cumplimiento de los descriptores de cada indicador. Después de finalizada la revisión, se adelantó una sesión conjunta de los jueces con el equipo de trabajo para discutir y aclarar las observaciones y sugerencias de los jueces.

Una vez ajustada esta primera versión, se sometió a una nueva revisión por parte de un panel de expertos con participación de once profesionales expertos en psicometría o evaluación, vinculados a diferentes instituciones de diez ciudades del país, quienes conformaron el equipo de trabajo para el diseño del *Sistema de Evaluación de la Calidad de las Pruebas Psicológicas*, en desarrollo actualmente. Los resultados de estas revisiones permitieron realizar ajustes tanto al instrumento y al instructivo para su uso.

Pilotaje y conformación de la versión final del instrumento

La primera versión del instrumento fue sometida también a una prueba piloto mediante la aplicación en la valoración de seis pruebas seleccionadas con base en los resultados del estudio para identificar los instrumentos más usados por los psicólogos en el país (Herrera & León, 2015). Se seleccionaron seis pruebas de las más usadas y se conformó un grupo de evaluadores diferente al que había participado en el desarrollo del instrumento, de manera que cada prueba fuera evaluada por un experto en el área de aplicación de la misma y un experto en psicometría. Los evaluadores participaron en un taller de capacitación sobre uso del instrumento y se les pidió que anotaran todas las observaciones o dudas que surgieran durante el proceso. Estas observaciones permitieron hacer nuevos ajustes tanto al instrumento como al instructivo que lo acompaña.

Resultados

Versión preliminar y pilotaje del instrumento

El primer resultado de este trabajo estuvo constituido por la versión preliminar del instrumento para la evaluación de la calidad de las pruebas y un instructivo con los fundamentos conceptuales y técnicos necesarios para evaluar los indicadores, así como las instrucciones de uso del instrumento (Herrera, et al., 2018a). La primera versión del instrumento quedó compuesta por los cinco (5) apartados que aparecen en la Tabla 2. El primero de ellos es una ficha técnica que describe en detalle la prueba evaluada, y los cuatro restantes incluían 47 indicadores de calidad que valoran el nivel de cumplimiento de cada uno de ellos con base en la información reportada en el manual y material de prueba.

Uno de los aspectos característicos de este instrumento es la inclusión de indicadores opcionales y obligatorios. Los obligatorios recogen información que se considera indispensable para garantizar un nivel de calidad aceptable en una prueba, cada uno de ellos se valora en una escala de cuatro niveles de cumplimiento: no cumple, cumplimiento mínimo, aceptable y adecuado. Se esperaría que una prueba debiera cumplir por lo menos en el nivel mínimo todos los indicadores obligatorios. Los indicadores opcionales hacen referencia a condiciones deseables en una prueba pero que en la actualidad no son muy frecuentes

en los desarrollos de estas y por tanto no suelen estar incluidos en los manuales. Estos solo tienen dos niveles de cumplimiento: adecuado y sobresaliente y pueden ser calificados con puntuaciones entre 4,0 y 5,0, como se muestra en la Tabla 3. La versión preliminar del instrumento contenía 32 indicadores obligatorios y 15 opcionales.

Los resultados de las revisiones por pares y del pilotaje de la versión preliminar evidenció la necesidad de hacer ajustes importantes sobre todo en tres aspectos. En primer lugar, se aumentó el número de ítems de la ficha técnica que pasó de 10 a 14, con el fin de recoger información detallada sobre la composición de la prueba en términos de dimensiones del constructo medido, número y tipo de ítems en cada una de ellas y procedimientos para las adaptaciones o traducciones en los casos en que aplica. En segundo lugar, se aumentó de 32 a 36 el número de indicadores obligatorios, se disminuyó de 15 a 12 el número de indicadores opcionales y se precisó la redacción de los criterios de calificación en 28 de ellos. Los indicadores nuevos hacen referencia a la actualización de los referentes conceptuales y de los estudios que soportan empíricamente la confiabilidad y la validez de la prueba, así como a las características de la muestra normativa y la actualización de las tablas de calificación; por el contrario, se disminuyó el número de indicadores relacionados con el análisis de los ítems por considerar que algunos de ellos parecían repetidos o no resultaban aplicables para la mayoría de las pruebas. Finalmente, una de las modificaciones más notorias fue la ampliación de las explicaciones técnicas y conceptuales del instructivo para uso del instrumento con el fin de detallar los criterios de calificación y reducir las diferencias que se observaron entre los profesionales que participaron en el pilotaje en cuanto a la interpretación de los mismos (León, 2017).

Descripción del instrumento final

La estructura general del instrumento final se muestra en la Tabla 2 y los indicadores y criterios de calificación aparecen en el anexo[1]. El instrumento definitivo incluye 48 indicadores de calidad, de los cuales 36 son obligatorios y 12 opcionales, como se muestra en la Tabla 2.

1 Puede consultar el anexo en: https://revistas.unal.edu.co/index.php/psicologia/article/view/93018

Tabla 2
Estructura general del instrumento definitivo de evaluación de la calidad técnica de las pruebas psicológicas

Apartado	Contenidos	Número de indicadores	
		Obligatorios	Opcionales
1. Descripción	Identificación de la prueba Características de la prueba Descripción general de la prueba		
2. Referentes conceptuales		6	2
3. Confiabilidad		7	2
4. Validez	Evidencia de validez del contenido de la prueba	4	2
	Evidencia basada en la estructura interna	3	
	Evidencia basada en la relación con otras variables	5	2
	Evidencia con base en el análisis de ítems*	3	1
5. Calificación y estandarización		8	3

Fuente: Adaptado de (Herrera et al., 2018b)

Tabla 3
Categorías y escala para evaluar la información de los indicadores

Indicadores	Categoría	Escala numérica
Obligatorios	No cumple (NC)	
	Cumplimiento mínimo	Entre 3.0 y 3.4
	Cumplimiento aceptable	Entre 3.5 y 3,9
	Cumplimiento adecuado	Entre 4.0 y 4,9
Opcionales	Adecuado	Entre 4,0 y 4,4
	Sobresaliente	Entre 4,5 y 5,0

En lo referente a los contenidos específicos evaluados por el instrumento, en el anexo[2] aparecen los indicadores y los descriptores de los niveles de cumplimiento para cada categoría en cada uno de ellos. El segundo apartado se ocupa de valorar la información que el manual brinda al usuario de la prueba sobre los referentes conceptuales que la sustentan. Los indicadores obligatorios valoran la información sobre el constructo que evalúa la prueba, su definición conceptual y operacional, la definición de las dimensiones que lo componen y su justificación de uso para la población objeto. Los dos indicadores opcionales se refieren a información que permite identificar el constructo como parte de un desarrollo estructurado científicamente, donde se entiende la evolución del concepto desde una perspectiva amplia.

El tercer apartado evalúa la información sobre los estudios de confiabilidad. Los indicadores obligatorios valoran la información sobre las fuentes de error aleatorio y el diseño y los resultados de los estudios empíricos que sustentan la precisión de la medida. Los indicadores opcionales hacen referencia a la justificación y la coherencia en la selección del diseño del estudio o estudio de confiabilidad y del estadístico o estadísticos para estimarla.

Para la definición de los indicadores de validez se adaptó la categorización de evidencias de validez propuesta por los últimos estándares de AERA, APA, NCME (2014). Este apartado incluye entonces cuatro subapartados según el tipo de evidencia de validez: relacionada con el contenido de la prueba, basada en la relación con otras variables, relacionada con estructura interna de la prueba y evidencia sobre el análisis de ítems. Los indicadores del último subapartado solo se califican si la prueba evaluada se compone de ítems o tareas.

Finalmente, el apartado 5 evalúa aspectos relacionados con el estudio o estudios que soportan la construcción de las tablas de calificación y la precisión de la información para calificar, trasformar los puntajes y reportar los resultados. Los indicadores obligatorios valoran la claridad en la definición de la población objeto, así como la calidad de la muestra normativa y los procedimientos para transformar e interpretar los puntajes obtenidos.

Conclusiones

El instrumento desarrollado es producto del trabajo de un amplio número de profesionales de todo el país, expertos en medición y evaluación o en alguna de las áreas aplicadas de la psicología, y por tanto recoge los acuerdos, productos de muchas discusiones, sobre las condiciones técnicas mínimas que un manual de prueba debería reportar para considerar que la misma cuenta con una calidad técnica adecuada. Además, dado que el desarrollo del instrumento se basó en una revisión cuidadosa de los modelos existentes a nivel internacional, recoge y capitaliza dichas experiencias.

A pesar de que uno de los propósitos del equipo de trabajo que participó en la primera fase el estudio fue desarrollar una herramienta de fácil uso para la mayoría de los profesionales de la psicología, los resultados de la aplicación piloto mostraron que su adecuado manejo implica una lectura detallada del manual técnico y una adecuada formación en temas relacionados no solo con el uso de las pruebas, sino con algunos conceptos y metodologías propias de la psicometría. A futuro se espera que el grupo de evaluadores esté conformado por profesionales que reciban una capacitación específica en el uso del instrumento y los elementos conceptuales y metodológicos que sustentan los criterios de calificación.

2 Puede consultar el anexo en: https://revistas.unal.edu.co/index.php/psicologia/article/view/93018

Además, aunque el objetivo principal del instrumento es proponer una herramienta que permita proveer información relevante para que los profesionales dispongan de elementos técnicos que apoyen su decisión sobre la elección de las pruebas en su ejercicio profesional, este puede convertirse en una guía que ayude a orientar la formación en psicometría por parte de los programas de psicología. Si bien el instructivo que acompaña al instrumento (Herrera et al., 2018b) no pretende ser un texto de psicometría o de evaluación, ni ser exhaustivo en el tratamiento de los temas, recoge elementos técnicos que cualquier usuario de pruebas debería conocer.

En lo referente a los resultados de las evaluaciones piloto de las seis pruebas más usadas, aunque el objetivo del ejercicio era pilotear la claridad y la comprensión del instrumento, vale la pena destacar dos resultados: Una amplia disparidad de los resultados en los cuatro apartes valorativos del instrumento, y la ausencia de tablas de calificación levantadas con muestras normativas colombianas en los manuales técnicos evaluados.

Finalmente, vale la pena destacar el avance que representa para el país contar con una primera versión de un Instrumento para Evaluar las Calidades Técnicas de las Pruebas adecuado para obtener información sobre los instrumentos que usan los profesionales en ejercicio. Aunque seguramente serán necesarias varias actualizaciones antes de disponer de un modelo maduro, este instrumento, producto de un trabajo colectivo muy arduo y que recoge muchas discusiones y mirada diferentes de los aspectos deseables en la principal herramienta de trabajo de los psicólogos, constituye una herramienta muy útil que seguramente impactará el ejercicio profesional en lo relacionado con el uso de pruebas.

Agradecimientos

"Este proyecto fue cofinanciado por el Colegio Colombiano de Psicólogos y por la Universidad Nacional de Colombia".

"Las autoras agradecen a las instituciones que se vincularon al proyecto mediante la delegación de alguno de sus profesionales, a todos los psicólogos que participaron tanto en el desarrollo del instrumento como en su revisión y pilotaje, y a las casas editoriales *PSEA S:A:S Psicólogos Especialistas asociados y Manual Moderno Colombia* por facilitar los materiales de prueba necesarios para el pilotaje del instrumento".

Referencias

American Psychological Association, American Educational Research Association & National Council on Measurement in Education [APA] (1999). *Standards for educational and psychological test y manuals*. Washington: American Psychological Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education [APA] (2014). *Standards for Psychological Testing*. Washington: American Educational Research Association.

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. European *Journal of Psychological Assessment*, *12*(1), 62–71. https://doi.org/10.1027/1015-5759.12.1.62

Bartram, D. (1998). The need for international guidelines on standards for test use: A review of European and international initiatives. *European Psychologist*, *3*(2), 155–163. DOI:10.1027/1016-9040.3.2.155

Bartram, D. (2002). Review model for the description and evaluation of psychological tests. Brussels: European Federation of Psychologists' Associations (EFPA).

Buros, O. K. (1938). *The 1938 Mental Measurements Yearbook*. New Brunswick: Rutgers University Press.

Buros Center of Testing (2014). *History of The Buros Center for Testing*. Recuperado de http://buros.org/history.

Carlson, J. F., & Geisinger, K. F. (2012). Test reviewing at the Buros Center for Testing. *International Journal of Testing*, *12*(2), 122–135. DOI: 10.1080/15305058.2012.661003.

Conselho Federal de Psicologia (23 de março de 2003), Resolução Nº 002, Define e regulamenta o uso, a elaboração e a comercialização de testes psicológicos e revoga a Resolução CFP nº 025/2001.

Evers, A. (1996). Regulations concerning test qualifications and test use in the Netherlands. *European Journal of Psychological Assessment*, *12*(2), 153–159. DOI: https://doi.org/10.1027/1015-5759.12.2.153

Evers, A. (2012). The internationalization of test reviewing: Trends, differences, and results. *International Journal of Testing*. *12*(2), 136–156. DOI: 10.1080/15305058.2012.658932

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and result. *International Journal of Testing*, *10*(4), 295–317. DOI: 10.1080/15305058.2010.518325.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., Frans, Ö., Gintiliené, G., Hagemeister, C., Halama, P., Iliescu, D., Jaworowska, A., Jiménez, P., Manthouli, M., Matesic, K., Schittekatte, M., Sümer, H. C., & Urbánek, T. (2012). Testing practices in the 21st century: Developments and European psychologists' opinions. *European Psychologist*, *17*(4), 300–319. DOI:10.1027/1016-9040/a000102

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, *25*(3) 283–291. DOI: 10.7334/psicothema2013.97

Hagemeister, C., Kersting, M. & Stemmler, G. (2012). Test reviewing in Germany. *International Journal of Testing*, *12*(2), 185–194. DOI:10.1080/15305058.2012.657922

Herrera, A. N. (2009). *Evaluación de la Calidad Técnica de las pruebas más frecuentemente Usadas por los psicólogos profesionales en Colombia. Propuesta plan de acción para la División de Medición Evaluación y Estadística Aplicada*. Bogotá: Documento inédito.

Herrera, A. N., & León, F. A. (Septiembre de 2015). *Retos en el uso de pruebas en Colombia*. Trabajo presentado en el III Congreso de Psicología de Colpsic–Ascofapsi. Bogotá, Colombia.

Herrera, A. N., León, F. A., Arias, E. M., Avendaño, B. M., Camargo, S. L., Cárdenas, L., Cervantes, V. H., Cortes, O. F., Cuevas, M. L., Charry, C., Duarte, L. H., Escobar, J., Espinosa, J. C., Meneses, A. L., Pineda, C. A., Rodríguez, O. R., Salas, R., Solís, G., Suárez, A. E., & Vera, A. (2018a). *Evaluación de la calidad técnica de las pruebas psicológicas usadas en Colombia. Instrumento para la valoración de calidades técnicas.* Bogotá: COLPSIC; Universidad Nacional de Colombia.

Herrera, A. N., León, F. A., Arias, E. M., Avendaño, B. M., Camargo, S. L., Cárdenas, L., Cervantes, V. H., Cortes, O. F., Cuevas, M. L., Charry, C., Duarte, L. H., Escobar, J., Espinosa, J. C., Meneses, A. L., Pineda, C. A., Rodríguez, O. R., Salas, R., Solís, G., Suárez, A. E., & Vera, A. (2018b). *Evaluación de la calidad técnica de las pruebas psicológicas usadas en Colombia. Instrumento para la valoración de calidades técnicas–Instructivo.* Bogotá: COLPSIC; Universidad Nacional de Colombia

International Test Commission [ITC] (2001). International guidelines for test use. *International Journal of Testing, 1*(2), 93–114. DOI: 10.1207/S15327574IJT0102_1

International Test Commission [ITC] (2014) ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *14*(3), 195–217. DOI: 10.1080/15305058.2014.918040

International Test Commission [ITC] (2016). The international test commission guidelines on the security of tests, examinations, and other assessments. *International Journal of Testing*, *16*(3), 181–204. DOI: 10.1080/15305058.2015.1111221

International Test Commission [ITC] (2018). ITC guidelines for translating and adapting tests. *International Journal of Testing*, *18*(2), 101–134. DOI: 10.1080/15305058.2017.1398166

International Test Commission [ITC] (2019). ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations. *International Journal of Testing*, *19*(4), 301–336. DOI: 10.1080/15305058.2019.1631024

Kersting, M., & Hornke, L. F. (2006). Improving the quality for proficiency assessment: The German Standardization Approach. *Psychology Science*, *48*(1), 85–98. Recuperado de https://psycnet.apa.org/record/2006-07428-006

León, F. A. (2017). *Evaluación piloto de la calidad técnica de las seis pruebas más usadas en Colombia* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá, Colombia.

Lindley, P. A., & Bartram, D. (2012). Use of the EFPA test review model by the UK and issues relating to the internationalization of test standards. *International Journal of Testing, 12*(2) 108–121. DOI: 10.1080/15305058.2011.652267

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J. R., & Zaal, J. N. (2001). Testing practices in European countries. *European Journal of Psychological Assessment*, *17*(3), 201–211. DOI:10.1027/1015-5759.17.3.201

Organización Internacional de Normalización (2011). Proceso para la evaluación de personas en las Organizaciones. Norma 10667. Recuperado de https://www.normas-iso.com/iso-10667-una-herramienta-internacional-en-rrhh-2/

Porto Noronha, A. P. (2012). *Sistema de Avaliacao dos Test Psicológicos de Conselho Federal de Psicologia*. Conferencia en el ciclo de capacitación para el diseño de un instrumento para valorar las calidades técnicas de las pruebas. Universidad Nacional de Colombia, Bogotá, Colombia

Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo*, *77*, 65–72. Recuperado de https://www.redalyc.org/pdf/778/77807709.pdf.

Schlueter, J. E., Carlson, J. F., Geisinger, K. F., & Murphy, L. L. (2013). *Pruebas publicadas en español. An index of Spanish tests in print*. Lincoln, Nebraska: The Buros Center for Testing. Recuperado de http://www.buros.org/pdfs/PPEsamplepages.pdf

Schlueter, J. E., Anderson, N. A., Carlson, J. F., & Geisinger, K. F. (Eds.) (2018). *Pruebas publicadas en español II. An index of Spanish tests in print*. Lincoln, Nebraska: The Buros Center for Testing.