

EVALUACIÓN PILOTO DE LA CALIDAD TÉCNICA DE SEIS DE LAS PRUEBAS
PSICOLÓGICAS MÁS USADAS EN COLOMBIA

FLOR ÁNGELA LEÓN GRISALES



UNIVERSIDAD
NACIONAL
DE COLOMBIA

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS HUMANAS
DEPARTAMENTO DE PSICOLOGÍA
BOGOTÁ D. 2017

EVALUACIÓN PILOTO DE LA CALIDAD TÉCNICA DE SEIS DE LAS PRUEBAS
PSICOLÓGICAS MÁS USADAS EN COLOMBIA

FLOR ÁNGELA LEÓN GRISALES

Trabajo de grado para optar al título de magister en Psicología
En la línea de Métodos e Instrumentos en la Investigación del Comportamiento

Dirigido por:

AURA NIDIA HERRERA ROJAS



UNIVERSIDAD
NACIONAL
DE COLOMBIA

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS HUMANAS
DEPARTAMENTO DE PSICOLOGÍA
BOGOTÁ D.C 2017

A mis padres Yamile y Carlos

que me dieron la vida

A mis amores Daniela y Víctor Manuel

que me han dado el amor para vivir la vida

A mi amada Universidad Nacional de Colombia

Donde aprendí una profesión y aprendí a servir para darle un

sentido a la vida.

Agradecimientos

En estos breves párrafos extiendo mis sinceros agradecimientos a todas las personas que participaron en las actividades del proyecto, no sin antes mencionar que este trabajo es un gran logro en mi vida profesional, porque su desarrollo me llevó a reflexionar sobre la importancia de las pruebas psicológicas en la vida de las personas y me abrió los ojos a la realidad sobre las prácticas profesionales en relación con la evaluación en psicología, por ello más que este documento quedan valiosos aprendizajes y una mirada crítica sobre el uso de las pruebas en Colombia.

El primer lugar quiero agradecer a la Doctora Aura Nidia Herrera, directora del proyecto y una profesora ejemplar que me dio la oportunidad de trabajar bajo su dirección.

En segundo lugar a la Dra. Claudia Sanín, presidente del Colegio Colombiano de Psicólogos, que con su aval apoyó las acciones y actividades del proyecto, abriendo un espacio muy importante para que se logaran los objetivos de este trabajo.

Y de manera muy especial a todos los colegas que participaron en la construcción del instrumento y a mis compañeros del laboratorio que me regalaron sus comentarios para alimentar este documento.

Tabla de Contenido

Tabla de Contenido	1
Índice de tablas	1
Introducción	2
Revisión bibliográfica.....	8
Evaluación de calidad de las pruebas	8
Antecedentes	9
Modelos de directrices o estándares.....	16
Modelos de evaluación de las pruebas	18
Algunas consideraciones conceptuales y metodológicas sobre los criterios técnicos para evaluar la calidad de las pruebas.....	30
Fundamentación conceptual.....	32
Propiedades métricas de la prueba.....	33
Estandarización de la prueba.....	46
Método	49
Fase 1: Identificación de las pruebas más usadas en Colombia	49
Población y participantes.....	49
Instrumento.....	49
Procedimiento.....	50
Fase 2 Desarrollo del instrumento de evaluación de calidad técnica de las pruebas.....	51
Participantes.....	51
Procedimiento.....	51
Fase 3 Evaluación piloto de seis de las pruebas más usadas por psicólogos colombianos ..	55
Participantes	55
Instrumentos.....	55
Procedimiento.....	56
Resultados	58
Identificación de las pruebas más usadas por los psicólogos colombianos.....	58
Identificación de las pruebas más usadas	59
Desarrollo de instrumento para evaluar la calidad técnica de las pruebas.....	61
Evaluación de calidad de seis de las pruebas más usadas por psicólogos colombianos.....	66

Evaluación de los referentes conceptuales	67
Evaluación de los estudios de confiabilidad.....	68
Evaluación de los indicadores de validez.....	69
Evaluación de la información sobre estandarización	72
Discusión y conclusiones	74
Referencias	79
ANEXOS	86

Índice de tablas

TABLA 1 CONTENIDOS DE LOS ESTÁNDARES DE LA APA, AERA Y NMCE (2014)	17
TABLA 2 MODELO DE EVALUACIÓN DEL CENTRO BUROS.	20
TABLA 3 CONTENIDOS EVALUADOS EN EL MODELO ALEMÁN.....	22
TABLA 4 MODELO DE LA EVALUACIÓN DE PRUEBAS EN HOLANDA.....	23
TABLA 5 ASPECTOS EVALUADOS EN EL INSTRUMENTO PARA REVISIÓN DE CALIDAD TÉCNICA DE LAS PRUEBAS DE LA EFPA	26
TABLA 6 MODELO BRASILEÑO DE EVALUACIÓN	28
TABLA 7 NOCIONES DE CONFIABILIDAD Y MÉTODOS DE ANÁLISIS EN LOS TRES MODELOS DE ANÁLISIS.	39
TABLA 8 IDENTIFICACIÓN DE TIPOS DE EVIDENCIA Y OBJETIVOS DE LOS ESTUDIOS DE VALIDEZ.....	42
TABLA 9 PERFILES DE LOS PARTICIPANTES EN EL PROCESO DE DESARROLLO DEL INSTRUMENTO.....	52
TABLA 10 PRUEBAS SELECCIONADAS SU EVALUACIÓN PILOTO DE CALIDAD TÉCNICA	56
TABLA 11 REPORTE FORMACIÓN DE POSGRADO DE LOS PARTICIPANTES EN LA CONSULTA DE USO DE PRUEBAS.	58
TABLA 12 LISTADO DE LAS PRUEBAS CON LOS MAYORES REPORTES DE USO	59
TABLA 13 ESTRUCTURA GENERAL DEL INSTRUMENTO DE EVALUACIÓN DE LA CALIDAD TÉCNICA DE LAS PRUEBAS PARA USO EN COLOMBIA	62
TABLA 14 INDICADORES EVALUADOS EN EL APARTADO EVALUACIÓN DE LOS REFERENTES CONCEPTUALES	63
TABLA 15 CRITERIOS DE CONFIABILIDAD PARA EVALUAR LA CALIDAD DE LA TÉCNICA DE LA PRUEBA	63
TABLA 16 INDICADORES PARA EVALUAR LA VALIDEZ	64
TABLA 17 INDICADORES DE EVALUACIÓN DE ESTANDARIZACIÓN DE LA PRUEBA	65
TABLA 18 CATEGORÍAS Y ESCALAS PARA EVALUAR LA INFORMACIÓN DISPONIBLE EN EL MANUAL.....	66
TABLA 19 RESULTADOS EVALUACIÓN DE APARTADO REFERENTES CONCEPTUALES	67
TABLA 20 PORCENTAJES DE INDICADORES DE REFERENTES CONCEPTUALES EN CADA CATEGORÍA DE CUMPLIMIENTO	67
TABLA 21 RESULTADOS EVALUACIÓN DE LOS INDICADORES DE CONFIABILIDAD DE LAS PRUEBAS EVALUADAS.	68
TABLA 22 PORCENTAJE DE INDICADORES DE CONFIABILIDAD EN CADA CATEGORÍA DE CALIFICACIÓN.....	69
TABLA 23 RESULTADOS DE LA EVALUACIÓN DE LOS INDICADORES DE VALIDEZ.....	70
TABLA 24 PORCENTAJES DE INDICADORES DE VALIDEZ EN POR CATEGORÍA.....	71
TABLA 25 RESULTADOS DE LA EVALUACIÓN DE LA ESTANDARIZACIÓN	72
TABLA 26 PORCENTAJES DE CUMPLIMIENTO DE INDICADORES DE ESTANDARIZACIÓN	73

Resumen

La calidad técnica de las pruebas es una cualidad fundamental que se relaciona con la pertinencia de su uso, en este trabajo se buscó reconocer la realidad sobre la calidad técnica de las pruebas usadas por los psicólogos en Colombia, para ello se realizó un trabajo por fases en la primera se identificó mediante encuesta las usadas, la segunda se construyó, un instrumento para evaluar la calidad técnica y se aplicó a las seis pruebas objetivas que de acuerdo al reporte de los profesionales son las más usadas.

Los resultados de la evaluación de calidad técnica y las repuestas de los psicólogos muestran que no tienen en cuenta la información técnica al momento de escoger las pruebas que utilizan y durante su uso no verifican la pertinencia respecto al objetivo de evaluación, usan pruebas sin estudios de confiabilidad y validez para población colombiana versiones que corresponden a ediciones de pruebas con más de 20 años de publicación.

A futuro es deseable que se desarrollen este tipo de evaluaciones de calidad en un proceso formal para divulgar y promover el uso de pruebas idóneas.

Palabras clave: calidad técnica, pruebas psicológicas, uso de pruebas, 16 pf-5, Wisc-IV, Inventario Millon´s de Personalidad, MMPI—2, IPV , BDI-II, estándar de calidad de pruebas.

Abstract

The technical quality of a test is a fundamental property related with its pertinence for specific uses. We study the technical quality of the tests used by psychologists in Colombia. These tests were first identified using a survey distributed through Colegio Colombiano de Psicólogos (the country's professional association to which all psychologists must affiliate). We developed an instrument for assessing the quality of tests and used it to evaluate the six tests most commonly used by professionals in the country.

From these evaluations and from responses to the survey, we found that test selection and use is not usually pertinent; that the tests have not been studied regarding their reliability and validity on Colombian population; and that the versions in use are around 20 years old,

although newer versions are available. This kind of quality assessments is thus relevant to promote and make appropriate tests known to professionals.

Keywords: technical quality, psychological tests, use of tests, 16 pf-5, Wisc-IV, Millon's Personality Inventory, MMPI-2, IPV, BDI-II, quality of evidence

Introducción

“los test son herramientas imprescindibles en la práctica psicológica en todo el mundo, y todo indica que va a seguir siéndolo en el futuro, por lo tanto no se debe ahorrar esfuerzos en mejorar su uso”

Muñiz, J., & Hambleton, R. K., (1996).

Los psicólogos, al igual que otros profesionales, requieren herramientas de medición para el desarrollo de sus actividades profesionales (Aiken, 2003; Anastasi & Urbina, 1998; Muñiz & Hambleton, 1996; Muñiz, Elosua, & Hambleton, 2013; Muñiz, Hernández, & Ponsoda, 2015; Oakland, 2004; Koene, 1997) y las pruebas o tests se han convertido en una de las más utilizadas. En consecuencia, el uso de pruebas psicológicas como instrumentos de medición es una práctica en todos los campos aplicados y profesionales de la psicología, puesto que proveen información para sustentar un diagnóstico, tomar decisiones de vinculación en el ámbito educativo o laboral, identificar riesgos o para planear estrategias de intervención, entre muchas otras aplicaciones.

Los resultados de las pruebas constituyen un elemento importante en la toma de algunas decisiones que pueden cambiar la vida de los evaluados, por ello su uso implica el cumplimiento de altos estándares de calidad de las mismas y una adecuada competencia, responsabilidad y sentido ético de quien las usa. La utilización indiscriminada de pruebas sin condiciones técnicas y el descuido de los procesos recomendados al momento de aplicarlas, afecta la validez de las decisiones que se sustentan en sus resultados. Varios autores (Bartram, 1998; Muñiz, Campillo, Fonseca, Fernández, & Peña, 2011) señalan que el desconocimiento general por parte de los profesionales de aspectos técnicos referentes a la fundamentación teórica y estadística de las pruebas es uno de los principales determinantes de la inadecuadas prácticas profesionales.

En el ámbito internacional se han adoptado preponderantemente dos estrategias para hacer frente a este reto: cualificación de los profesionales y evaluación de la calidad técnica de las

pruebas. Por una parte, los usuarios deben entrenarse para la aplicación y calificación de las pruebas, y por otra, las asociaciones gremiales promueven la adopción de normas de calidad por parte de quienes desarrollan las pruebas y divulgan ampliamente la información técnica de las mismas. Con excepción de Brasil, en Latinoamérica no se han adoptado estrategias para evaluar la calidad de las pruebas, ni se han reglamentado procesos o procedimientos para la adaptación, validación y uso de las pruebas por parte de los profesionales en ejercicio.

Mediante la Ley 1090 de 2006, en Colombia se dispone del marco normativo que reglamenta el ejercicio profesional de la psicología en el país y que se refiere explícitamente, en su capítulo VI: “uso de material psicotécnico”, a los principios fundamentales del uso de las pruebas como herramientas científicas. De acuerdo con esta normatividad, el desarrollo, selección y uso de pruebas debe seguir un alto sentido de responsabilidad, implementando criterios rigurosos que garanticen que éstas cumplan con los propósitos para los cuales se construyen. Además, otros apartados hacen alusión tácita a la importancia de la calidad de las pruebas como herramientas usadas en los procesos profesionales; así por ejemplo en el título V, se enuncia como un derecho profesional: “*contar con la tecnología e insumos adecuados para el desempeño eficiente*”; y en el artículo 46 se menciona como responsabilidad del psicólogo, la verificación de las cualidades técnicas de las pruebas que usa:

“Cuando el psicólogo construye o estandariza tests psicológicos, inventarios, listados de chequeo, u otros instrumentos técnicos, debe utilizar los procedimientos científicos debidamente comprobados. Dichos tests deben cumplir con las normas propias para la construcción de instrumentos, estandarización, validez y confiabilidad” (Ley 1090 de 2006).

Aunque existe este marco legal, en la cotidianidad las prácticas profesionales no se ajustan a estos principios (Herrera, 2013; León, 2013; 2017). Si bien existe una preocupación sobre el uso inadecuado de las pruebas, no se han puesto en marcha estrategias que alerte a los profesionales sobre la importancia de la calidad técnica de las pruebas y sus posibles limitaciones. Para la fecha de desarrollo de este trabajo no hay una reglamentación o política emitida por una entidad oficial sobre los criterios de calidad que debería cumplir una prueba para ser usada en Colombia, de manera que los psicólogos no tienen referentes o directrices

para seleccionar las pruebas que usan. Tal decisión se sustenta, en la mayoría de los casos, en los criterios brindados por la institución en la que recibió su título profesional, situación que se agrava considerando que el diagnóstico sobre el nivel de formación del pregrado en áreas metodológicas y cuantitativas, juzgadas a partir de los resultados de las pruebas SABER-PRO de psicología, muestra deficiencias importantes (Herrera, 2017).

Este diagnóstico muy general muestra la urgente necesidad llenar el vacío normativo y constituye una de las razones por las que se inició el proyecto marco desde el cual se desarrolló este trabajo: “Evaluación de las calidades técnicas de las pruebas más frecuentemente usadas en el ejercicio profesional de los psicólogos en Colombia”; formulado desde la presidencia de la desaparecida División de Evaluación y Medición del Colegio Colombiano de Psicólogos (Herrera, 2009).

Este trabajo de basa en algunas experiencias internacionales y parte del supuesto de que la evaluación de la calidad técnica de las pruebas y la amplia divulgación de sus resultados son estrategias que promueven el ejercicio responsable de la evaluación en educación y en diferentes ámbitos aplicados de la psicología. Si los profesionales cuentan con información asequible, clara y confiable sobre los instrumentos de que dispone para su trabajo, podrá seleccionarlos a partir de juicios fundamentados sobre la pertinencia, utilidad y limitaciones de los mismos como fuentes de información para sus necesidades específicas.

El propósito de este trabajo es entonces dar un paso inicial en el diseño de un procesos de evaluación de la calidad técnica de las pruebas usadas en el país, y tiene como objetivo general, brindar información sobre las calidades técnicas de algunas de las pruebas más usadas en Colombia mediante el desarrollo y uso de un instrumento que permita evaluarlas, y que se constituya un modelo para el país. Para el logro de éste, se definieron tres objetivos específicos:

1. Mostrar el panorama nacional sobre del uso de las pruebas psicológicas, identificando las más frecuentemente utilizadas por los psicólogos profesionales en Colombia y algunas prácticas de uso de las mismas.

2. Proponer un instrumento para evaluar la calidad técnica de las pruebas psicológicas para uso en Colombia.

3. Hacer un diagnóstico inicial de la calidad de las pruebas usadas en Colombia, a partir de la evaluación de las seis pruebas más usadas en el país utilizando el instrumento desarrollado.

Reconociendo que el buen uso de las pruebas es un deber ético de los psicólogos en su práctica individual, y su vigilancia es una responsabilidad de las asociaciones gremiales como el Colegio Colombiano de Psicólogos, se espera que este trabajo sea un ejemplo del tipo evaluaciones de calidad indispensables para el levantamiento y divulgación de información técnica sobre las pruebas. Se espera proponer además, un conjunto de criterios que puedan acogerse para avalar el uso de pruebas con nuestra población, y que los resultados de la evaluación piloto de la calidad técnica de seis de las pruebas más usadas por los psicólogos colombianos, sea un ejemplo de análisis sobre la pertinencia del uso de las pruebas en los procesos de evaluación.

En el desarrollo de este trabajo se implementaron algunas directrices recomendadas por la Comisión Internacional de Pruebas, (International Test Commission, ITC) y se adaptó la metodología propuesta por Herrera en 2009. Se desarrollaron tres fases; en la primera se identificaron las pruebas más usadas mediante una encuesta en línea, en la segunda se construyó el instrumento para evaluar la calidad técnica de las pruebas y en la tercera se aplicó el instrumento para realizar la revisión de seis (6) pruebas identificadas en la primera fase.

Revisión bibliográfica

Para desarrollar los elementos conceptuales que den cuenta del objetivo principal de este trabajo, este documento desarrolla dos temas relacionados con los proceso de evaluación de la calidad de las pruebas psicológicas, los antecedentes internacionales y principales modelos y procesos que se han desarrollado para la evaluación de calidad de las pruebas y los criterios o estándares técnicos que se tienen en cuenta para avalar la calidad de las mismas. Este último tema supone una revisión general de algunos conceptos y procedimientos psicométricos que sustentan el desarrollo de un sistema de estándares de calidad para las pruebas.

En términos generales los estándares de calidad son criterios definidos a partir de normas o acuerdos, que describen los lineamientos que garantizan la idoneidad de un producto o servicio y sirven para valorar sus cualidades; en otras palabras, son reglas, guías o descripciones que aseguran que los productos, procesos y servicios cumplan su propósito con interoperabilidad y condiciones de calidad (DANE, 2009; Infraestructura Colombiana de Datos Espaciales, 2015).

Los estándares de calidad para la evaluación de las pruebas psicológicas son las normas y los principios técnicos que deben considerarse en la construcción y desarrollo de estudios para sustentar teórica y metodológicamente el modelo de medida que ofrecen; al igual que en otras áreas los estándares se desarrollan y aplican para garantizar una alta calidad de los productos. En el caso particular de las pruebas psicológicas, los estándares son los criterios a considerar en el diseño, construcción, aplicación y calificación de las mismas y la interpretación de sus resultados (Buros Center for Testing, 2014).

Evaluación de calidad de las pruebas

La búsqueda de literatura sobre estándares o estrategias para evaluación de la calidad técnica de las pruebas arroja dos tipos de trabajos. Una categoría consiste en publicaciones sobre recomendaciones, directrices o estándares que deben seguirse en la construcción y desarrollo de instrumentos para garantizar que el producto sea de calidad. Una segunda categoría de trabajos hace referencia al desarrollo de instrumentos y procedimientos para

evaluar la calidad de la prueba ya construida, a partir de la información disponible en los manuales y materiales de prueba. En este documentos los trabajos de la primera categoría se denominarán modelos de directrices o estándares mientras que los segundos serán los modelos de evaluación de pruebas. En este aparte se revisan algunos antecedentes históricos del desarrollo de estrategias generales para evaluar la calidad de las pruebas y posteriormente, se presentan las principales características de estas dos categorías de trabajos.

Antecedentes

El interés a nivel mundial por la regulación de la calidad de las pruebas se convirtió en una discusión notable desde finales de la década de los 70; debido en parte a las controversias sobre la justicia en los procesos de evaluación, que llevaron la mirada a los asuntos técnicos relacionados con las pruebas como herramientas de evaluación. Esto derivó en algunos procesos de revisión por parte de asociaciones profesionales que en diferentes países establecieron comités y mesas de trabajo para desarrollar la evaluación de las pruebas; ejercicios de observación reflexiva y crítica de las condiciones métricas de las pruebas; posibles gracias a acuerdos locales para gestionar la calidad de las mismas. En este aparte se revisan los antecedentes históricos

1. Norte América: Del sistema Buros a los estándares de la Asociación Americana de Psicología

La mayoría de revisiones sobre antecedentes históricos relacionados con la evaluación de la calidad de las pruebas psicológicas, mencionan como evento pionero la aparición del primer *Committee on Mental Measurements* de la *American Psychological Association* (en adelante, APA) en 1895 (Evers, Sijtsma, Lucassen, & Meijer, 2010; Evers, 2012); sin embargo, la primera publicación que recopila información sobre la calidad de las pruebas adoptando un primer sistema de evaluación de las mismas, apareció en los años 30 del siglo pasado, gracias al entonces estudiante de la Universidad de Columbia, Oscar Krisen Buros. Inspirada en el trabajo de Kelly Truman (Buros Center for Testing, 2014), en 1938 Buros logró publicar la primera edición del anuario de pruebas mentales “*Mental Measurements Yearbooks*” (MMY), que hoy se publica de manera periódica cada 3 años, y divulga información técnica de las pruebas (Carlson & Geisinger, 2012; Buros Center for Testing, 2016).

El sistema Buros es el precursor de los procesos de monitoreo de las pruebas a nivel mundial; en sus inicios mediante una metodología de foro con participación de diferentes expertos se comentaban los aspectos de calidad de las pruebas, nutriendo las primeras evaluaciones que se publicaron en la primera edición del MMY. Oscar Buros trabajó más de 40 años como editor de este compendio de evaluaciones, tras su muerte, el Instituto Buros continuó dicha publicación que para 2016 contaba 13 ediciones con información sobre la evaluación de más de 2000 pruebas (Brandes, 2008; Carlson & Geisinger, 2012; Buros Center for Testing, 2016).

Después de la muerte de Oscar Buros su viuda, Luella Gubrud, en 1994, transfirió los archivos del MMY a la Universidad de Nebraska; donde se estableció el Centro de pruebas Buros (*Buros Center for Testing*); que en la actualidad continua la revisión de la calidad de las pruebas disponibles en el comercio, mediante un protocolo estructurado (Buros Center for Testing, 2017) y edita y publica este famoso compendio que constituye la experiencia de evaluación de pruebas con mayor trayectoria, considerada una fuente independiente, precisa, completa y fidedigna de información técnica (Brandes, 2008, Carlson & Geisinger, 2012).

El otro referente Norte Americano para la verificación de la calidad de las pruebas son las directrices conocidas como Estándares para Evaluación Psicológica y Educativa, que apareció en 1954 cuando la APA publicó la primera versión de las “*Technical Recommendations for Psychological Test and Diagnostic Techniques*”. La segunda versión titulada “*Technical Recommendations for Achievement Test*” se publicó en 1955 por la *American Educational Research Association* (en adelante, AERA) y fue preparada por un Comité de Estándares conformado por esta asociación y el *National Council on Measurement in Education* (en adelante NCME). En 1966 estas dos organizaciones se aliaron con la APA para convocar a un Comité de expertos para el desarrollo de las versiones posteriores. En consecuencia, desde su tercera edición esta publicación es preparada por un comité que representa las tres instituciones líderes en evaluación en Estados Unidos: AERA, APA) y NCME.

Esta publicación es un compendio de criterios deseables que orientan a desarrolladores, comercializadores y profesionales usuarios, en los procesos de construcción y selección de pruebas utilizadas en psicología y educación. Debido a que en su formulación participa un

amplio comité experto, la publicación recoge los principales avances metodológicos y conceptuales, a partir de los cuales actualiza las recomendaciones para el desarrollo de pruebas y goza de amplio reconocimiento entre la comunidad académica y las asociaciones gremiales en Estados Unidos. Además, a falta de publicaciones locales, esta publicación se ha convertido en el modelo de referencia para América Latina; la versión más reciente se publicó en 2014, después de 5 años de trabajo para la actualización de las directrices divulgadas en la quinta edición de 1999.

2. Europa: De evaluaciones locales al de la Federación de Asociaciones Europeas de Psicología

Según Cardinet (1995) el primer intento por establecer un sistema de estándares en Europa se desarrolló en Suiza en los años sesenta, iniciativa que fracasó en parte por la ausencia de un contexto legal para desarrollar estrategias de control que permitiera la puesta en marcha de un modelo de evaluación. Sin embargo, en 1959 la Asociación Holandesa de Psicología (*Dutch Psychological Association*, NIP) había creado una Comisión de Investigación de Pruebas con participación de delegados de las principales facultades de psicología locales; este grupo se convirtió posteriormente en el Comité Holandés de Asuntos de Pruebas (*Commissie Testaangelegenheden Nederland - COTAN*) que aún se encarga de promover la calidad de las pruebas. De acuerdo con Evers, Sijtsma, Lucassen, y Meijer (2010) la primera publicación con calificación de las calidades de las pruebas fue preparada por este comité y apareció en 1969.

Este mismo año inició la revisión de pruebas en Holanda mediante una metodología cualitativa que clasificaba las pruebas en seis categorías, desde muy buenas hasta muy pobres. Diez años más tarde, en 1979, la NIP conformó el grupo para trabajar en la mejora de las prácticas de uso de pruebas, equipo que consolidó los criterios de evaluación y la metodología que se implementó durante 13 años (Arne, Klaas, Lucassen, & Meijer, 2010). En 1982 se inició la aplicación de un protocolo con criterios técnicos de acuerdo con los Estándares de la APA vigentes para esa época; modificación que subsanó las falencias de la metodología inicial que no revisaba de manera específica aspectos de validez y confiabilidad; este protocolo valoraba

cinco aspectos: los principios de la construcción de pruebas, la calidad de la prueba, materiales e instrucciones, normas, confiabilidad y validez.

La metodología se actualizó en dos oportunidades posteriores, en 2000 y 2007; en la primera se modificaron los criterios de revisión de material de prueba y se ampliaron los criterios para revisar los aspectos de validez agrupados en dos apartados específicos: validez de criterio y validez de constructo; el protocolo quedó conformado entonces por siete (7) componentes evaluados. En la segunda actualización que se dio a conocer en 2009 se ampliaron los criterios incluyendo actualizaciones sobre avances teóricos y metodologías de análisis psicométricos; se establecieron criterios referentes al desarrollo de pruebas en general, incluyendo aspectos relacionados con el uso de formas computarizadas y la actualización de los datos de referencia y normas. De manera que la vigencia es un criterio excluyente y únicamente se da visto bueno a pruebas que sustenten las tablas normativas con máximo 15 años de antigüedad.

Actualmente el COTAN difunde información técnica de las pruebas por medio de un boletín en el que los profesionales pueden consultar los aspectos técnicos de las mismas (Evers, Sijtsma, Lucassen, & Meijer, 2010).

Los problemas asociados a la falta de control en la distribución de las pruebas y las consecuencias de la falta de regulación en la calidad de las mismas en Europa; fueron las razones por las cuales se estableció en 1976 la *International Test Commission* (en adelante, ITC), mediante la cual se consolidó la propuesta que se venía gestando en Suiza en la década de los 60. Después de varios acercamientos, algunas asociaciones gremiales del orden nacional y local se unieron para intercambiar información sobre estrategias para mejorar las prácticas de uso de pruebas y divulgar los procesos de revisión de mismas, y conformaron lo que es hoy la agremiación más grande especializada en el tema de la calidad de las pruebas (Oakland, 2004; Oaklang, Poortinga, Schlegel, & Hambleton, 2001; Cardinet, 1995)

Otro proceso de revisión de pruebas europeo surgió en 1986 en el Reino Unido gracias a la Sociedad Británica de Psicología, (*British Psychological Society*, BPS). Consciente de la falta de preparación de los profesionales en el uso de las pruebas, la BPS inició procesos de

certificación profesional, estrategia que se complementó con la valoración de la calidad técnica de las pruebas para uso en los contextos organizacionales (Bartram, Lindley, Foster, & Marshall, 1992; Bartram, 1992). Los procesos de cualificación de profesionales se desarrolló a través de casas comercializadoras y editores, quienes los certificaba para el uso de pruebas específicas; por diferentes controversias entre grupos gremiales, se cambió el enfoque y se conformó el *Psychological Testing Center* (PTC, en adelante) que se encarga de valorar la calidad técnica de los pruebas. El PTC funciona con el auspicio del Consejo de Administración de la Sociedad Británica de Psicología, y hace más de 20 años lidera estrategias de certificación de profesionales, evaluación de las pruebas y registro de las pruebas para promover estándares técnicos en el diseño y desarrollo de las mismas.

Por la misma época y por razones similares a las del Reino Unido, en Alemania en 1986 se estableció el Comité Permanente de Pruebas, organización encargada de promover buenas prácticas de uso de pruebas; inició con acciones enfocadas a la certificación por medio de cursos, que se complementaron con el proyecto de creación de la norma la DIN 33430 que apareció en 2001. Esta constituye una norma legal reglamentaria de los aspectos éticos, las condiciones de cualificación para el uso y la regulación de los aspectos técnicos del uso de pruebas (Hagemeister, Kersting, & Stemmler, 2012). La norma DIN 33430 fue el marco para el desarrollo del sistema alemán *Test Beurteilung System des Test kuratoriums* (TBS-TK); Sistema de Evaluación Técnica de la Comisión de las Pruebas; conocido también como DIN, que inició el proceso de revisión de pruebas en 2006 (Kersting, 2006; Moosbrugger et al, 2009)

Las experiencias del Reino Unido, Holanda y Alemania fueron los principales insumos para el modelo de la Federación de Asociaciones Europeas de Psicología (*European Federation Psychological Associations*, EFPA), que recogió en un instrumento, los desarrollos y fundamentación conceptual adelantada por estas tres asociaciones locales para la revisión de las pruebas, y que se consolidó como un instrumento guía para la revisión de las pruebas en los países que conforman la Unión Europea (Muñiz & Bartram, 2007; Bartram, 2001, 2006).

La EFPA conformó un comité de pruebas para promover las buenas prácticas de uso de pruebas incluyendo en sus acciones la promoción del valor de la calidad técnica; por lo cual

promueven el instrumento modelo para evaluación de las pruebas, como herramienta que puede ser adaptada por las diferentes asociaciones europeas locales. Este inventario se publicó por primera vez en la sitio web (www.efpa.eu) en 2002 (Bartram 2001, 2002), y tiene dos actualizaciones en las que incluyeron criterios de calidad para el uso de recursos tecnológicos en la evaluación, tales como las aplicaciones computarizadas, informes automáticos e informes de análisis desde modelos de Teoría de Respuesta al Ítem (Lindley, Bartram, y Kennedy, 2008).

3. América latina: El Modelo Brasileño y la propuesta mexicana

Mientras en Estados Unidos se publican las directrices con mayor acogida en América y funciona el Centro Buros que publica el anuario evaluación de pruebas, y en Europa la EFPA promueve el uso de un herramienta de evaluación calidad técnica en los países de la Unión; en los países de centro y sur América los avances han sido menos significativos.

Al momento de realizar este documento el único sistema de evaluación de la calidad de pruebas en la región, es el brasileño: Satepsi, que inició su desarrollo en década de los 90 y se estableció oficialmente en 2003 mediante la resolución 002, del *Conselho Federal de Psicologia*. El sistema parte de una regulación normativa que introdujo la exigencia de criterios mínimos obligatorios de calidad técnica para el uso de pruebas psicológicas en Brasil y estableció la metodología de evaluación y los criterios para dar aval de uso en territorio brasileño (Weschler, 2003, 2011)

Aunque fue recibida con resistencia por su carácter restrictivo en el uso de prueba, la entrada en vigencia de la ley configuró un escenario que ha transformado las prácticas de construcción, adaptación y uso de las pruebas psicológicas, promovió el desarrollo de estudios con población brasileña para el desarrollo de parámetros de interpretación, impulsó el desarrollo de pruebas y transformó las relaciones entre comercializadores, académicos y profesionales. El resultado de este proceso de evaluación está en línea en la página web de SATEPSI: www.satepsi.cfp.org.br (Porto, 2012).

El modelo brasileño definió criterios mínimos a partir de la revisión de la propuesta española publicada en 2000 (Muñiz et al., 2011) y desarrolló un instrumento considerando las

directrices de los estándares vigentes para la evaluación educativa y psicológica de la APA. Logró entonces consolidar un instrumento para evaluar las pruebas objetivas y proyectivas el cual ha sido utilizado para revisar más de 110 pruebas hasta el momento; los resultados se divulgan al público mediante portal web del sistema *satepsi* disponible en línea en [www http://satepsi.cfp.org.br/](http://satepsi.cfp.org.br/) (Weschler, 2013).

Finalmente, vale la pena mencionar la propuesta de directrices para el desarrollo de pruebas disponible en el libro: “Estándares de Calidad Para Pruebas Objetivas” de Agustín Tristán y Rafael Vidal, que constituye la única propuesta de estándares para el desarrollo de pruebas en español (Tristán y Vidal, 2006). Aunque fue concebida y preparada en el Instituto de Evaluación e Ingeniería Avanzada de México, la primera edición de esta publicación apareció en 2006 editada por la Cooperativa Editorial Magisterio de Colombia y la segunda edición se encuentra en preparación.

Esta publicación ofrece una cuidadosa revisión de los elementos conceptuales y metodológicos importantes para el desarrollo de pruebas, y una guía práctica que dispone de diversos formatos para la revisión de indicadores de calidad en el proceso. Es un modelo de revisión de aspectos de forma y fondo que se deben considerar en la construcción de pruebas, apoyados en comentarios y recomendaciones que permiten identificar los aspectos prácticos asociados a los diferentes criterios. Estos estándares se han utilizado para el desarrollo de más de 60 pruebas aplicadas de México, El Salvador, Colombia, Perú y la República Dominicana, desde nivel preescolar hasta certificación profesional.

Los antecedentes actuales evidencian que el desarrollo de sistemas de evaluación de calidad de pruebas es incipiente en Latinoamérica, considerando el número de países que no cuentan con mecanismos de control o vigilancia de calidad de la amplia cantidad de pruebas disponibles. Esta realidad puede explicarse debido a la complejidad del proceso para el desarrollo de un sistema que conlleva años de trabajo y la superación de diferentes retos: exige la comunicación entre academia, profesionales y comercializadores e implica la concertación de los actores involucrados en el desarrollo de las pruebas y las políticas de control y acompañamiento al ejercicio profesional

Modelos de directrices o estándares

El objetivo fundamental de este tipo de propuestas es divulgar criterios, directrices o recomendaciones, generalmente acordados entre grupos académicos y profesionales involucrados en tareas de desarrollo de pruebas, que se consideran condiciones necesarias para que las prueba construida tenga una adecuada calidad o para que su aplicación, calificación e interpretación sea adecuada.

Los tres modelos de estándares o directrices de esta primera categoría son: *The Standards for Educational and Psychological Testing* de la APA, AERA y NMCE (1999, 2014), las directrices de la *International Tests Commission* y la propuesta del Instituto de Evaluación e Ingeniería Avanzada de México (Tristán y Vidal, 2006). Los dos primeros son resultado de consensos en los que participan equipos académicos que se centran en revisar los asuntos conceptuales y técnicos relevantes para la construcción, desarrollo y uso de las pruebas en diferentes contextos. En general, éstos se divulgan entre profesionales en ejercicio por medio de documentos que describen los criterios de calidad.

Los Estándares de la APA, AERA y NMCE incluyen aspectos conceptuales y recomendaciones técnicas de una amplia gama de temas desde la conceptualización de la prueba hasta la aplicación en contextos particulares o para usos específicos. La versión del 2014 recoge las principales controversias académicas sobre la validez de las pruebas y la equidad de las medidas, temas que han generado acaloradas discusiones en las últimas décadas. En particular, sobre el tema de validez brinda algunas directrices para los procedimientos de validación a través de algunas definiciones de cinco tipos de evidencias de validez. La tabla 1 presenta la organización de los contenidos de la versión 2014.

La ITC divulga por medio de documentos separados algunas directrices que se constituyen en guías orientadoras sobre diferentes aspectos para el desarrollo, características, aspectos técnicos y uso de pruebas. Cada documentos trata un tema específico, brinda con claridad guías de utilidad para los usuarios y desarrolladores de pruebas y de divulga de manera gratuita a través de la página WEB de la ITC (www.intestcom.org). Muchos de éstos se han

traducido a diferentes idiomas y se han convertido en guías muy útiles no solo para los países europeos (International Test Commission, 2017).

Los temas que se han desarrollan en los documentos descargables son:

1. Directrices de la ITC sobre la adaptación de pruebas
2. Directrices de la ITC sobre el uso de las pruebas.
3. Directrices de la ITC para la evaluación computarizada
4. Directrices sobre la seguridad de los test, exámenes y otras evaluaciones.
5. Directrices de la ITC de control de la calidad
6. Directrices sobre la eliminación de la prueba
7. Guía para los evaluados sobre sus derechos
8. Declaración del ITC sobre el uso de pruebas para la investigación.

Tabla 1

Contenidos de los estándares de la APA, AERA y NMCE (2014)

Apartado	Capítulos
Parte I: Fundamentos	Validez. Confiabilidad / precisión y errores de medición. Equidad en las pruebas.
Parte II: Operaciones	Diseño y desarrollo pruebas. Calificación, escalas, las normas, la puntuación y puntajes de corte. Administración de la prueba, la puntuación, la presentación de informes y la interpretación. Documentación de apoyo para las pruebas. Los derechos y responsabilidades de los examinandos. Los derechos y responsabilidades de los usuarios de prueba.
Parte III: Probando aplicaciones	Pruebas y evaluación psicológica. Las pruebas y la acreditación del lugar de trabajo. Pruebas y evaluación educativa. Usos de las pruebas para la evaluación de programas, estudios de políticas y la rendición de cuentas.

El único referente publicado en español es la publicación sobre estándares de pruebas de Tristán & Vidal (2006) que presenta los criterios a considerar en el proceso de desarrollo de las pruebas desde la identificación de los actores responsables del proceso hasta la presentación de resultados e informes de resultados de la prueba. Incluye una serie de guías en forma de protocolos y listas de chequeo que tienen como objetivo facilitar la planeación de las actividades necesarias para el desarrollo de la prueba y verificar rápidamente el cumplimiento de estas recomendaciones. Esta publicación organiza sus contenidos en los siguientes apartes:

1. Órganos responsables
2. Manual Técnico y planeación
3. Validez asociada con la prueba
4. Reactivos y objetividad
5. Confiabilidad relativa a la prueba
6. Construcción de las pruebas
7. Interpretación de resultados
8. Materiales de la prueba
9. Proceso de aplicación y logística
10. Presentación de resultados y uso
11. Promoción y contratación

Modelos de evaluación de las pruebas

La evaluación de la calidad de las pruebas es un proceso orientado a la supervisión de las herramientas usadas por los profesionales en los procesos de evaluación; requiere articular decisiones políticas, requerimientos técnicos y consideraciones éticas con recursos económicos, humanos e institucionales. La implementación de un sistema de evaluación conlleva acuerdos, discusiones, y esfuerzos gremiales que se concretan en la existencia de una infraestructura que permita la gestión del proceso por parte de entidades competentes, con

reconocimiento, autoridad y capacidad para divulgar y ofrecer diferentes tipos de acceso a los resultados de las evaluaciones (Lim, Geranpayeh, Khalifa, & Buckendahl, 2013).

Los procesos desarrollados por las asociaciones locales para evaluar la calidad de las pruebas han definido indicadores que operacionalizan los modelos y que suelen actualizarse periódicamente para recoger los avances conceptuales y metodológicos, además, las actualizaciones suelen modificar los criterios de evaluación según las necesidades percibidas y el impacto de las evaluaciones. Los procesos de revisión de calidad de pruebas psicológicas referenciados en este trabajo, son supervisados por agremiaciones profesionales psicológicas, que por medio de comités o mesas de trabajo, disponen recursos económicos, conforman listas de profesionales idóneos y supervisan la evaluación.

La evaluación se implementa por medio de protocolos de revisión en los que participan por lo menos dos psicólogos con la dirección de un árbitro que retroalimenta el proceso y consolida el informe técnico sobre la prueba. En todos los modelos revisados la divulgación de los resultados es parte fundamental de las estrategias de promoción de mejores prácticas de uso de las pruebas y sólo en Brasil la evaluación es usada como criterio para el aval de uso de las pruebas (Lindley & Bartram, 2012; Evers, et al., 2013).

El sistema Buros.

El método de revisión que se implementa en el Centro Buros incluye 11 pasos en los cuales se cuenta con la participación de una red de expertos que elaboran y revisan el concepto técnico, producto de la revisión de la prueba. Los pasos del proceso son: (1) identificación de las pruebas a revisar, (2) obtención de pruebas y consolidación de las descripciones de pruebas, (3) revisión de requisitos mínimos, (4) identificación de perfiles de evaluadores pertinentes, (5) selección de los revisores, (6) revisión por parte de los evaluadores seleccionados, (7) retroalimentación a los comentarios por parte del desarrollador, (8) corrección y actualización de conceptos con la información enviada por los desarrolladores como respuesta a los comentarios, (9) aprobación de las correcciones por parte del evaluador, (10) revisión de conceptos por parte de editores y (11) publicación.

En la actualidad, en el proceso participan como evaluadores, expertos de diferentes nacionalidades, quienes reciben las guías que orientan el desarrollo de la revisión y los comentarios. Esta documentación presenta criterios construidos considerando los estándares de la APA, AERA & NMEC (2014).

Tabla 2

Modelo de evaluación del centro Bueros.

Sección	Contenidos
Descripción	Resumen de uso propuesto, población objetivo y usos previstos de los resultados de las pruebas, incluye una visión general de procedimientos de administración y documentación disponible y los procedimientos de puntuación.
Desarrollo	Resumen del proceso de desarrollo de la prueba, fundamento teórico, dimensiones de pruebas y pruebas piloto (si corresponde) y detalles de los procesos de selección de elementos / retención.
Técnica Documentación de la información técnica dividida a menudo en tres subsecciones	Estandarización Información sobre la muestra normativa, incluyendo concepto de adecuación de muestra con la población para la cual se ofrece, también puede referirse a las normas para diferentes grupos de género o étnico/cultura.
	Confiabilidad Comentarios respecto a la calidad de la evidencia de consistencia de la puntuación, incluye la descripción y magnitud de las estimaciones de fiabilidad.
	Validez Comentarios sobre la información de las evidencias que sustentan las interpretaciones y los usos potenciales de los resultados de la prueba, incluidos los procedimientos o estudios destinados a investigar la idoneidad de las interpretaciones.
Comentario	Presenta las fortalezas y debilidades de la prueba en general, es una valoración descriptiva de los evaluadores, puede incluir comentarios sobre los estudios y la investigación que sustenta la prueba.
Resumen	Presenta las conclusiones y recomendaciones relacionadas con los usos apropiados de la prueba.

El concepto técnico sobre la prueba evaluada está compuesto por cinco apartados: descripción, resumen del desarrollo de la prueba, fundamentación técnica, comentarios sobre fortalezas y debilidades y conclusiones de la revisión, como se muestra en la tabla 2. La

revisión y edición para publicación de los conceptos en el Centro Buros es realizada por estudiantes con entrenamiento y el proceso incluye la posibilidad de corrección y comentarios por parte de los editores, de manera que el concepto que se publica es revisado por las editores de las pruebas (Carlson & Geisinger, 2012).

Este proceso integra la experticia profesional de los evaluadores y los criterios técnicos, de manera que pone a disposición de los usuarios, los comentarios para orientar el uso adecuado de la prueba. Además, la corrección, revisión y aprobación que se realizan durante la producción del informe final de la pareja de expertos, permiten control y conciliación de los criterios para concretar un concepto cualitativo que apoye la toma de decisiones para la selección de las pruebas.

El proceso alemán. Norma DIN

El Colegio Oficial de Psicólogos de Alemania y la Sociedad Alemana de Psicología unieron esfuerzos para que el sistema de revisión de pruebas alemán (Testbeurteilungssystem des Diagnostik- und Testkuratoriums, TBS -TK) desarrolle y disponga en línea las evaluaciones de las pruebas. En la actualidad el proceso se desarrolla sistemáticamente en 13 pasos que van desde la selección de las pruebas hasta la publicación de los resultados de la evaluación.

Sin embargo, a diferencia de otros sistemas, la evaluación se realiza solo si la prueba cumple con los pre-requisitos exigidos por la norma DIN 33430, los cuales se verifican en el momento de la solicitud de certificación de la prueba, mediante las listas de chequeo disponibles en la página web (www.zpid.de/index.wahl=Testkuratorium) del Instituto Leibniz de Información Psicología (Das Leibniz-Zentrum für Psychologische Information und Dokumentation, ZPID), entidad encargada de los procesos de evaluación de las pruebas. Posterior a la verificación de los requisitos mínimos, una pareja anónima de expertos del área de uso de la prueba, realiza la evaluación siguiendo el protocolo que incluye los criterios de la guía de control (Kersting, M., 2008).

La tabla 3 muestra los aspectos evaluados en el modelo alemán; el modelo de evaluación es semiestructurado y combina la revisión de criterios definidos y las apreciaciones y

comentarios de los evaluadores; mediante un documento de máximo 9000 caracteres; a partir de una escala cualitativa se dan a conocer los comentarios sobre cada uno de los aspectos evaluados, de acuerdo con la clasificación de las pruebas adoptado por la ZDIN. Los conceptos son revisados por un experto de la comisión delegada de pruebas y se unifican después de conciliar los comentarios de cada evaluador; éstos se envían al desarrollador para que tenga oportunidad de revisión y réplica por parte de los editores, después de lo cual se consolida la versión para publicación (Eignor, 1999).

Tabla 3

Contenidos evaluados en el modelo alemán

Aspecto evaluado	Descripción
	Acerca de la descripción del objetivo de la prueba y el uso de la misma
	Elementos conceptuales que fundamenta la construcción de la prueba
3. Características Técnica	Objetividad Estandarización (calibración) Confiabilidad Validez Otros factores de calidad (Susceptibilidad, incorruptibilidad y escalas)
4 Evaluación final Recomendación	Resumen de los resultados de las anteriores

Fuente adaptada de. Testkuratorium (2010). TBS-TK-Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009. Psychologische Rundschau, 61 (1), 52-56. http://www.kersting-internet.de/pdf/tbs_tk_2010_rundschau_61_52-56.pdf (PDF)

La norma alemana DIN desarrolló una clasificación de pruebas y unos criterios para evaluar las pruebas de cada categoría, esta información puede ser consultada por los profesionales por medio de la base PSYNDEX; e incluye diferentes tipos de revisiones: descripciones detalladas de los procesos de desarrollo de las pruebas, descripciones cortas de pruebas, una amplia gama de procesos diagnósticos como pruebas, escalas, cuestionarios, técnicas de entrevista, métodos de observación, protocolos de prueba, métodos de diagnóstico asistido por ordenador e instrumentos de diagnóstico.

El sistema de evaluación holandés. Asociación Holandesa de Psicólogos.

Los revisores de pruebas en Holanda son anónimos al igual que en Alemania, la COTAN convoca una pareja de expertos en el área de uso de la prueba, quienes aplican el protocolo y emiten un concepto en cada uno de los aspectos evaluados, los resultados de las revisiones se compilan en un informe técnico. Cada aspecto del protocolo se califica en una escala del 1 a 3, la puntuación se suma y pondera en una puntuación general, que se interpreta por medio de una escala cualitativa de calificación; a partir de la revisión de la información disponible en el manual técnico el evaluador puede valorar cada indicador como "suficiente" (3), "inadecuado" (2) e "insuficiente" (1). La tabla 4 presenta las generalidades de los componentes evaluados por el sistema holandés.

Tabla 4

Modelo de la evaluación de pruebas en Holanda

Aspecto evaluado	Descripción
Bases teóricas de la prueba	Conceptos teóricos desde los cuales se fundamenta el constructo evaluado.
La calidad de los materiales de la pruebas	Este apartado revisa los implementos para la aplicación de las pruebas, documentos, formato, plantillas de puntuación y elementos para ejecución, y las instrucciones, sustento de estandarización y adecuación de las instrucciones.
Comprensibilidad del manual.	Exhaustividad de las instrucciones para el éxito de la administración de las pruebas, inclusión de descripciones de casos, disponibilidad de indicaciones para la interpretación del puntaje, declaraciones sobre la calificación del usuario
Normas	La calidad de las normas y vigencia de los datos normativos suministrados
Confiabilidad	La magnitud de los coeficientes de fiabilidad crítica, y a continuación, la calidad de la salida investigación de la fiabilidad
Validez de constructo.	Valora la calidad y resultados de la investigación realizada para probar los usos indicados de forma explícita y obtener los datos que apoyan la validez de constructo
Validez de criterio	Valora los resultados y calidad de la investigación para demostrar la validez de criterio.

Fuente: http://www.iops.nl/wp-content/uploads/2015/02/Cotan_2015_RobMeijer.pdf descargado en abril de 2015

Si hay discrepancias entre los evaluadores, se acuerda una reunión para conciliar los puntos de controversia y en caso de no lograr el acuerdo, se remite a un tercer evaluador quien toma

la decisión sobre el asunto donde no hubo acuerdo. El proceso es supervisado por un editor y el concepto consolidado se envía al editor de la prueba para que responda a las observaciones, si hay lugar a ajustes en el informe, estos se hacen antes de la publicación (Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. 2010).

Desde 1961 el Centro de Recursos de Evaluación Holandés y el COTAN divulgan la información de la evaluación de las pruebas y de investigaciones de uso de las mismas, en la página web <http://www.cotandocumentatie.nl/>; ésta se actualiza mensualmente con descripciones y nuevas evaluaciones de calidad de pruebas, actualmente dispone para consulta información de la evaluaciones de más de 700 pruebas disponibles en los Países Bajos (Evers, A., Klaas, S., Lucassen, W., Meijer, R. 2010). La publicación divulga los conceptos de pruebas que cuentan con investigación con población holandesa en normas, fiabilidad o validez incluye extractos de las investigaciones donde se han utilizado.

El Sistema del Centro de Evaluación del Reino Unido

En el Reino Unido el PTC evalúa y pone a disposición de los usuarios la información técnica de las pruebas; esta institución funciona como parte de la organización que acoge el Comité de Normas de Ensayo (STC) de la Sociedad Británica de Psicología (BPS) y es la encargada de las actividades relacionadas con la divulgación de la información sobre la valoración de las pruebas psicológicas; aunque desde su fundación el enfoque principal de sus actividades ha sido la cualificación de los profesionales, la estrategia requirió la creación de un proceso de evaluación y clasificación de las pruebas.

Este proceso se viene implementando desde los noventa, y desde 2013 se modificó acogiendo el protocolo de evaluación de calidad europeo propuesto por la EPFA. El proceso inicia cuando los editores diligencian una solicitud, adjuntan el material disponible para los usuarios y cancelan los costos; únicamente si cumplen los criterios de referencia mínimos para la inscripción de la prueba, se inicia el proceso. Sólo se evalúan pruebas que cumplan con el nivel de "adecuado" según la definición de los criterios del modelo derivado del examen de EFPA, en términos de validez, confiabilidad y normas o cualquier otra información necesaria para la interpretación significativa de las puntuaciones. Un editor senior verifica estos

requisitos, en caso de no cumplir los requerimientos se clasifica en un sistema de semáforo: verde, ámbar y rojo, clasificación que es notificada al editor para que subsane las fallas y someta de nuevo la prueba a evaluación.

El cumplimiento de los requisitos exigidos otorga un certificado de registro de la prueba que informa al público que ésta reúne los criterios mínimos de referencia; además, posterior al proceso de evaluación se dispone del resumen de la valoración consolidada por el editor senior del centro y revisada por el editor de la prueba.

Entre las estrategias orientadas a mejorar las prácticas de uso de pruebas, el PTC dispone de una página web (<https://ptc.bps.org.uk/>) en la que proporciona acceso a la información sobre las mismas a los profesionales y evaluados; esta información muestra los puntos fuertes y las debilidades de las diferentes pruebas en el mercado, de manera que se convierte en un criterio importante para la selección de las mismas.

Proceso modelo para países de la Unión Europea

El desarrollo de los procesos de evaluación en Holanda, Reino Unido y Alemania impulsó el interés en la evaluación de las pruebas en los países de la Unión Europea. Mientras en el Reino Unido la evaluación fue parte de la estrategia de cualificación de los profesionales, en Holanda y Alemania los esfuerzos se enmarcaron en una tendencia cultural de promoción de calidad.

Durante la reunión del Comité Permanente de Pruebas y Ensayos de la EFPA en 2001, se conformó el grupo de trabajo para el Estándar Europeo y se acordó desarrollar un proyecto para promover criterios de evaluación de calidad que pudieran ser útiles para toda Europa. Representantes de Reino Unido, Suecia y Países Bajos acordaron trabajar para establecer las directrices y desarrollaron una guía de evaluación ilustrativa que compilara las experiencias de los diferentes países europeos (Bartram, 2006).

El grupo de trabajo se conformó entonces con representantes de Holanda, Alemania, Noruega, Reino Unido, España, y Suecia; miembros de los comités encargados de promover buenas prácticas de uso de las pruebas, en sus respectivos países. Como resultado de este

trabajo conjunto se desarrolló el instrumento modelo basado en las recomendaciones de la *International Tests Commission* (ITC), un prototipo de instrumento oficial para la Unión Europea que retomó los criterios de los modelos británico y holandés. La versión actual "Modelo de Revisión de la EFPA para la Descripción y Evaluación de Pruebas Psicológicas" se publicó en julio de 2013, después de la revisión de la versión de 2008 (ver tabla 5).

Tabla 5

Aspectos evaluados en el instrumento para revisión de calidad técnica de las pruebas de la EFPA

Aspecto evaluado	Información revisada
Descripción del instrumento	Descripción general
	Clasificación
	Medición y puntuación
	Informes generados por computadora
	Condiciones y costos de suministro
Calidad de la explicación de los fundamentos la presentación y la información proporcionada	Calidad de la explicación de los fundamentos
	Adecuación de la documentación disponible para el usuario
	Calidad de las instrucciones de procedimiento previstas para el usuario
Calidad de los materiales de ensayo	Concepto sobre la información
	Calidad de los materiales de la prueba de papel y lápiz
	Calidad de los materiales de prueba de análisis por ordenador o Ensayos basados en la Web
Normas	Concepto sobre materiales de prueba
	Interpretación referida a la norma
	Criterio de interpretación que se hace referencia
Confiabilidad	Concepto sobre las normas
	Aspectos técnicos de los estudios
Validez	Concepto sobre confiabilidad
	Evidencias de validez de Constructo
	Evidencias de validez de Criterio
Calidad de los informes generados por computadora	Validez general
Evaluación final	Conclusión

Aspecto evaluado	Información revisada
Concepto de uso	

La EFPA (2013) propone un proceso de revisión similar a la habitual evaluación por pares de trabajos y proyectos científicos, involucra al menos dos revisores independientes y un editor de consultoría quien se encarga de supervisar las revisiones y puede convocar a un tercer revisor si se encuentran discrepancias significativas entre las dos opiniones. Las organizaciones nacionales pueden incluir variaciones en el procedimiento, siempre y cuando se garantice la cualificación e independencia de los revisores.

La EFPA recomienda que los informes de evaluación se dirijan a los usuarios de prueba cualificados de acuerdo con la normas de certificación que se ha implementado en Europa, también deben ser comprensibles para los académicos, autores de prueba y especialistas en psicometría y pruebas psicológicas.

La tabla 5 se muestra los aspectos que evalúa el instrumento modelo de la EFPA, una estructura con dos apartes, una descriptiva y una de valoración técnica; la primera parte incluye 5 temas y recoge la información que describe las características generales de la prueba; la segunda parte contempla 7 grandes apartados que valoran los aspectos técnicos de la prueba. En la actualidad la EFPA adelanta acciones para promover la adaptación del instrumento en diferentes países, impulsando a las asociaciones locales de psicología para que adapten el protocolo de evaluación; además, ofrece asesoría a las mismas para que gradualmente se unifiquen los criterios de valoración, de manera que a largo plazo las normas sean comparables en los distintos países de la Unión Europea.

Sistema Satepsi de Brasil

En la actualidad el único sistema oficial sur americano que evalúa la calidad técnica de las pruebas es Satepsi en Brasil. El desarrollo de este sistema de la Federación Brasileira de Psicología (FBP), es una experiencia ejemplar en Latinoamérica; para su desarrollo la FBP convocó a psicólogos de los 23 estados federales quienes en dos años consolidaron el protocolo que hizo parte de la resolución 002 de 2003, norma que oficializó los criterios para la revisión técnica de las pruebas.

Esta normatividad es la consolidación de una década de trabajo liderado por académicos de diferentes universidades brasileñas que en la década de los 90 del siglo pasado trabajaron en el desarrollo de pruebas y promovieron la discusión sobre la necesidad de mejorar la calidad de las mismas, discusión que dio paso a la fundación del Instituto Brasileño de Evaluación Psicológica (IBAP en adelante) en agosto de 2002.

El IBAP se ocupa de todo lo relacionado con la evaluación psicológica: capacita a profesionales, acoge el listado de revisores de calidad de las pruebas; presta asesoría especializada en el desarrollo de pruebas y estudios de validación y además, se involucra en la planificación y realización de investigaciones y publicaciones. Estas actividades las realiza en alianza con diferentes organizaciones, escuelas y clínicas de salud mental que ofrecen servicios de evaluación. Cuenta, además, con una revista técnica-científica llamada “Evaluación Psicológica”, que se distribuye gratuitamente a los miembros del IBAP y es la principal publicación la producción académica especializada en evaluación en Latinoamérica.

El proceso en Brasil al igual que en Europa, lo realizan pares expertos que revisan el material de la prueba disponible, y si hay diferencias en conceptos se solicita la intervención de un consultor que arbitra los conceptos. Toda prueba para aprobación de uso se somete a un proceso que dura aproximadamente 6 meses.

La revisión de calidad Brasileña sigue un protocolo de tres (3) apartes que examinan seis (6) aspectos de calidad a partir de indicadores con criterios mínimos de valoración; este protocolo es una adaptación del modelo propuesto para la revisión en España (Prieto & Muñiz, 2000) y consideró las directrices técnicas de la ITC, la Asociación Americana de Psicología APA, y la Asociación Canadiense de Psicología CPA. Ver tabla 6 con la información del modelo de evaluación que se reglamentó.

Tabla 6

Modelo brasileño de evaluación

Capítulo	Indicadores y comentarios
Base teórica del instrumento	La definición el constructo que se evalúa y propósito del instrumento e información de los contextos de uso propuesto.

Capítulo	Indicadores y comentarios
Validez	Estudios que aportan evidencia empírica de validez y exactitud de las interpretaciones propuestas de resultados de las pruebas, se valora la justificación de los procedimientos específicos utilizados en la investigación.
Sustento psicométrico	Información disponible sobre el análisis de los datos empíricos de funcionamiento de la medida, las propiedades psicométricas de los ítems que componen el instrumento.
Estandarización y calificación	<p>Información sobre procedimientos de corrección e interpretación de los resultados, explicando la lógica subyacente en el procedimiento, dependiendo del sistema de interpretación adoptada, que puede ser:</p> <ol style="list-style-type: none"> a) debe incluir información sobre las características de la muestra usada en el estudio de estandarización de manera clara y completa, preferentemente mediante la comparación con estimaciones nacionales, lo que demuestre la representatividad del grupo de referencia utilizado para el levantamiento de los datos o tablas de transformación de las puntuaciones. b) debe sustentar la interpretación estándar de referencia, en cuyo caso, explicar la base teórica y justificar la lógica del procedimiento de interpretación utilizado (para pruebas de enfoque psicoanalítico)
Estandarización aplicación	Información sobre los procedimientos para establecer las normas de corrección y las condiciones en que la prueba se debe aplicar para garantizar procedimientos uniformes que intervienen en su solicitud
Manual técnico	<p>Calidad de la compilación de la información anterior, así como otros que son importantes en un manual que contenga al menos información sobre:</p> <ol style="list-style-type: none"> a) el aspecto técnico y científico, que describe los estudios y fundamentos empíricos sobre el instrumento. b) el aspecto práctico, explicando la aplicación, corrección e interpretación de los resultados. c) los medios científicos de la literatura que sustentan la interpretación de los resultados.

Fuente elaborado a partir de la RESOLUÇÃO CFP N° 002/2003 del Conselho Federal De Psicologia

En general, esta revisión muestra que los diferentes sistemas o procesos incluyen criterios de calidad consensuados en las asociaciones locales, y su intención primaria es emitir un concepto que oriente a los profesionales en ejercicio y decida sobre la conveniencia de usar o

no la prueba. Los criterios de evaluación que se usan verifican el cumplimiento de las condiciones técnicas evalúan la evidencia relacionada con los procesos de construcción de la prueba y de los estudios para estimar su fundamentación teórica y psicométrica de manera se garanticen resultados que den cuenta de la capacidad de la prueba para medir el constructo. Con excepción de Brasil, el concepto no restringe el uso de la prueba puesto que la intención principal es exhortar a los profesionales a utilizar pruebas que cumplan los criterios de calidad.

Algunas consideraciones conceptuales y metodológicas sobre los criterios técnicos para evaluar la calidad de las pruebas

Para dimensionar el significado de evaluar la calidad de las herramientas de medición usadas por los psicólogos vale la pena traer a colación la definición de prueba propuesta en los Estándares para la evaluación psicológica y educativa:

“...a test, broadly defined, is a set of task or stimuli designed to elicit responses that provide a sample of an examinee’s behavior or performance in a specified domain”

(AERA,APA,NCME 2014 pág. 33)

Las pruebas son herramientas de medida, y su desarrollo implica la adopción de una postura teórica, el diseño de una escala para categorizar el atributo evaluado, la obtención de evidencias empíricas de su validez y confiabilidad, la determinación de un proceso de aplicación adecuado y el establecimiento de directrices para la interpretación de resultados; toda esta información debe documentarse en el manual técnico, que es el material que apoya el uso profesional de la prueba y orienta su uso adecuado y la correcta interpretación de los resultados.

La valoración de calidad de una prueba implica la revisión sistemática del material disponible para los usuarios con el objetivo de identificar las cualidades de la prueba considerando los materiales de los cuales se compone y la fundamentación técnica de la misma como instrumento de medición; en un proceso estructurado para verificar las características técnicas y la rigurosidad de los procesos de desarrollo de la prueba, el evaluador juzga a partir del manual técnico el cumplimiento de los requisitos metodológicos

que requiere el desarrollo de un sistema de medida para un constructo (Bartram, 2012; Carlson & Geisinger, 2012; Evers, y otros, 2013; Lindey & Bartram, 2013; Center Buros 2014).

El desarrollo de una prueba psicológica implica la sinergia de procesos académicos y prácticos que deben seguirse de manera rigurosa para garantizar que la escala de valoración del constructo sea robusta; en consecuencia la revisión de calidad de las pruebas implica que el evaluador comprenda cómo el incumplimiento de los requerimientos metodológicos afecta la calidad de la prueba y que tenga en mente los desafíos propios de sustentar de forma teórica y empírica un escala de medida para una variable latente. La evaluación de los indicadores de calidad en cualquiera de los procesos requiere experticia técnica para identificar como las operaciones empíricas y formales se articularon en el desarrollo de la prueba y la elaboración de la escala de medida (Stevens S. S., 1946; Stevens S. S., 1951; Martínez, 1996). La figura 1 presenta las etapas mínimas que se implementan en el desarrollo de una prueba.

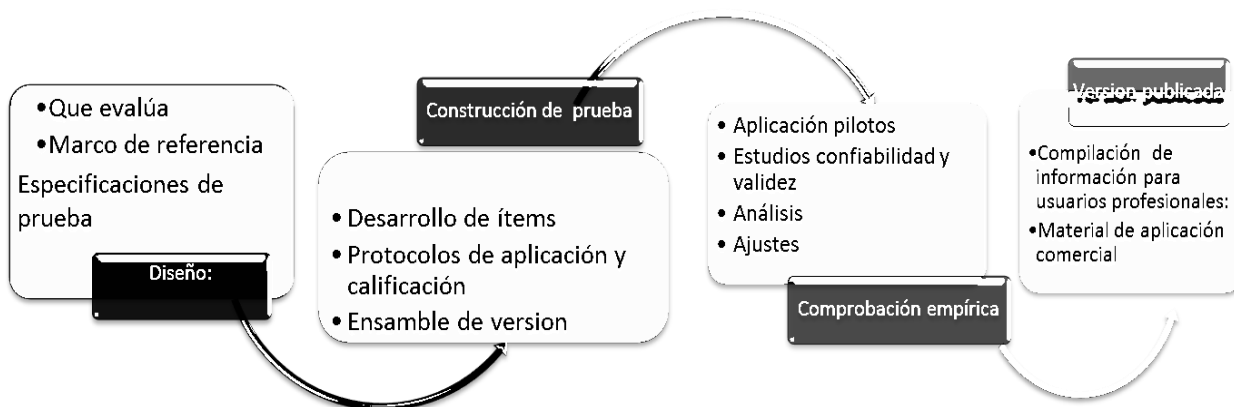


Figura 1. Esquema general de las etapas de desarrollo de una prueba psicológica

En síntesis, una prueba puede entenderse como un sistema de valoración de un constructo, y en el manual técnico debe incluirse la información que permita a los usuarios reconocer los conceptos que la sustentan, para que como usuario calificado pueda interpretar los resultados después de un juicio fundado de la pertinencia de su uso (Nunnally & Bernstein, 1995).

Aunque la estructura del protocolo y los indicadores de valoración varían, todos los sistemas de evaluación referenciados en este trabajo articulan la valoración de tres aspectos: fundamentación conceptual, propiedades métricas e información sobre estandarización de la

prueba. La fundamentación conceptual se entiende como el sustento teórico del constructo evaluado; las propiedades métricas como los resultados de los estudios empíricos para demostrar confiabilidad y validez o pertinencia de las interpretaciones de los resultados y la información de estandarización hace referencia a la disponibilidad de las instrucciones para aplicación y características de las muestras normativas o de referencia para la calificación de la prueba. En los siguientes apartados se describirán los fundamentos teóricos y metodológicos de estos aspectos.

Fundamentación conceptual.

En los modelos y procesos de revisión la primera etapa es la verificación de la fundamentación conceptual o marco de referencia que sustenta el constructo evaluado, un elemento fundamental para el uso de los resultados en un proceso de evaluación; de forma resumida se puede decir que se valoran tres características: suficiencia, precisión y actualización de las referencias conceptuales que dan soporte a la prueba.

Suficiencia. En general se espera que el manual técnico incluya información teórica y conceptual de los conceptos que describan de forma completa y con un nivel adecuado de explicación los constructos que aborda la prueba; incluyendo las aclaraciones pertinentes que permitan a un usuario profesional comprender la argumentación desde la cual se vinculan las conductas observables con el constructo evaluado.

Actualización Se espera que la información conceptual que sustenta la interpretación de los puntajes, se actualice con regularidad para garantizar la pertinencia del uso de la prueba, este criterio de actualización de la información se relaciona con vigencia de las posturas científicas que dan soporte a su uso y la vida útil de las pruebas. La información teórica y conceptual del constructo debe sustentarse en avances científicos vigentes; el usuario debe poder identificar con claridad la actualización de las posturas teóricas desde las cuales se definió la evaluación del constructo, de manera que pueda juzgar la pertinencia de su uso, identificando las limitaciones y avances teóricos que sustenten la interpretación.

La actualización resulta de especial importancia cuando evalúan constructos que pueden cambiar según el entorno social y cultural; sucede en especial en pruebas que usan respuestas

que se asocian a conductas que se extinguen o modifican por efecto del medio social o cultural, de manera que con el tiempo disminuye la sensibilidad de la prueba para estimar la magnitud del constructo. Un ejemplo de ello es el famoso “efecto Flynn” descrito por Richard Herrnstein y Charles Murray (Flynn, J. R., 2007; 2008; 2009a; 2009 b).

Precisión. El sustento conceptual debe dar cuenta de las dimensiones y variables de interés que pretende evaluar la prueba, de manera que el usuario profesional pueda revisar los elementos teóricos relevantes para la interpretación del puntaje a lo largo de toda la escala.

Propiedades métricas de la prueba.

De la misma manera que se valora el soporte conceptual, o marco de referencia teórico que describe el constructo objetivo de evaluación, se revisan la pertinencia, rigurosidad y resultados de los estudios que probaron las características métricas de la prueba: la confiabilidad y la validez (Evers, 2012).

En los manuales técnicos se incluye por lo general un capítulo de sustento estadístico de la prueba, en el que se referencian los resultados de los estudios empíricos que sustentan el funcionamiento de la medida que ofrece la escala de calificación de la prueba. Sobre tales estudios deben referenciarse aspectos como las metodologías empleadas en aplicaciones piloto, los procesos de verificación de calidad de los ítems, las características de las muestras con las que se obtuvieron evidencias de confiabilidad y validez y los procedimientos de análisis estadísticos de los datos, de manera que se documenten las evidencias que sustenten la conveniencia del uso de la escala que propone la prueba.

Los diseños y metodologías implementadas en los estudios para probar el funcionamiento de la prueba deben ser coherentes con la población para la cual se diseñó, con el procedimiento de aplicación que propone y con los aspectos relevantes de la teoría que sustenta su uso. De esta manera la evidencias reportada por los estudios sustenta la pertinencia de la escala de medida que se propone. En general, en la valoración de estos se consideran los aspectos prácticos y la coherencia de los procedimientos respecto al constructo: el número de participantes, características de la población, procedimientos de aplicación y la rigurosidad en

la verificación de los supuestos estadísticos para la aplicación de los algoritmos con los que se analizaron los datos.

Lo más frecuente en los diferentes protocolos de evaluación es que se espere como mínima información para dar cuenta de las características métricas de la prueba, reporte sobre el proceso de desarrollo de la misma y la documentación de por lo menos dos estudios; uno en el que se recolecten datos para analizar la precisión de los resultados, la confiabilidad, y otro sobre la validez, de manera que se presenten la descripción de los procedimientos, la población, los resultados y el análisis de los mismos.

Para la revisión de la pertinencia de los estudios psicométricos, es necesario identificar el modelo teórico bajo el cual desarrollaron los análisis. Cuando el modelo de análisis usado por los desarrolladores es Teoría Clásica de los Test (en adelante, TCT) los estudios reportados analizan las respuestas de los individuos como un puntaje global por cada aspecto que evalúa; es decir, se estima la medida del atributo a partir de todas las respuestas a la prueba o sus sub dimensiones partiendo del supuesto básico que asimila el comportamiento del atributo con las características de un modelo lineal donde la habilidad se entiende como un continuo, en el cual la cantidad de aciertos es proporcional al atributo. De manera similar la dificultad y discriminación de los ítems se estiman a partir del conjunto de respuestas del grupo total, de forma que los parámetros de dificultad y discriminación dependen de las características de la población con la cual se realizó la aplicación, limitación que debe ser considerada para el uso de la prueba en poblaciones diferentes (Muñiz, 2010).

Los estudios que analizan la información desde la TCT deben realizarse con sujetos que tengan la mayor similitud con la población objetivo; y es deseable que los datos provengan de las respuestas de un grupo amplio de sujetos, de manera que se garantice para los análisis de confiabilidad la mayor variabilidad posible; desde este modelo la estimación del error aleatorio de medida es una magnitud constante que tiene el mismo efecto a lo largo de toda la escala de la prueba.

Si el modelo de análisis corresponde al de la Teoría de la Generalizabilidad (en adelante TG) en los estudios de confiabilidad se espera que se identifiquen diferentes fuentes de error y

los estudios deben incluir metodologías de control de condiciones experimentales asociadas a las posibles fuentes de error. De esta forma, se espera que los estudios analicen la variabilidad de los resultados a partir del uso de la prueba en diferentes condiciones, de manera que se estime el impacto de la condición controlada en la precisión del resultado. En general, estos estudios son complejos, exigen el control de variables que puedan afectar el comportamiento del constructo, e implican la selección rigurosa de participantes y condiciones para responder a los supuestos conceptuales desde los cuales se diseñen, resultan apropiados para pruebas diagnósticas que requieran comprobar las diferencias significativas entre grupos, mostrando la sensibilidad de la clasificación que ofrece la prueba.

Finalmente, si el modelo de análisis es el de Teoría de Respuesta al Ítem (en adelante TRI) los estudios deben presentar datos sobre el comportamiento de cada ítem respecto al modelo elegido para el análisis, el ajuste del modelo se relaciona con la sensibilidad del ítem en un nivel determinado de atributo y la función de información es la inversa del error de medida como una función de la magnitud de atributo; en general se acepta que para este tipo de análisis los grupos tengan al menos 200 individuos y la calibración de los ítems supone la disponibilidad de software específicos especializados.

Los dos capítulos obligatorios en los procesos de revisión de calidad de las pruebas verifican los procedimientos y resultados que sustenten la confiabilidad y la validez. Los dos apartes siguientes de este documento revisan brevemente ambos tópicos iniciando por el de la confiabilidad como cualidad imprescindible de la medida y necesaria para la validez (Prieto & Delgado, 2010).

Confiabilidad

La confiabilidad es un aspecto fundamental de la calidad de las medidas psicológicas ya que se relaciona con algunos conceptos claves en medición: error de medida, fuentes de error, error típico de medida, nociones que se han desarrollado para explicar la variabilidad de los puntajes observados y estimar el rango en el cual puede fluctuar las estimaciones de la magnitud del atributo; es decir de la precisión o exactitud de la estimación depende la adecuada interpretación del resultado.

Martínez (1996) define la confiabilidad como la propiedad métrica que se relaciona con la capacidad de la prueba para ofrecer una estimación de la magnitud del atributo con la menor cantidad de error posible, de manera que el resultado observado sea lo más cercano posible a la realidad; y la importancia de su análisis radica en que es una condición indispensable de la medida; y puede entenderse de forma general como la precisión o consistencia del resultado obtenido a través del uso de la herramienta de medición a lo largo de varias mediciones. En consecuencia, el reporte de alguna estimación del error de medida de la prueba, es una condición indispensable para hacer una adecuada interpretación de los resultados de la misma para un individuo o grupo de individuos.

Además de la estimación del error, el manual técnico de una prueba debe incluir información sobre los estudios en los cuales se analiza el comportamiento empírico de la prueba, procesos de análisis del error que incluyen los resultados. Los estudios deben demostrar que la escala propuesta funciona de forma razonable para considerar que el resultado es una estimación del atributo que se acerca al valor real (puntaje verdadero en la jerga de la TCT) y que la clasificación, categoría o puntaje, realmente describe con precisión la magnitud del atributo. La información que se incluye generalmente en los manuales son los índices o coeficientes de confiabilidad, que ofrecen una estimación plausible de la probabilidad de error en el resultado de la medición.

La selección del procedimiento para estimar la confiabilidad debe hacerse a partir de la identificación de la fuente de error relevante para la prueba y el uso propuesto para la misma; por ejemplo, si se trata de una prueba que conduce a diagnósticos o ubicación del examinado en algún sistema de categorías, el error más importante es el error de clasificación puesto que evalúa la consistencia del resultado en la ubicación del examinado en la misma categoría. Sin lugar a dudas, el índice de confiabilidad más reportado es el coeficiente de confiabilidad *Alpha de Cronbach* (Cronbach, 1951, 1955) que estima el error aleatorio debido a la inconsistencia entre los ítems, tareas o elementos que componen la prueba, fuente de error que, desde la TCT, se considera relevante en todas las pruebas compuestas por ítems.

La interpretación del índice de confiabilidad depende de los valores posibles del estadígrafo a partir del cual se calcula, por ejemplo para el Alpha de Cronbach, el valor fluctúa de menos

infinito a uno, y cuanto más cerca a uno, se interpreta como más confiable (Cronbach, 1951, 1955); varios autores plantean algunas reglas para juzgar estos resultados pero en general se consideran adecuados los valores superiores a 0.7, y se recomienda que para uso de pruebas en procesos diagnósticos éstos deben ser superiores a 0,8 (Martínez, 1996).

Además del alpha de Cronbach, que se reporta en la gran mayoría de manuales de prueba, los procedimientos de análisis que se referencian para sustentar la confiabilidad dependen del modelo de análisis que se adopte y de la fuente de error relevante. Bajo la TCT los más frecuentes son los que usan la correlación entre los puntajes de aplicaciones repetidas o entre formas de la prueba o entre grupos de ítems, para mostrar la estabilidad de las medidas o la baja variabilidad de los puntajes, desde la presunción que una adecuada asociación entre los resultados en las aplicaciones es evidencia de la precisión o consistencia de la medida. Estas correlaciones tienen las limitaciones propias de este estadístico y como norma general para valorarlas como evidencia de la consistencia de la prueba, se sugiere interpretarlas en términos de porcentaje de varianza compartida entre las dos medidas. El *valor p* asociado al estadístico puede ser muy poco informativo cuando se tienen tamaños de muestra grandes mientras que el cuadrado de la correlación ($100r^2$) como porcentaje de varianza, es una medida del tamaño del efecto que resulta más útil en la gran mayoría de los casos.

Si se adoptan modelos basados en la TRI, muy escasos aún en los manuales de prueba actuales, la información básica para evaluar la consistencia de los resultados de la medida es la Función de Información (FI) para los Ítems (Función de Información del Ítem, FII) o para la prueba total (Función de información del Test, FIT). La TRI no asume independencia del error de medida y la magnitud de atributo de manera que el error se expresa como una función de esta última y la función de información es la inversa del error. Generalmente no se reporta de manera individual la FII para cada uno de los ítems; puesto que el objetivo es sustentar el nivel de precisión de la prueba total, resulta más útil la FIT que se obtiene a partir de la suma de las FII para los ítems que la componen (Baker, 2001). La figura 2 es una representación típica del error de medida y la FIT para una prueba.

Además de la identificación de la o las fuentes de error relevantes para la medida específica, en la revisión de los estudios de confiabilidad para valoración de la calidad técnica

de las pruebas, el evaluador debe considerar tres aspectos: las características de los individuos que participan, el tamaño del grupo y las condiciones de recolección de los datos. En cuanto a las características de los individuos el aspecto más importante es la similitud con la población a la que está dirigida al menos en las variables que se consideren importantes o relacionadas con el atributo que mide la prueba; en cuanto al tamaño de los grupos que participan, se espera que la muestra de respuestas permita el cumplimiento de los supuestos estadísticos para que los resultados resulten extrapolables, y en cuanto a las condiciones de recolección se espera que puedan asimilarse con los procedimientos de aplicación indicados en el manual, de manera que las condiciones de recolección sean coherentes con el procedimiento de aplicación propuesto.

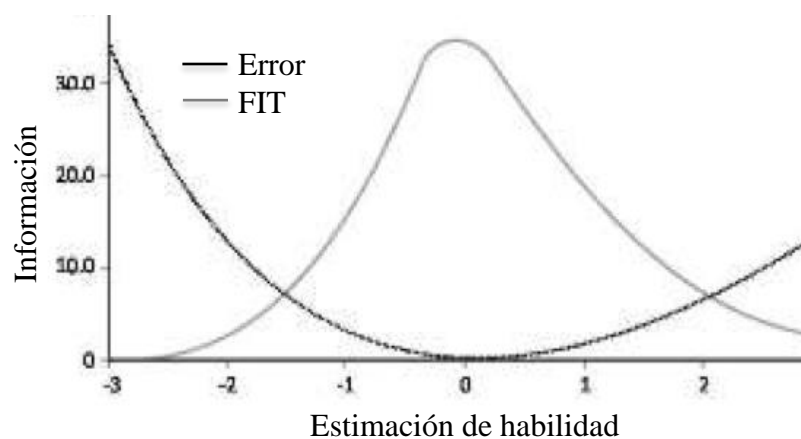


Figura 2 Representación del error de medida y la función de información del test (FIT) de una prueba analizada con modelo TRI.

En síntesis, la confiabilidad es un aspecto central en la evaluación de la calidad técnica de una prueba y la información clara y completa sobre los estudios de confiabilidad en los manuales técnicos facilita y orienta al usuario sobre la pertinencia del uso de la misma. Además, la consideración de la precisión de la medida en la interpretación del resultado según el margen de error posible es central en la interpretación del resultado para un examinado o grupo de examinados. En consecuencia, los manuales de prueba deben incluir información de referencia sobre estudios de confiabilidad, la fuente o fuentes de error relevantes, el diseño implementado por los desarrolladores para recoger los datos, la metodología de análisis para estimar la precisión de los puntajes, la manera como se estimaron el error y la confiabilidad y

los valores obtenidos. La tabla 7 presenta algunos procedimientos de análisis de los datos y el posible estimador de acuerdo con modelos teóricos de análisis (APA, 2014; Aiken, 2003; Anastasi & Urbina, 1998; Cohen & Swerdlik, 2000)

Tabla 7

Nociones de confiabilidad y métodos de análisis en los tres modelos de análisis.

Modelo teórico	Concepto de confiabilidad	Método de análisis	Diseño de recogida de datos	Estimador
Teoría clásica de los test	Precisión de la medida asociada a la variabilidad debida a error aleatorios	Modelo lineal	Test- retest	Error estándar de medida.
		Correlaciones	Formas paralelas	Índice de confiabilidad
		Modelo de regresión	Consistencia interna Consistencia entre observadores	Coefficiente de confiabilidad.
Teoría de la generalizabilidad	Grado de generalización de una medida	Análisis de varianza	Ensayos en diferentes condiciones	Coefficiente de generalizabilidad
Teoría de respuesta al ítem	Amplitud del rango de la función de información del ítem y de la prueba	Modelos probabilísticos Función de máxima verosimilitud	Aplicación piloto a evaluados	Función de información de ítem o prueba

Validez

El otro aspecto métrico imprescindible para juzgar la calidad de una prueba y que debe sustentarse como fundamentación estadística o técnica en los manuales, es la validez. Los estudios sobre la validez de los resultados son un aspecto imprescindible de la calidad técnica de la prueba, ya que si bien la confiabilidad es necesaria; el uso de los resultados y la interpretación del puntaje, solo tienen sentido si hay evidencias empíricas que permitan considerar que las inferencias a partir de éstos son correctas.

El concepto de validez ha cambiado mucho a lo largo del tiempo, ha sido objeto de múltiples discusiones y aún hoy no hay completo acuerdo sobre el mismo, entre académicos, entre las organizaciones que se encargan de la evaluación de la calidad de las pruebas ni entre

los diversos sistemas de estándares. Para este trabajo se entiende la validez como la definen los estándares para la evaluación educativa y psicológica: “*el grado en cual las evidencias y la teoría soportan las interpretaciones de los puntajes de la prueba*” (traducción propia).

Validity refers to the degree to which evidence and theory support the interpretations of the test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests (AERA, APA, NCME, 2014. P. 11)

La validez es la característica de la medida que se relaciona con el objeto medido, en las pruebas psicológicas con el constructo sobre el cual se pretende tener una estimación de magnitud; al igual que la confiabilidad es una cualidad abstracta, compleja de estudiar. De acuerdo con Muñiz; el meollo de la validez se relaciona con la respuesta a dos preguntas ¿cuáles son los argumentos y datos que sustentan las inferencias? y ¿cómo se recogen las evidencias teóricas y prácticas necesarias para afirmar que determinada inferencia es válida? (Muñiz, 2003); El desarrollo de prueba implica realizar los estudios para obtener evidencias que sustenten la interpretación propuesta de los resultados de la prueba.

Los resultados de los estudios de validez deben sustentar las inferencias propuestas a partir de los resultados de las pruebas y quienes se dedican a la adaptación de pruebas deben articular los procesos de validación, para obtener este tipo de evidencias de manera que se pueda considerar que son interpretables en las poblaciones para las cuales se realiza. (Borsboom, & Markus, 2013; Kane 2013 a, Kane 2013b). De acuerdo con la versión actual de los estándares de la AERA, APA, NCME (2014), las evidencias de validez se clasifican de acuerdo con la fuente de información por lo que una prueba con estudios psicométricos actualizados debería en primer lugar, justificar el tipo de evidencia relevante para la prueba según su objeto, uso propuesto o características de la población a la que va dirigida; y, además, incluir en el manual técnico, estudios que den cuenta de la recolección y análisis de información que se pueda clasificar en alguno de los siguientes cinco tipos de evidencias de validez:

- (1) basadas en el contenido de la prueba,

- (2) basada en procesos de respuesta,
- (3) basada en estructura interna,
- (4) basada en la relación con otras variables,
- (5) basadas en las consecuencias del uso de la prueba.

Aunque no está libre de discusiones y por el contrario, constituye un aspecto de amplio debate (Camargo, 2017), la concepción actual de validez considera el contexto y el uso de la prueba, tanto en la noción misma de validez como en los procesos de validación. Una consecuencia de adoptar esta postura es que en los estudios se validan las interpretaciones de los resultados para los usos propuestos, y por lo tanto para cada uso propuesto en el manual debería referenciarse resultados de por lo menos un estudio que sustente las interpretaciones.

Para que un estudio sustente la validez de una interpretación propuesta, la metodología implementada para obtener los datos y analizarlos debe ser coherente con el tipo de población y con el objetivo de uso, es decir los participantes en el estudio deben tener características relevantes en la manifestación del constructo, de manera se garantice que los resultados son extrapolables a la población a la que se dirige su uso. Por ejemplo el estudio de validez de resultados de una prueba para uso clínico que proponga un diagnóstico o clasificación en categorías, debería incluir el análisis de datos de poblaciones clínica y control donde se demuestre la especificidad y/o sensibilidad del diagnóstico.

Puesto que existen muchas posibles preguntas relevantes a responder alrededor de las interpretaciones de los puntajes de la prueba, la concepción actual de validez hace énfasis en que la validación es un proceso sin fin y los procesos de validación funcionan como procedimientos de prueba de hipótesis; por ello, admite la recolección de todo tipo de evidencias (Hubley & Zumbo, 2011). Además, la validez es una cuestión de grado; más y más variada evidencia puede juzgarse como mayor soporte de validez de las inferencias que se derivan de los usos de la prueba; sin embargo, esto crea un reto a la hora de hacer las evaluaciones de calidad puesto que genera la pregunta ampliamente formulada por los desarrolladores de prueba: *How much is necessary and how much is enough?*

Dado este amplio panorama en los posibles procedimientos de validación, los criterios para evaluar los estudios que den soporte de validez pueden ser los mismos que se utilizan para evaluar cualquier estudio que pretenda someter a prueba una hipótesis empíricas. Esta clasificación de evidencia de validez es aún reciente como para esperar que las pruebas disponibles hoy en el mercado, la hayan recogido en sus manuales y los instrumentos de evaluación de la calidad de las pruebas las hayan incluido; sin embargo, lo que parece ser determinante a la hora de evaluar la calidad técnica de la prueba en lo relacionado con los estudios de validez es la coherencia y la relevancia de los resultados de los estudios para soportar la solidez conceptual y empírica de la interpretación propuesta a cada categoría o puntaje de la escala. Esto requiere una adecuada fundamentación metodológica y una mirada crítica que permita identificar los posibles diseños, requerimientos, alcances y limitaciones de los estudios para valorar la evidencias de validez de la interpretación propuesta por los autores, los propósitos de uso de la prueba, la población a la que se dirige y el constructo como un elemento central en el diseño de estos estudios. La tabla 8 presenta algunos de los tipos de estudios posibles y las metodologías de análisis de datos que se encuentran con frecuencia asociados a los estudios de validez desde esta perspectiva.

Los criterios mínimos más frecuentemente utilizados para la evaluación de la calidad de la pruebas en lo relacionado con la validez siguen siendo dos procedimientos muy clásicos: La valoración de la representatividad de los contenidos de la prueba y los elementos que la componen respecto del constructo evaluado, a través de la evaluación por jueces; y la identificación de la correspondencia entre las dimensiones definidas teóricamente y las agrupaciones empíricas observadas a partir de las respuestas de una muestra de examinados, a través de algún procedimiento de análisis factorial.

Tabla 8

Identificación de tipos de evidencia y objetivos de los estudios de validez

Tipo de evidencia	Objetivo	Estudios comunes	Metodologías
Contenido	Análisis de la relación entre contenido de la prueba y el constructo	Valoración jueces expertos. Observación sistemática	Análisis lógico o empírico de la representatividad del

Tipo de evidencia	Objetivo	Estudios comunes	Metodologías
			constructo en el los ítems de la prueba
Procesos de respuesta	Explicitar el proceso respuesta para verificar su relación con asunción que hace el constructo del mismo	Entrevista cognitivas-análisis de las respuestas documentación del proceso de respuesta y desempeño.	Comprobación de la relación entre el desempeño evaluado y el proceso de respuesta.
Estructura interna	El grado de relación entre los ítems que componen la prueba y las dimensiones del constructo	Análisis factorial, Funcionamiento diferencial de los ítems	Análisis de correlación y/o variación de conjuntos de respuesta
Relación con otras variables	Relación con un criterio externo que permita análisis lógico del constructo	Correlacionales, Comparación de grupos, Diseños experimentales, Análisis de funcionamiento diferencial	Comparación de similitud o divergencia de las respuestas respecto a constructos relacionados o sobre los cuales se formula una hipótesis de comprobación
Consecuencias del uso de la prueba	Relación con el uso de los resultados en los procesos de evaluación	No se han referenciado en manuales aún.	Dependerán de los usos como elemento de referencia.

En síntesis, confiabilidad y validez son los aspectos métricos de la prueba, centrales en la evaluación técnica de la misma para sustentar la precisión de la medida y la evidencia tanto teórica como empírica que soporta la interpretación de sus resultados. Sin embargo, cuando las pruebas están compuestas por elementos, ítems o tareas, estas propiedades métricas dependen en buena parte, o son una configuración de las propiedades de los elementos que las componen. La evaluación de las características de los elementos de prueba es lo que se conoce como análisis de ítems desde la TCT o como calibración de los ítems desde la TRI.

Análisis de ítems

En las pruebas desarrolladas bajo la TCT, lo mínimo esperado sobre el análisis de los ítems de prueba es alguna evidencia de su dificultad y de su relación con la prueba total. Desde esta perspectiva la dificultad del ítem hace referencia a la media de las puntuaciones en el mismo o la proporción de respuesta correcta en una muestra de examinados. En el proceso de desarrollo de la prueba, normalmente se adopta algún criterio para la selección de los elementos que conforman la prueba definitiva, según los objetivos de la misma. Aunque con frecuencia se privilegian ítems que muestren dificultad media, por ejemplo proporciones de acierto entre 0,3 y 0,7 para ítems de calificación dicótoma, este criterio puede cambiar dependiendo de las características deseables en la prueba. En cualquier caso el manual técnico debería reportar los criterios adoptados y su coherencia con los objetivos de la prueba.

Además de la dificultad, una propiedad importante en los ítems desde la TCT, es su relación con la prueba total, conocida como discriminación y evaluada normalmente a través de algunas medidas de asociación entre las puntuaciones en el ítem y el puntaje total en la prueba. Las estimaciones más frecuentemente utilizadas son la diferencia de dificultad del ítem para grupos de alto y bajo rendimiento en la prueba (Prieto & Delgado, 2003), la correlación biserial puntual (Linacre, 2008) y el Brodgen-Clemens (Brodgen, 1949) cuando se trata de ítems de calificación dicótoma. Nuevamente, se espera que la pertinencia de estas medidas y los criterios para la decisión sobre la composición de la prueba, estén explícitamente reportados en los manuales técnicos de prueba.

Cuando se adopta un modelo de la TRI para el desarrollo de la prueba, se espera que los manuales de prueba reporten como mínimo, además de la FIT revisada en el aparte de confiabilidad, los criterios para la selección del modelo específico, la verificación de los supuestos del modelo elegido y los resultados de la calibración de los ítems con el ajuste del modelo. En general los modelos de la TRI estiman la probabilidad de acertar en un ítem o de puntuar en una categoría de respuesta, en función de la magnitud de atributo del evaluado considerando los parámetros (dificulta, discriminación y pseudoazar) del ítem. La representación gráfica la probabilidad de acierto en función de la magnitud de atributo se conoce como la Curva Característica del Item (CCI) como se ilustra en la figura 3. En algunas

ocasiones resulta útil reportar además, la Curva Característica del Test (CCT) que representa la puntuación esperada en la prueba total, en función de la magnitud de atributo.

Los diferentes modelos de la TRI dependen de tres criterios: la escala de calificación de las preguntas, el supuesto sobre la dimensionalidad del atributo medido por la prueba y el número de parámetros del ítem necesarios para explicar la probabilidad de acierto en el mismo. De acuerdo con el primer criterio los modelos pueden ser de ítem dicótomos, de crédito parcial, de respuesta graduada o continuos; según la dimensionalidad del atributo los modelos pueden ser unidimensionales o multidimensionales y según el número de parámetros de los ítems, pueden ser de uno, dos o tres parámetros.

Aunque buena parte de las pruebas de uso frecuente hoy sustentan sus propiedades métricas con base en supuestos y procedimientos de la TCT, dentro de los que usan TRI, el modelo más frecuentemente utilizado es el modelo de Rasch (Rasch, 1960). Según este modelo la probabilidad de acierto en un ítem puede estimarse de manera muy sencilla en términos de la diferencia entre la habilidad del examinado y la dificultad del ítem expresada en término de magnitud de atributo (Tristán López, 1998; 2001). Dado que la habilidad del examinado y la dificultad del ítem se expresan en la misma métrica, de manera muy intuitiva puede entenderse que, a medida que la habilidad del examinado sea mayor que la que el ítem requiere para ser respondido, la probabilidad de acierto es mayor. La figura 3 representa las CCI de tres ítems siguiendo un modelo de Rasch.

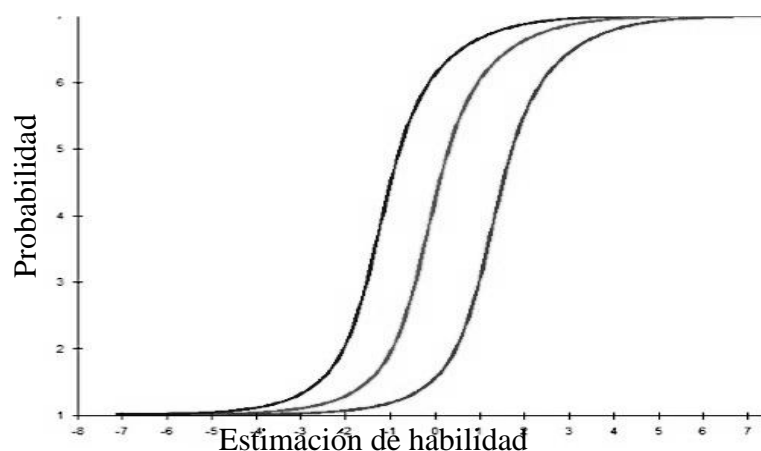


Figura 3. Representación de las CCI para tres ítems ajustando un modelo de Rasch

Además de los criterios ya mencionados, en la valoración de los estudios de análisis de ítems basados en la TRI para sustentar la calidad técnica de una prueba, resulta de gran importancia el tamaño y la calidad de la muestra de estudio. En general, una adecuada calibración de los ítems con modelos de la TRI requieren tamaños de muestra grandes; además, aunque matemáticamente puede demostrarse la invarianza de las estimaciones de los parámetros de los modelos TRI, la utilidad de los resultados de la calibración de los ítems depende en buena parte de la calidad de la muestra en términos de la información que logre recoger sobre las variables importantes de la población a la que se dirige la prueba.

Estandarización de la prueba

Un tercer aspecto que se valora en la revisión técnica es la estandarización o información disponible para la aplicación, calificación e interpretación de los resultados de la prueba. La estandarización de la prueba es un aspecto fundamental de su calidad técnica puesto que los procedimientos de aplicación y calificación son parte del proceso de medición. En consecuencia, en las pruebas estandarizadas es indispensable seguir los procedimientos al pie de la letra para garantizar la comparabilidad del desempeño de los evaluados con la escala que propone la prueba.

Los manuales suelen incluir un apartado sobre los procedimientos de aplicación en el cual se explican las condiciones en las cuales debe realizarse la aplicación y se dan instrucciones específicas para que el aplicador desarrolle el paso a paso, controle las variables o condiciones que pueden interferir con el resultado y obtenga las calificaciones tanto en puntaje bruto como en las transformaciones del mismo. Los estándares o protocolos de evaluación de calidad de las pruebas suelen valorar la suficiencia, utilidad y claridad de esta información.

Respecto a la claridad se espera que las instrucciones para la asignación de puntajes resulten comprensibles y orienten al profesional en casos en los cuales se pueden generar ambivalencias en la interpretación de la respuestas; en cuanto a la suficiencia se espera que se presenten el paso a paso de la calificación de manera que cada uno de los aspectos del constructo puedan ser calificados; finalmente, la pertinencia hace referencia a la coherencia de

la información en relación a la interpretación de las respuestas con los referentes conceptuales que fundamentan el constructo y los usos propuestos para la prueba.

En cuanto al proceso de calificación se espera que el manual incluya instrucciones detalladas, para obtener el resultado de acuerdo con la escala que propone la prueba para estimar la magnitud del atributo; si por ejemplo la prueba propone categorías diagnósticas se espera que se reporte claramente los estudios que sustentan los puntos de corte y explique claramente qué indicadores se afilian a cada a clasificación de manera que la interpretación del resultado pueda sustentarse desde el marco teórico y conceptual de referencia. Desde la TCT, los procedimientos más frecuentes para establecer los puntos de corte se basan en la comparación de grupos (clínico y control, por ejemplo) que permita evaluar la capacidad de la prueba para clasificarlos correctamente.

Lo más frecuente es el uso de tablas de conversión de puntajes es escalas estándar que permitan comparar los resultados entre diferentes pruebas o diferentes poblaciones; es decir, el uso de normas de calificación que faciliten la interpretación en términos de comparación con una muestra normativa. Las escalas más utilizadas son las transformaciones conocidas como Z o T , con medias de 100 y 50 , y desviaciones estándar de 10 y 15 , y las escalas de percentiles (Sarriá, 1999). Sea cual sea el procedimiento de calificación, la información sobre el mismo debe incluir las explicaciones que permitan identificar la relación de la conducta seleccionada como indicador y el sistema de puntuación de la escala; de manera que el resultado numérico o categórico sea una estimación coherente con el modelo teórico que sustenta la prueba. El manual debe brindar al profesional la explicación para entender el proceso de calificación y la asignación del puntaje a partir de unas reglas estructuradas del sistema de medida del constructo.

Un aspecto fundamental para valorar la calidad de los estudios que soportan la elaboración de los baremos o tablas de conversión, es la calidad y el tamaño de la muestra empleada. Ya que la validez y utilidad de las interpretaciones de los resultados dependen de la comparación del desempeño del examinado con la muestra normativa, la procedencia de tal muestra constituye un criterio muy importante para evaluar la calidad técnica de la prueba. El manual técnico debe incluir la composición de la o las muestras en términos de su procedencia,

variables sociodemográficas y demás información que se considere relevante según las características de la prueba, y de los subgrupos identificados en caso de existir variaciones significativas en el comportamiento de los resultados según alguna o algunas variables.

Además de la claridad y la suficiencia de la información para describir las muestras y los procedimientos, esta información debe poderse ubicar en el tiempo para que el lector decida sobre la vigencia de la misma respecto al proceso de evaluación en el que utiliza la prueba. Desde que se describió el efecto Flynn (Flynn, 1984, 2008) se cuenta con evidencia empírica que llama la atención sobre la importancia de la actualización de las tablas de calificación y los posibles cambios en las interpretaciones y decisiones que se tomen con base en los resultados de la pruebas (Rossi-Casé, Neer, y Lopetegui, 2001; Rusell, 2007).

En síntesis, la estandarización de la prueba hacer referencia tanto a las condiciones de aplicación, calificación e interpretación como a la construcción de baremos, tablas de conversión o categorías en las que pueden expresarse los resultados de los examinados. La calidad técnica de la prueba depende en parte de la claridad, suficiencia y pertinencia de la información reportada en el manual para calificar el desempeño de los examinados y obtener los resultados. Resultan de especial importancia las características de la o las muestras normativas para la construcción de los criterios de calificación, la identificación de posibles subgrupos y la actualización de los estudios para levantar los baremos.

Método

Para el cumplimiento del objetivo principal de este trabajo se adaptó y se complementó la propuesta metodológica general del proyecto original de Herrera (2009) en tres fases, que se ejecutaron entre 2012 y diciembre de 2016. Cada una de las fases tenía como propósito general alcanzar uno de los objetivos específicos de la investigación. En este capítulo se describen con detalle el procedimiento adoptado y las actividades que cubrió cada una de las fases.

Fase 1: Identificación de las pruebas más usadas en Colombia

El objetivo de esta fase fue lograr un panorama general del estado actual del uso de pruebas psicológicas en el país, identificando las más frecuentemente reportadas por los psicólogos profesionales en Colombia y algunas prácticas de uso de los manuales técnicos. El resultado obtenido fue un listado de pruebas a partir del cual se seleccionaron las que se evaluaron en la tercera fase del estudio, y una información general sobre algunas prácticas de los psicólogos en el uso de las mismas.

Población y participantes.

La población para esta fase del estudio fueron los psicólogos colombianos en ejercicio profesional entre 2013 y 2014. Participaron 1226 psicólogos entre enero de 2013 y septiembre de 2014, quienes aceptaron la invitación formulada a través del Colegio Colombiano de Psicólogos (COLPSIC, en adelante) para diligenciar la encuesta en línea. Esta permaneció disponible en la página web www.colpsic.org.co durante veinte meses.

Respondieron 938 (76%) mujeres 275 (22%) hombres, el porcentaje restante de los no reportaron información sobre su género. En el momento en que respondieron la encuesta, los participantes se encontraban en ejercicio profesional en 114 municipios o ciudades colombianas. Participaron psicólogos graduados desde 1968, sin embargo, 78% tenía diez años o menos de ejercicio profesional.

Instrumento.

Se diseñó una encuesta de preguntas abiertas que permitía a los profesionales reportar hasta cinco pruebas usadas en su ejercicio cotidiano. Se solicitó específicamente que informara cuáles pruebas utilizaba con mayor frecuencia en sus actividades cotidianas, y que respondiera seis 6 preguntas relacionadas con el manejo de los manuales técnicos de cada una de ellas. Estas preguntas fueron:

1. ¿Tiene acceso al manual técnico de la prueba?
2. ¿Consulta el manual técnico de la prueba para su aplicación y calificación?
3. ¿Aplica el protocolo indicado por el manual en la aplicación de la prueba?
4. ¿Conoce información sobre la validez de la prueba?
5. ¿Conoce información sobre la confiabilidad de la prueba?
6. ¿Conoce información sobre el uso de la prueba en Colombia?

La encuesta tuvo dos versiones, la primera versión estuvo disponible desde septiembre de 2012 hasta febrero de 2013, momento en el cual se cambió por una versión con un entorno gráfico más dinámico, que estuvo disponible para diligenciamiento hasta diciembre de 2014. Ver apéndice A

Procedimiento.

La encuesta fue diseñada por la autora del este trabajo de acuerdo con los objetivos de esta fase del estudio, fue revisada y corregida por la directora del proyecto marco (Herrera, 2009), y la presidencia del Colegio Colombiano de Psicólogos aprobó su publicación y divulgación en la página web institucional. El profesional administrador del sitio web de COLPSIC fue el responsable del diseño gráfico y la publicación.

Por medio de correos electrónicos y publicidad en la página se invitó masivamente a los psicólogos colegiados de todo el país a participar en la consulta. Durante la duración de esa fase del estudio se repitió la invitación en 5 oportunidades y al finalizar se consolidaron las respuestas de los participantes en una base de datos con 3717 registros de 1226 personas.

La información de reporte de uso de pruebas se registró mediante texto libre, por lo que el procesamiento de las respuestas de los participantes requirió verificación individual de los nombres; base que después se depuró y codificó, para la obtención de un listado de

aproximadamente 500 nombres diferentes de pruebas. Los datos fueron analizados mediante estadísticos descriptivos con el paquete SPSS Versión 19.

Fase 2 Desarrollo del instrumento de evaluación de calidad técnica de las pruebas

El objetivo de esta fase diseñar y construir un instrumento para evaluar la calidad técnica de las pruebas siguiendo estándares internacionales, que recogieran las principales avances y discusiones actuales sobre el desarrollo y uso de pruebas. Este instrumento debería permitir la recolección de la información sobre las características técnicas de las pruebas que se seleccionaron para responder al objetivo general de este trabajo.

Participantes.

Se conformaron dos grupos de participantes, uno de autores del instrumento y otro de jueces expertos para su validación. El primero estuvo conformados por 23 psicólogos profesionales con experiencia en evaluación, docentes e investigadores quienes al momento del desarrollo de las actividades se encontraban vinculados a 12 universidades de las diferentes regiones del país, y al ICFES.

El grupo de jueces expertos estuvo conformado por cinco académicos: 3 profesionales de amplio reconocimiento nacional y trayectoria en medición y evaluación, y dos invitados internacionales líderes de los procesos de evaluación de calidad de las pruebas, en sus respectivos países. Este grupo de expertos hizo la revisión del instrumento y sugirió modificaciones que fueron consideradas por el grupo de autores.

Procedimiento.

Para la construcción del instrumento de valoración de calidad técnica de pruebas se adelantaron las siguientes actividades:

1. Conformación de grupo de expertos
2. Capacitación específica
3. Diseño del instrumento y desarrollo de instrumento
4. Revisión de jueces expertos.
5. Consolidación versión definitiva

Conformación de grupo de autores

En primer lugar se definieron los perfiles de los autores del instrumento para lo cual se acogieron los criterios propuestos en un documento preliminar que de la extinta División de Evaluación y Estadística Aplicada de COLPSIC (Herrera, 2008) para la vinculación de psicólogos a la división. Con tales criterios se consultó la información sobre profesionales con experiencia en evaluación y psicometría, disponibles en la página del Departamento Administrativo de Ciencia, Tecnología e Innovación, COLCIENCIAS y se solicitaron nombres a las directivas de la Asociación Colombiana de Facultades de Psicología, ASCOFAPSI, y de COLPSIC. En la tabla 9 se muestran los diferentes perfiles.

Tabla 9

Perfiles de los participantes en el proceso de desarrollo del instrumento

Categoría	Experiencia y formación académica
Experto	Cinco años o más años de experiencia profesional en diseño, construcción, traducción o adaptación, estandarización, selección, aplicación e interpretación de pruebas psicológicas y con experiencia docente en asignaturas relacionadas con medición a nivel de pregrado en Psicología
Especialista	Formación avanzada a nivel de especialización en áreas como matemáticas o estadística en cualquiera de sus especialidades, análisis de datos, psicometría, medición, evaluación o similares. La formación avanzada puede sustituirse por otro título de pregrado en áreas como matemáticas, estadísticas o similares.
Investigador	Psicólogo con formación avanzada a nivel de maestría o doctorado en psicología cuantitativa, matemáticas o estadística, métodos psicométricos o estadísticos, métodos de investigación en psicología o ciencias del comportamiento o similares.

Fuente: Tomado de (Herrera, 2008). Documento inédito de la presidencia de la División de Medición, Evaluación y Estadística Aplicada de Colpsic.

Los profesionales fueron contactados e invitados a dos sesiones de capacitación en las que se presentó el proyecto y se les extendió la invitación a participar en el mismo. Quienes aceptaron la invitación se vincularon al proyecto general mediante un acuerdo de cooperación y confidencialidad firmado por todo el equipo de investigación de la segunda fase del estudio; los autores del instrumento se vincularon como co-investigadores es esta fase del proyecto. En

el apéndice B se encuentra la lista de participantes en la contraportada del instructivo. Participaron profesionales de Bogotá, Medellín, Popayán, Barranquilla, Bucaramanga y Tunja

Capacitación específica

Se convocó a los autores y otros profesionales del área a dos actividades de capacitación específica, con invitados internacionales, estas actividades de alto nivel académico fueron patrocinadas por la Universidad Nacional y se consideran una retribución a los profesionales que participaron en el proyecto.

La primera actividad de capacitación fue un *Curso taller: "Introduction IRT: the Rasch model for measurement"*, con el profesor Kendon J. Conrad, PhD, de la Universidad de Illinois experto investigador del departamento de políticas de salud. Tuvo una duración de cuatro días, el taller abordó contenidos relacionados con la construcción de pruebas con modelo análisis TRI y la metodología del mapeo conceptual para la construcción de pruebas; durante esta actividad se inició la gestión para conformar el grupo de expertos.

La segunda sesión de capacitación fue el curso *"Consideraciones sobre la calidad de las pruebas para el desarrollo de los procesos de evaluación de pruebas psicológicas"* con la profesora Ana Paula Porto Noronha Ph.D. de la Universidad São Francisco (USF). Campus Itatiba. Durante 5 días se trabajó conceptos técnicos de calidad de pruebas y el modelo de evaluación brasileño. Durante esta actividad se consolidó el grupo de autores y se inició la programación de actividades para la construcción del instrumento.

Diseño y desarrollo del instrumento.

El instrumento se diseñó y construyó en dos años y medio de trabajo entre 2012 y 2015, durante los cuales se hicieron trece (13) sesiones presenciales de una duración de al menos, 5 horas de trabajo conjunto de trabajo y trabajo individual o en subgrupos. El desarrollo del instrumento cubrió dos etapas. En la primera, que tuvo una duración de 3 sesiones, se presentó un panorama general de las estructuras y contenidos de los protocolos desarrollados en otros países, se definió la estructura general del instrumento y se conformaron 4 mesas de trabajo para desarrollar los apartados técnicos. Esta etapa finalizó con un simposio en el cual las

mesas presentaron la propuesta inicial de los cuatro apartados que conformarían el instrumento: referentes conceptuales, confiabilidad, validez y estandarización

A partir de la cuarta sesión de trabajo y durante los dos años siguientes, las mesas de trabajo adelantaron revisiones teóricas y metodológicas para definir los indicadores, su distribución en los apartados y los criterios de calidad técnica de una prueba en cada una en uno de los apartados previamente acordados. Esta fase requirió sesiones de trabajo separadas para cada mesa de trabajo algunas de las cuales se hicieron virtuales y sesiones de trabajo conjuntas del grupo completo. Estas últimas se realizaron en las instalaciones de COLPSIC y en ellas se discutieron y revisaron indicadores y descriptores integrando la versión propuesta por la mesa de trabajo con los aportes de los miembros de las otras mesas. Después de año y medio de trabajo se llegó a la primera versión.

Revisión por expertos

Esta primera versión tuvo una revisión de forma y diseño gráfico preliminar, y se pasó a revisión por parte del grupo de expertos a quienes se les solicitó una mirada reflexiva y crítica; este proceso duró dos días en jornada continua. Los expertos hicieron observaciones sobre los descriptores, sobre los criterios y sobre el instructivo para el uso del instrumento; información que se dio a conocer a los autores en una sesión plenaria en la cual los autores tuvieron la oportunidad de discutir y justificar sus puntos de vista respecto a las sugerencias de los jueces expertos. Cada mesa de trabajo hizo los ajustes que consideró necesarios según los comentarios de la revisión y se consolidó la versión nueva del instrumento.

Consolidación de la versión final del instrumento.

Después de conocidos y discutidas las observaciones y recomendaciones del grupo de expertos, el grupo de expertos introdujo algunos cambios a la versión original del instrumento. Estos se realizaron durante 2 sesiones de trabajo conjunto en las que se revisaron y aprobaron los cambios por consenso.

Simultáneamente con las sesiones de desarrollo y revisión del instrumento se elaboró el instructivo para su uso en el que se consolidan las instrucciones para los evaluadores de las pruebas y se recogieron las observaciones de los autores y expertos. Ver apéndice B

Fase 3 Evaluación piloto de seis de las pruebas más usadas por psicólogos colombianos

El objetivo de esta fase fue hacer un diagnóstico inicial de la calidad de las pruebas usadas en el país, a partir de la evaluación piloto de seis pruebas de las más usadas en el país utilizando el instrumento desarrollado.

Participantes

Para esta fase se conformó un grupo de evaluadores compuesto por profesionales en ejercicio en diferentes áreas aplicadas de la psicología y algunos con experiencia en medición y evaluación. Además de la autora de este trabajo, se convocó a psicólogos con amplia experiencia en evaluación y psicometría, y profesionales con experiencia en las áreas de aplicación de las pruebas seleccionadas: clínica, educativa y organizacional.

Instrumentos.

En esta fase los profesionales usaron el instructivo y el instrumento de valoración técnica desarrollado en la fase dos y las pruebas seleccionadas para la evaluación. Ver tabla 10. Las pruebas seleccionadas para el ejercicio de evaluación de calidad técnica fueron las primeras de la lista siempre y cuando fueran objetivas puesto que el instrumento no está diseñado para pruebas proyectivas. Sin embargo, para efectos de este proceso no se evaluó el 16 PF en su versión original que fue la prueba más frecuentemente reportada porque, de acuerdo con TEA, distribuidor legal de la prueba, desde el año 2010 dejó de distribuirse por su obsolescencia y la versión vigente es el 16PF5. En la tabla 10 se presentan las pruebas más frecuentemente reportadas en la primera fase del estudio.

Tabla 10**Pruebas seleccionadas su evaluación piloto de calidad técnica**

Prueba	Edición	Autor	Año	ISBN
Escala Wechsler de Inteligencia para Niños IV WISC IV	Traducción al español para México	David Wechsler	2007	970-729-261-X
Inventario Multifásico de la Personalidad Minnesota MMPI- 2	Traducción al español para México	S.R. Hathaway y J.C.Mc Kinley	1995	978-950-12-303-4
Inventario de Personalidad para vendedores IPV	Adaptación española Nicolás Seisdedos y Agustín Cordero	Les editions du centre de psychologie appliquee	1996	84-7174-813-4
Cuestionario factorial de la personalidad 16 pf 5	10ª edición, revisada y ampliada	Raymond, Karen & Heather Cattell	1995	978-84-15262-86-2
Inventario Millon's de Estilos de personalidad MIPS	Traducción Adolfo Negrotto, adaptación y supervisión María Martina Casullo, Alicia Cayssials	Teodore Millon	1996	978-950-12-1303-4
Inventario de Depresión Beck BDI-II	Adaptación española	Aaron T .Beck, Robert A Steer, Gregory K Brown	1996	978-84-939315-1-3

Procedimiento

En primer momento se seleccionaron las pruebas que serían evaluadas a partir de la lista de las más reportadas por los profesionales que participaron en la primera fase del estudio, y se solicitó su préstamo a las casas comercializadoras dueñas de los derechos de distribución en el país. Simultáneamente se convocó al grupo de evaluadores con experiencia en el área de uso de la prueba y experiencia en evaluación o psicometría, quienes recibieron una inducción en el uso del instrumento y posteriormente revisaron los manuales técnicos de las pruebas seleccionadas y las evaluaron haciendo uso del mismo.

El grupo de evaluadores trabajó durante dos sesiones de trabajo presencial de cuatro horas cada una. Los evaluadores recibieron información sobre los antecedentes del proyecto y las

instrucciones para el manejo del material para la evaluación: el instructivo, el instrumento y un ejemplar de prueba que debían evaluar. Las evaluaciones se asignaron de manera que cada prueba tuviera al menos dos revisiones, una de un profesional experto en el área aplicada de uso predominante de la prueba, y una de un experto en evaluación o psicometría. Los evaluadores trabajaron individualmente; revisaron los manuales técnicos de las pruebas y diligenciaron parte del instrumento de evaluación desarrollado en la fase dos. Por motivos de tiempo y la confidencialidad del instrumento no se permitió retirar el material de las sesiones de trabajo. La extensión de los manuales y la falta de familiaridad con el instrumento impidieron que los evaluadores terminaran completamente las evaluaciones, por lo cual la autora de este trabajo completó las evaluaciones considerando las evaluaciones y los comentarios de estos profesionales.

Los resultados de las evaluaciones se procesaron mediante distribuciones de frecuencias del nivel en que quedó ubicada cada prueba en cada uno de los indicadores de calidad del instrumento.

Resultados

Los resultados del trabajo se presentarán separadamente para cada una de las fases presentadas previamente.

Identificación de las pruebas más usadas por los psicólogos colombianos

Una vez terminada la primera aplicación de la encuesta en línea se encontró que de los 1226 participantes, el 98% reportó información sobre el área principal de ejercicio profesional, el 60% informó la ciudad donde ejerce y el 50% reportó tener formación de posgrado. La tabla 11 muestra los resultados de la distribución de la muestra de participantes; sin embargo, es necesario aclarar que aunque los participantes ejercen su profesión en 113 municipios colombianos diferentes, en la tabla solo se reportan las ciudades en las cuales había al menos 10 participantes.

Tabla 11

Reporte formación de posgrado de los participantes en la consulta de uso de pruebas.

Variable	Áreas de ejercicio	Frecuencia	Porcentaje
Área de ejercicio profesional	Clínica	402	32,8
	Organizacional	386	31,5
	Educativa	177	14,4
	Docencia Universitaria	64	5,2
	Social	49	4,0
	Jurídica o Forense	42	3,4
	Investigación	21	1,7
	Militar	11	0,9
	De la Salud	10	0,8
	Neuropsicología	7	0,6
Formación de posgrado	Doctorado	18	1,5
	Especialización	390	31,8
	Maestría	227	18,5
Ciudad de ejercicio profesional	Bogotá D.C.	248	20,23
	Medellín	68	5,55
	Barranquilla	36	2,94

Variable	Áreas de ejercicio	Frecuencia	Porcentaje
	Cali	34	2,77
	Bucaramanga	31	2,53
	Pereira	26	2,12
	Neiva	16	1,31
	Ibagué	14	1,14
	Armenia	13	1,06
	Cartagena	13	1,06
	San Juan de Pasto	13	1,06

Identificación de las pruebas más usadas

La encuesta permitió a los participantes reportar el uso de hasta cinco (5) pruebas y una vez terminado el tiempo para diligenciarla, se encontraron 3717 registros de las respuestas de 1226 profesionales, de 680 pruebas diferentes, datos que fueron revisados y depurados para identificar posibles inconsistencias en la información o errores de digitación. Se excluyeron respuestas que no se referían a pruebas sino a técnicas de observación o respuestas que no especificaban el nombre de una prueba, en total se procesaron 3626 reportes. La tabla 12 presenta el listado de pruebas que tuvieron hasta 20 reportes en la consulta; con base en estos resultados se seleccionaron las seis que se evaluaron.

Tabla 12

Listado de las pruebas con los mayores reportes de uso

Prueba	Reportes	Tipo
Wartegg	562	Proyectivo
16 PF	422	Objetivo
Valanti	176	Proyectivo
Test de Machover	156	Proyectivo
Dibujo de la Familia	119	Proyectivo
Test de la Figura Humana	117	Proyectivo
IPV-Inventario de Personalidad para Vendedores	89	Objetivo
MMPI	69	Objetivo
WISC IV	66	Objetivo
16 PF-5	60	Objetivo
MMPI-2	51	Objetivo

Test Gestáltico Visomotor de Bender	49	Desempeño
CMT-Cuestionario de Motivación para el trabajo	48	Objetivo
WISC	39	Objetivo
Inventario de depresión de Beck	34	Objetivo
WAIS	33	Objetivo
Kuder	33	Objetivo
Evaluación Neuropsicológica Infantil (ENI)	30	Objetivo
H-T-P	26	Proyectivo
WISC R	26	Objetivo
Minimult	25	Objetivo
Eros	23	Objetivo
Inventario Clínico Multiaxial de Millon's (MCMI)	22	Objetivo
Persona bajo la lluvia	21	Proyectivo
DISC	20	Objetivo
Test Naipes	20	Objetivo

De acuerdo con este resultado, cinco de las seis pruebas más frecuentemente mencionadas son proyectivas y la más usada de las objetivas es el 16PF; sin embargo, teniendo en cuenta la disponibilidad para la distribución por parte de las casas comerciales, la versión original del 16 PF original no se evaluó por su obsolescencia. Excluyendo el 16PF, las seis pruebas objetivas más seleccionadas fueron: la Cuestionario Factorial de la Personalidad 16PF5; la Escala Wechsler de Inteligencia para Niños IV, WISC IV; el Inventario Multifásico de la Personalidad Minnesota MMPI-2; el Inventario de Personalidad para vendedores IPV; el Inventario Millon's de Estilos de Personalidad MIPS y el Inventario de Depresión BDI- II;

La figura 4 muestra los resultados de las preguntas adicionales que respondieron los participantes sobre el acceso a información técnica de las pruebas. Los resultados muestran que el 82% de los participantes reportó tener acceso al manual técnico de las pruebas que usa, el 91% se apoya en él para la aplicación, y solo el 6% lo usa en los procesos de calificación. Además, menos del 40% afirman conocer estudios de validez y confiabilidad sobre las pruebas que usa y menos del 20% reconoce la existencia de estudios con población colombiana.

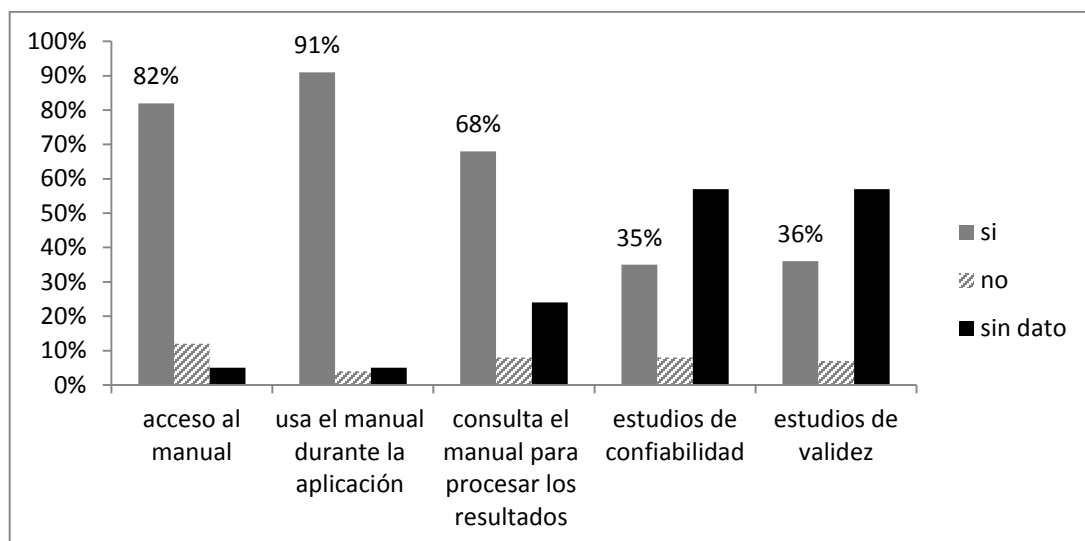


Figura 4. Distribución de las respuestas en la segunda parte de la encuesta

Desarrollo de instrumento para evaluar la calidad técnica de las pruebas

El instrumento definitivo quedó conformado por cinco apartes: Información de la prueba, referentes conceptuales, confiabilidad, validez y estandarización. El primer apartado es una ficha que recopila la información de identificación y descripción de la prueba en la cual el evaluador consigna los datos sobre la identificación de la prueba (nombre, versión, registro ISBN), datos de los autores de la versión evaluada y de versiones anteriores, la identificación del constructo que evalúa, la población a la cual está diseñada, el modelo psicométrico que la sustenta y las generalidades de la metodología de aplicación y calificación. Al final de este apartado el evaluador puede consignar sus consideraciones o comentarios sobre la prueba (ver apéndice B).

Los siguientes apartados valoran la calidad técnica de la prueba por medio de 45 indicadores; de los cuales 31 son obligatorios y 14 opcionales, los obligatorios valoran las características que por conceso de los autores del instrumento, se consideran fundamentales para el uso adecuado de una prueba. Cada indicador se valoran con una escala de cuatro categorías: no cumple, mínimo, aceptable y bueno, de manera que se para que una prueba muestre un nivel de calidad técnica aceptable, debe cumplir por lo menos en el nivel mínimo de todos los indicadores obligatorios.

Tabla 13**Estructura general del instrumento de evaluación de la calidad técnica de las pruebas para uso en Colombia**

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		5	2
3. Confiabilidad		6	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*	2*	5
5. Calificación y estandarización		7	2

Fuente: tomado de (Herrera y León, 2015) instructivo del instrumento para la valoración de calidades técnicas

Nota: *: indicadores que se evalúan solamente si la prueba está conformada por ítems o elementos.

Los 14 indicadores opcionales consideran cualidades de valor agregado que describen condiciones deseables en una prueba pero que en la actualidad no son muy frecuentes en los desarrollos de las mismas y por tanto no suelen estar incluidos en los manuales.

La evaluación de las características métricas de la prueba inicia en el segundo apartado, que valora los referentes conceptuales por medio de 5 indicadores obligatorios. Aquí se valora la información disponible en el manual sobre los referentes conceptuales y teóricos que fundamentan el constructo evaluado, y que representa el sustento de interpretación de la medida. La tabla 14 presenta los criterios que se evalúan sobre referente conceptual de la prueba.

Tabla 14**Indicadores evaluados en el apartado evaluación de los referentes conceptuales**

Criterios evaluados en el apartado referente conceptual de la prueba	
2.1	Definición conceptual del constructo a medir
2.2	Definición operacional del constructo a medir
2.3	Definición de las dimensiones del constructo y de las relaciones entre ellas
2.4	Descripción de los usos previstos para el instrumento
2.5	Sustento teórico para el uso de la prueba en la población(es) objeto.
2.6	Contextualización histórica del constructo*.
2.7	Estructuración teórica de conocimientos acerca del constructo medido*.

* Indicadores opcionales

Los cinco indicadores obligatorios identifican la rigurosidad con la cual se espera que el manual presente la información teórica sobre el constructo evaluado por la prueba, identificado y definiendo claramente el concepto, sus dimensiones y su operacionalización, así como su justificación de uso para la población. Los dos indicadores opcionales refieren información que permite identificar el constructo como parte de un desarrollo estructurado científicamente, donde se entiende la evolución del concepto desde una perspectiva amplia.

Tabla 15**Criterios de confiabilidad para evaluar la calidad de la técnica de la prueba**

Indicadores para la evaluación de la confiabilidad	
3.1	Identificación de la o las fuentes de error de medición
3.2	Descripción del estudio para estimar la confiabilidad
3.3	Estimación del error de medida
3.4	Tamaño de muestra de los estudios de confiabilidad
3.5	Valor del estadístico
3.6	Interpretación de resultados de confiabilidad
3.7	Modelo psicométrico para el estudio de confiabilidad*
3.8	Escogencia del método estadístico para estimar la confiabilidad*

* Indicadores opcionales

El tercer apartado del instrumento aborda la información sobre los estudios de confiabilidad, en 8 indicadores. Los 6 indicadores obligatorios valoran la pertinencia del

diseño y los resultados de los estudios empíricos que sustentan la precisión de la medida que ofrece la prueba. Además, los 2 opcionales hacen referencia a la justificación y la coherencia en la selección del diseño del estudio o estudio de confiabilidad y del estadístico para estimarla. La tabla 15 presenta los criterios que se evalúan en el apartado.

Tabla 16

Indicadores para evaluar la validez

Criterios de evaluación de las evidencias de validez		
Evidencias basadas en el contenido de la prueba	4.1	Estructura conceptual de la prueba
	4.2	Revisión por expertos idóneos
	4.3	Número de expertos y contenido de la consulta.
	4.4	Resultados de evaluación por expertos
	4.5	Capacitación a expertos*
	4.6	Tabla de especificaciones*
Evidencia de validez basada en la estructura interna de la prueba	4.7	Estructura interna de la prueba
	4.8	Tamaño de muestra
	4.9	Resultados del estudio
Evidencia de validez con base en la relación con otras variables	4.10	Estrategia de validación
	4.11	Selección de la(s) variable(s) criterio
	4.12	Tamaño de muestra
	4.13	Análisis estadístico empleado
	4.14	Resultados de los estadísticos
	4.15*	Funcionamiento diferencial del ítems
4.16*	Análisis de sesgo de la prueba	
Evidencia con base en el análisis de ítems	4.17	Modelo psicométrico utilizado
	4.18	Estimación de dificultad
	4.19*	Estadísticos para análisis de ítems
	4.20*	Análisis de las opciones de respuesta
	4.21*	Parámetros del modelo de análisis
	4.22*	Ajuste del modelo
4.23*	Función de información	

* Indicadores opcionales.

El cuarto apartado valora los estudios de validez que se reportan en el manual. Este apartado es el más complejo del instrumento e incluye 4 sub-apartados, que evalúan cuatro tipos de evidencias de validez: de contenido de la prueba, de relación con otras variables, estructura interna y el análisis de ítems. En este apartado se califican las características de los

estudios que brindan sustento a la interpretación de los resultados en el manual; consta de 12 indicadores obligatorios y 9 opcionales clasificados según los tipos de evidencia de acuerdo con la clasificación propuesta en la sexta versión de estándares de AERA, APA, NCME (2014). El último apartado sobre análisis de ítems solo se diligencia si la prueba evaluada se compone de ítems o tareas. La tabla 16 presenta los criterios que se valoran en este apartado.

El último apartado hace referencia a la calificación y estandarización; por medio de 9 indicadores que valoran la suficiencia de la información para la aplicación y calificación de la prueba y la calidad de los estudios que soportan la elaboración de los baremos de calificación. Un criterio muy discutido fue la procedencia de la muestra de tales estudios que finalmente quedaron descritas como “muestras nacionales”. La tabla 17 presenta los criterios para evaluar este aspecto técnico de la prueba.

Tabla 17

Indicadores de evaluación de estandarización de la prueba

Criterios para evaluar la información para aplicación y calificación de la prueba	
5.1	Descripción de material de prueba
5.2	Instrucciones para la aplicación
5.3	Escalas y sistema de transformación de las puntuaciones
5.4	Descripción de la población objetivo y la(s) muestra(s) normativas
5.5	Tamaño de la(s) muestra(s) normativa(s)
5.6	Descripción del ámbito de aplicación y la calidad de los datos en que se basan los baremos
5.7	Protocolo de interpretación
5.8	Condiciones especiales de aplicación*
5.9	Protocolo de presentación de resultados*

Nota * indicadores opcionales

El instrumento se diseñó para que un evaluador califique cada indicador después de revisar el manual técnico de la prueba y los materiales disponibles para los usuarios profesionales. Para diligenciarlo, el evaluador selecciona el mejor descriptor de las características de información, si considera que la prueba cumple con el criterio, asigna una puntuación numérica entre 3 y 5 siguiendo las instrucciones que se incluyen detalladamente en el

instructivo. Los indicadores opcionales solo se califican si el manual reporta información sobre el mismo y la calificación numérica está en el rango de 4.0 a 5.0. Dado que estos indicadores hacen referencia a aspectos que darían “un valor agregado” a la prueba en términos de su calidad técnica, la calificación numérica equivale, como mínimo, al máximo nivel de los indicadores obligatorios. La tabla 18 presenta la escala numérica para los tipos de indicadores.

Tabla 18

Categorías y escalas para evaluar la información disponible en el manual

Tipo de indicador	Categoría de calificación	Escala
Obligatorio	No cumple (NC)	
	Cumplimiento mínimo (M)	Entre 3.0 y 3.5
	Cumplimiento aceptable (A)	Entre 3.6 y 4.0
	Cumplimiento bueno (B)	Entre 4.0 y 5.0
Opcionales	Bueno (B)	Entre 4.0 y 4.5
	Sobresaliente	Entre 4.6 y 5.0

En general estas categorías se acordaron en consenso después de revisar la conveniencia del uso de diferentes palabras que podrían denotar la calidad técnica de la prueba. Por ejemplo, para el caso de la categoría “No cumple”, se consideraron y discutieron opciones como “Deficiente” o “Insuficiente”, pero teniendo en cuenta que se trata de un instrumento de carácter orientativo se dejó el término más descriptivo y menos valorativo. Las categorías se complementan con la puntuación numérica que es un estimativo donde se considera que el máximo posible es 5.0, los rangos de los valores que se definieron para las categorías siguen criterios convencionales muy utilizados en la valoración numérica. Para la primera se acordó dejar el rango de 0 a 2,9 generalmente entendido como “desaprobado” mientras que las demás se distribuyen entre 3 y 5. El instructivo del instrumento que es el documento de apoyo para la aplicación del mismo se presenta en el apéndice B.

Evaluación de calidad de seis de las pruebas más usadas por psicólogos colombianos

En este aparte se presentan los resultados generales de la valoración técnica de las pruebas seleccionadas, la evaluaciones completas se encuentran disponibles en el apéndice C

Evaluación de los referentes conceptuales

La tabla 19 presenta los resultados de la valoración de los indicadores sobre referentes conceptuales de las 6 pruebas. Aunque en diferente nivel de calidad, todos los manuales de las pruebas incluyen información sobre referentes conceptuales del constructo, sin embargo, la única que los cumple en un nivel alto es Escala Wechsler de Inteligencia para Niños IV WISC IV. El Inventario de Personalidad para Vendedores (IPV) y el Inventario de Depresión Beck BDI-II no alcanzan el nivel máximo en ninguno de los indicadores, si bien reportan información ésta no se desarrolla suficientemente.

Tabla 19

Resultados evaluación de apartado referentes conceptuales

Indicadores	WISC IV	16PF-5	BDI II	IPV	MMPI 2	MIPS
2.1 Definición conceptual del constructo a medir	B	M	A	M	M	B
2.2 Definición operacional del constructo a medir	B	B	M	M	B	A
2.3 Definición de las dimensiones del constructo y las relaciones entre ellas	B	B	M	M	B	B
2.4 Descripción de los usos previstos para el instrumento.	B	M	A	A	M	M
2.5 Sustento teórico para el uso de la prueba en la población(es) objeto.	B	A	A	M	A	A
2.6 Contextualización histórica del constructo*	S	-	-	-	B	S
2.7 Estructuración teórica de conocimientos acerca del constructo medido*	B	-	-	-	-	S

Categorías indicadores obligatorios: NC= no cumple, M=mínimo puntaje entre 3.0 y 3.4, A=Aceptable puntaje entre 3.5 y 3.9 B= Bueno puntaje entre 4.0 y 5.0.,

Categorías indicadores opcionales B= Bueno puntaje entre 4.0 y 4.5 y .S= Sobresaliente 4.6 y 5.0

La tabla 20, es un resumen de la evaluación de la calidad de los referentes conceptuales, evidencia el porcentaje de indicadores calificados en cada una de las categorías de calificación para cada una de las pruebas.

Tabla 20

Porcentajes de indicadores de referentes conceptuales en cada categoría de cumplimiento

Prueba	Indicadores obligatorios (5)				I. opcionales (2)	
	No cumple	Mínimo	Aceptable	Buena	Buena	Sobresaliente
WISC IV	-	-	-	<u>100%</u>	50%	50%
16PF5	-	40%	20%	40%	-	-
BDI II	-	40%	60%	-	-	-
IPV	-	80%	20%	-	-	-
MMPI 2	-	40%	20%	40%	50%	-
MIPS	-	20%	40%	40%	-	<u>100%</u>

Evaluación de los estudios de confiabilidad

La tabla 21 presenta los resultados de la evaluación de los criterios de confiabilidad de las pruebas, en este aspecto la única prueba que incluye información sobre identificación de fuentes de error es el WISC-IV, que se desarrolló cuando estaba vigente la quinta versión del estándar para evaluación psicológica y educativa, las 4 restantes no incluyen este tipo de información. Ninguna de las pruebas cuenta con estudios de confiabilidad sobre respuestas de población colombiana y todas obtuvieron la valoración mínima en este indicador que corresponde a tamaño de los grupos entre 300 y 500 participantes de habla castellana.

La tabla 22 presenta los porcentajes de indicadores de confiabilidad en cada una de las categorías de la escala de calificación.

En general respecto a la información que sustenta la confiabilidad de los resultados de la prueba se evidencia que el WISC_IV cumple con los indicadores obligatorias y opcionales en cuanto a la información de confiabilidad que se incluye en el manual, seguido por BDI II, que aun siendo un instrumento de revisión de síntomas o nosológico, sustenta con varios estudios la confiabilidad de los resultados. El peor evaluado es el IPV, que no cumple con la mitad de los indicadores obligatorios, seguido por el MIPS y el 16 PF-5.

Tabla 21

Resultados evaluación de los indicadores de confiabilidad de las pruebas evaluadas

Indicadores	WISC IV	16PF5	BDI II	IPV	MMPI 2	MIPS
-------------	---------	-------	--------	-----	--------	------

Indicadores	WISC IV	16PF5	BDI II	IPV	MMPI 2	MIPS
3.1 Identificación de la o las fuentes de error de medición.	B	NC	NC	NC	NC	NC
3.2 Descripción del estudio para estimar la confiabilidad	B	B	M	M	M	B
3.3 Estimación del error de medida	B	NC	M	M	M	NC
3.4 Tamaño de muestra de estudios de confiabilidad	M	M	M	M	M	M
3.5 Valor del estadístico	B	A	M	NC	A	A
3.6 Interpretación de resultados de confiabilidad	B	A	A	NC	M	A
3.7 Modelo psicométrico para el estudio de confiabilidad*	B	-	-	-	-	B
3.8 Escogencia del método estadístico para estimar la confiabilidad*	B	-	B	-	-	B

Categorías indicadores obligatorios: NC= no cumple, M=mínimo puntaje entre 3.0 y 3.4, A=Aceptable puntaje entre 3.5 y 3.9 B= Bueno puntaje entre 4.0 y 5.0.

Categorías indicadores opcionales B= Bueno puntaje entre 4.0 y 4.5 y .S= Sobresaliente 4.6 y 5.0

Tabla 21

Porcentaje de indicadores de confiabilidad en cada categoría de calificación

Prueba	Indicadores Obligatorios (6)				I. Opcionales (2)	
	No Cumple	Mínimo	Aceptable	Bueno	Bueno	Sobresaliente
WISC IV	-	17%	-	83%	100%	-
16PF5	33%	17%	33%	17%	-	-
BDI II	17%	67%	17%	-	50%	-
IPV	50%	50%	-	-	-	-
MMPI 2	17%	67%	17%	-	-	-
MIPS	33%	17%	33%	17%	100%	-

Evaluación de los indicadores de validez

La tabla 23 presenta los resultados de la evaluación de los estudios de validez que se reportan en los manuales, de nuevo la prueba que mejor desarrolla la información es el WISC-IV. En general las cinco pruebas restantes no cumplen con los mínimos; de manera que se

pone en evidencia que no incluyen información que permita al usuario comprender las justificaciones desde las cuales los desarrolladores o responsables de la adaptación pretenden que se interpreten los resultados.

Tabla 22

Resultados de la evaluación de los indicadores de validez

Indicadores	WISC IV	16PF5	BDI II	IPV	MMPI 2	MIPS
Evidencia basada en el contenido de la prueba						
4.1 Estructura conceptual de la prueba	B	B	A	A	B	NC
4.2 Revisión por expertos idóneos	B	M	M	NC	B	M
4.3 Número de expertos y contenido de la consulta.	B	M	M	NC	NC	B
4.4 Resultados de evaluación por expertos	B	NC	NC	NC	A	M
4.5 Capacitación a expertos*	B	-	-	-	-	-
4.6 Tabla de especificaciones*	-	-	B	-	-	B
Evidencia basada en la estructura interna de la prueba						
4.7 Estructura interna de la prueba	B	B	B	M	M	A
4.8 Tamaño de muestra	B	M	M	M	M	NC
4.9 Resultados del estudio	B	B	B	M	M	B
Evidencia de validez basada en la relación con otras variables						
4.10 Estrategia de validación	B	B	B	M	B	B
4.11 Selección de la(s) variable(s) criterio	B	A	M	M	NC	A
4.12 Tamaño de muestra	NC	M	A	M	NC	NC
4.13 Análisis estadístico empleado	B	M	B	M	NC	M
4.14 Resultados de los estadísticos	B	A	B	NC	NC	A
4.15 Funcionamiento diferencial del ítems	S	-	B	-	-	-
4.16 Análisis de sesgo de la prueba	S	-	-	-	-	-
Evidencia con base en el análisis de ítems						
4.17 Modelo psicométrico utilizado	B	NC	NC	NC	M	M
4.18 Estimación de dificultad	M	NC	NC	NC	NC	NC
4.19 Estadísticos para análisis de ítems	S	-	S	-	-	-
4.20 Análisis de las opciones de respuesta	-	-	S	-	B	-
4.21 Parámetros del modelo de análisis	-	-	-	-	-	-

Indicadores	WISC IV	16PF5	BDI II	IPV	MMPI 2	MIPS
4.22 Ajuste del modelo	-	-	-	-	-	-
4.23 Función de información	-	-	-	-	-	-

Categorías indicadores obligatorios: NC= no cumple, M=mínimo puntaje entre 3.0 y 3.4, A=Aceptable puntaje entre 3.5 y 3.9 B= Bueno puntaje entre 4.0 y 5.0.

Categorías indicadores opcionales B= Bueno puntaje entre 4.0 y 4.5 y .S= Sobresaliente 4.6 y 5.0

La tabla 24 presenta los porcentajes indicadores de validez en cada categoría la escala de calificación. Estos porcentajes por categoría de nuevo evidencian que el WISC-IV incluye información que documenta los estudios para la obtención de los cuatro tipos de evidencias; en general, los manuales de las otras cinco pruebas incluyen alguna información aunque escasa.

Tabla 23

Porcentajes de indicadores de validez en por categoría

Evidencia basada en el contenido de la prueba

Prueba	Indicadores Obligatorios				I. Opcionales (2)	
	No cumple	Mínimo	Aceptable	Bueno	Bueno	Sobresaliente
WISC IV	-	-	0%	100%	50%	-
16PF5	25%	50%	0%	25%	-	-
BDI II	25%	50%	25%	-	50%	50%
IPV	75%	-	25%	-	-	-
MMPI 2	25%	-	25%	50%	-	-
MIPS	25%	50%	-	25%	50%	-

Evidencia basada en la estructura interna de la prueba

	No cumple	Mínimo	Aceptable	Bueno
WISC IV	-	-	-	100%
16PF5	-	33%	-	66%
BDI II	-	33%	-	66%
IPV	100%	0%	-	-
MMPI 2	100%	-	-	-
MIPS	33%	-	33%	33%

Evidencia de validez basada en la relación con otras variables

Prueba	Indicadores Obligatorios				I. Opcionales	
	No cumple	Mínimo	Aceptable	Bueno	Bueno	Sobresaliente
WISC IV	20%	-	-	80%	-	100%
16PF5	-	40%	20%	40%	-	-

BDI II	20%	20%	20%	60%	50%	-
IPV	20%	80%	-	-	-	-
MMPI 2	80%	-	-	20%	-	-
MIPS	20%	20%	40%	40%	-	-

Evidencia con base en el análisis de ítems**

Prueba	Indicadores obligatorios			I Opcionales	
	No cumple	Mínimo	Aceptable	Bueno	Bueno Sobresaliente
WISC IV	-	50%	0%	50%	- 20%
16PF5	-	-	-	-	-
BDI II	-	-	-	-	100
IPV	-	-	-	-	-
MMPI 2	-	20%	-	-	20%
MIPS	-	-	-	-	-

** Este apartado es opcional, debe aplicar e apruebas de conocimiento en particular,

Evaluación de la información sobre estandarización

La tabla 25 presenta los resultados de la evaluación la información que incluyen los manuales para orientar al usuario profesional sobre los procesos de aplicación y calificación de la prueba, los resultados de este apartado son de particular sensibilidad para la evaluación de la calidad de las pruebas y la pertinencia de su uso con población colombiana.

Los resultados de este apartado tienen particular importancia para este proceso ya que las pruebas evaluadas no cuentan con adaptación o baremos de calificación para Colombia, en este aspecto técnico se da relevancia a las características de los grupos con los cuales se sustenta las tablas normativas para la asignación de calificación. En la valoración del criterio sobre los tamaños de las muestras, las pruebas evaluadas cumplen con el mínimo para sustentar los baremos con población de habla castellana pero no colombiana.

Tabla 24 Resultados de la evaluación de la estandarización

Indicadores	WISC IV	16PF5	BDI II	IPV	MMPI 2	MIPS
5.1 Descripción de material de prueba	B	A	A	M	M	M
5.2 Instrucciones para la aplicación	B	B	A	M	M	M

Indicadores	WISC IV	16PF5	BDI II	IPV	MMPI 2	MIPS
5.3 Escalas y sistema de transformación de las puntuaciones	B	NC	A	NC	A	M
5.4 Descripción de la población objetivo y la(s) muestra(s) normativas	A	NC	A	M	M	M
5.5 Tamaño de la(s) muestra(s) normativa(s)	M	M	M	M	M	NC
5.6 Descripción del ámbito de aplicación y la calidad de los datos en que se basan los baremos	B	M	A	NC	NC	M
5.7 Protocolo de interpretación	B	B	A	M	A	NC
5.8 Condiciones especiales de aplicación*	S	-	-	-	-	-
5.9 Protocolo de presentación de resultados*	S	-	-	B	B	-

Categorías indicadores obligatorios: NC= no cumple, M=mínimo puntaje entre 3.0 y 3.4, A=Aceptable puntaje entre 3.5 y 3.9 B= Bueno puntaje entre 4.0 y 5.0.

Categorías indicadores opcionales B= Bueno puntaje entre 4.0 y 4.5 y .S= Sobresaliente 4.6 y 5.0

La tabla 26 muestra el porcentaje de indicadores del capítulo de estandarización, ubicados en cada una de las categorías de calificación para todas las pruebas.

Tabla 25

Porcentajes de cumplimiento de indicadores de estandarización

Prueba	Indicadores Obligatorios				I. Opcionales	
	No cumple	Mínimo	Aceptable	Bueno	Bueno	Sobresaliente
WISC IV	-	14%	14%	71%	-	100%
16PF5	28%	28%	14%	28%	-	-
BDI II	-	15%	85%	-	-	-
IPV	28%	72%	-	-	50%	-
MMPI 2	14%	57%	28%	-	50%	-
MIPS	28%	72%	-	-	-	-

Discusión y conclusiones

El objetivo general de este trabajo es ofrecer información sobre las calidades técnicas de algunas de las pruebas más usadas en Colombia mediante el desarrollo y uso de un instrumento que a futuro pueda considerarse un modelo nacional. Para el logro del mismo se identificaron seis pruebas de las más reportadas por los psicólogos en ejercicio, se diseñó un instrumento con criterios de evaluación de las calidades técnicas de las pruebas y finalmente, las pruebas seleccionadas se sometieron a evaluación utilizando el instrumento.

Un primer hallazgo que vale la pena comentar tiene que ver con las respuestas de los participantes a las preguntas sobre acceso a la información técnica de las pruebas que utilizan frecuentemente; en ellas se evidenció la poca importancia que los profesionales parecen otorgar a la revisión de la información de las propiedades métricas de las pruebas ya sea por desconocimiento o por descuido, y posiblemente en su uso no se hace una reflexión sobre el error posible en la interpretación ni las limitaciones de las mismas. Si bien, esto parecería un aspecto que atañe más a la cualificación de los profesionales que a la de los instrumentos, cabe esperar que una adecuada divulgación de los resultados de una investigación como la presente, contribuya a llamar la atención de los profesionales sobre el tema.

En segundo lugar, en lo referente a los resultados de la evaluación de las pruebas vale destacar la disparidad de los resultados en los cuatro apartes valorativos del instrumento y entre las pruebas evaluadas. En lo relacionado con los referentes conceptuales las seis pruebas cumplieron los indicadores obligatorios aunque los niveles de cumplimiento fueron muy diferentes. Sin duda la única que cumplió con los criterios fue el WISC-IV, que es la prueba de publicación más reciente, en 2005, y que se desarrolló considerando cada uno de los estándares de evaluación educativa y psicológica de APA, AERA, NMCE (1999), versión vigente para el momento de su desarrollo.

Las otras 5 pruebas evaluadas son ediciones publicadas entre 1995 o 1996, y la información en los manuales no se estructura claramente, encontrar información que responda a preguntas claves puede resultar dispendioso, en particular sobre la valoración de las definiciones operacionales y del desarrollo teórico del constructo.

En el caso de MMPI-II y el IPV, la combinación de la presentación de referentes conceptuales con información de las versiones anteriores y comentarios de la aplicación hizo difícil valorar algunos de los indicadores. En general, la revisión evidenció que en los manuales de las seis pruebas se desarrolla el elemento conceptual, pero que la rigurosidad, extensión, profundidad y nivel de detalle de la información es una elección de los editores y que varios casos omiten información valiosa.

Una característica que se relaciona con el detalle de la información conceptual es el modelo de construcción de las pruebas. Los manuales del 16 PF-5, MMPI 2 y BDI-II que fueron desarrolladas mediante enlace empírico, no se incluye la información de referentes conceptuales y al ser actualizaciones de versiones con más de 20 años de uso, no se documenta el procedimiento respondiendo a estándares vigentes. Lo mismo ocurre con la información relacionada con las diferentes evidencias de validez, las 5 pruebas de la década de los 90 omiten información sobre este tipo de estudios.

En cuanto a la evaluación de los aspectos métricos se evidencia que las pruebas aunque incluyen datos específicos de los indicadores requieren mayor detalle sobre los diseños de los estudios; es frecuente que cuando se hacen traducciones de una prueba, se reporten los resultados de la versión original, pareciera que se da por sentado que el valor estadístico hallado en los estudios originales son extrapolables a otras versiones o poblaciones. Sin embargo, dado que el instrumento diseñado para la evaluación de las pruebas no incluyó una indicación específica que permita al evaluador direccionar la valoración sobre los estudios que sustenten la traducción o adaptación, es posible que se hayan tomado los valores de los estudios originales como lo presentan la mayoría de los manuales.

En lo referente a la información sobre confiabilidad, las 5 pruebas de los años 90 mostraron una clara deficiencia al no incluir información sobre fuentes de error aleatorio relevante para los resultados de la prueba.

Respecto a la evaluación de validez es el aspecto en que se evidencian mayores falencias de las pruebas con manuales de menor extensión. En el caso del IPV se reportan los estudios iniciales de las versiones originales que datan de la década de los 60 y en las traducciones

descuidan la rigurosidad de los estudios para validar su uso en otras poblaciones. De igual manera el manual del 16 PF-5 se no se incluye información descriptiva de los estudios con esta versión sino los estudios realizados con la versión original. El manual sustenta el uso de la prueba en la trayectoria y los estudios que a lo largo de 40 años se desarrollaron con la versión original omitiendo la inclusión de información técnica específica de esta versión.

Respecto a la estandarización ninguna de las pruebas cuenta con tablas de calificación construidas a partir de estudios con población colombiana; son versiones de adaptaciones para México, Argentina y España. Todas ellas cumplieron en el mínimo con la condición establecida en cuanto a tamaño de muestra: 300 personas de habla castellana. Aunque en Colombia pueden existir estudios con estas prueba, estos no se incluyen en los manuales. Una manera de subsanar esta debilidad de los manuales puede ser el desarrollo de separatas con datos para Colombia.

Hay que anotar que este trabajo evaluó una muestra de las pruebas más usadas en todas las áreas de psicología, respetando los derechos de autor y tomó como criterio la disponibilidad en el mercado, sin embargo, se evidenció que los psicólogos colombianos utilizan versiones anteriores de estas pruebas, tal como ocurre con el 16 PF. La obsolescencia de los instrumentos utilizados incide directamente en la calidad de los resultados en términos de confiabilidad y validez, lo cual puede considerarse una falta ética y es conveniente divulgar esta información para movilizar un cambio respecto a la selección de las pruebas.

El uso de versiones no actualizadas está ligado a la disponibilidad en internet de pruebas que incluso se pueden descargar automatizadas y que no cuentan con la información de respaldo teórico para interpretar los resultados más allá de los párrafos prediseñados en el desarrollo de la aplicación automatizada.

Si se adoptara como criterio para uso de una prueba, que cumpliera la totalidad de los indicadores obligatorios, como lo sugirió el grupo de autores del instrumento, de las 6 pruebas seleccionadas solamente el WISC-IV resultaría recomendable, ya un 95% de los indicadores fueron calificados con nivel bueno de calidad; mientras que las otras 5 pruebas requieren actualizaciones importantes.

Debe mencionarse además que a pesar del esfuerzo de parte del grupo de autores del Instrumento para la Evaluación de las Calidades Técnicas de las Pruebas, para desarrollar una herramienta de fácil uso para la mayoría de los profesionales de la psicología, el ejercicio de revisión implica una lectura detallada del manual técnico y una adecuada formación en temas relacionados no solo con el uso de la pruebas, sino con algunos conceptos y metodologías propias de la psicometría, esto conlleva cierta dificultad para contar con evaluadores capacitados y con la disponibilidad para hacerlo.

Los informes técnicos que son el mecanismo de divulgación de los resultados de este tipo de evaluación de calidad de las pruebas, requieren la participación de no menos de 5 profesionales entrenados, dos evaluadores expertos en el área de uso de la prueba, uno medición que nutra y retroalimente sus comentarios, un coordinador experto en medición que revise y compare detalladamente la evaluación de estos evaluadores, y un árbitro que pueda hacer conceso en casos de controversias entre valoraciones, y recibir los comentarios de las casas editoriales para llegar a una versión pública de evaluación.

La designación de una mesa o comité de trabajo en este tipo de evaluaciones es una necesidad urgente para Colombia, y la organización llamada a atenderla es el Colegio Colombiano de Psicólogos. Se espera que la información sobre las calidades de las pruebas que se desarrolló en este trabajo permita a los psicólogos colombianos una reflexión sobre las prácticas y el uso de pruebas. Es deseable también que los informes de evaluación sirvan como un llamado de atención a las casas comercializadoras y la universidades para que se promuevan mecanismos para el desarrollo de las separatas técnicas para Colombia con reporte de estudios que den cuenta de la confiabilidad, la validez y la variación a partir de estudios con población colombiana.

Finalmente, vale la pena destacar el avance que representa para el gremio colombiano de psicólogos, contar con una primera herramienta para evaluar la calidad técnica de las pruebas. El Instrumento desarrollado en este trabajo es adecuado para iniciar el ejercicio de producir información y documentar con detalle las calidades técnicas los instrumentos que usan los profesionales en ejercicio. Futuras versiones deberían incluir indicadores relacionados con la vigencia y actualización de los referentes conceptuales, los estudios de confiabilidad y validez

y de las tablas de calificación. Aunque este aspecto fue discutido por parte del grupo de autores no se concertaron indicadores al respecto, por consenso se consideró que se debe procurar que las primeras evaluaciones sean un acercamiento más pedagógico que tenga efectos en la práctica profesional sin provocar rechazo a la evaluación de los instrumentos.

Referencias

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for Psychological Testing*. Washington.
- Aiken, L. R. (2003). *Test Psicológicos y evaluación*. México.
- Anastasi, A., & Urbina, S. (1998). *Test Psicológicos*. México: Prentice Hall.
- Arne, E. (1996). Regulations Concerning Test Qualifications and Test Use in the Netherlands. *European Journal of Psychological Assessment*, 12, 153-159.
- Baker, F. B., (2001). *The Basics of Item Response Theory*. USA: ERIC: Clearinghouse on Assessment and Evaluation.
- Bartram, D. (1995). Predicting Adverse Impact in Selection Testing. *International Journal of Selection and Assessment*, 52-61.
- Bartram, D. (1998). The need for international guidelines on standards for test use: A review of European and international initiatives. *European Psychologist*, 2, 155-163.
- Bartram, D., Lindley, J., Foster, J., & Marshall, L. (1992). Review of Psychometric Test (Level A) for Assessment in vocational Training. Leicester: BPS Books.
- Bartram, D. (2001). Guidelines for Test Users: A Review of National and International Initiatives. *European Journal of Psychological Assessment*, 17(3), 173-186.
- Bartram, D. (2002). EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Notes for Reviewers. Brussels: EFPA Standing Committee on Tests and Testing.
- Bartram, D. (2006). The Internationalization of Testing and New Models of Test Delivery on the Internet. *International Journal of Testing*, Vol. 6, No. 2: 121-131
- Bartram, D. (2011). Contributions of the EFPA Standing Committee on Test and Testing to Standards and Good Practice. *European Psychologist*, 16, págs. 149-159.
- Bartram, D. (2012). Concluding Thoughts on the Internationalization of Test Reviews. *International Journal of Testing* (12), 195-201.
- Borsboom, D., & Markus, K. A. (2013). Truth and Evidence in Validity Theory. *Journal of Educational Measurement*, 50(1), 110-114.

- Brandes, J. (2008). *Mental Measurements Yearbook*. Recuperado de <http://uclibrary.troy.edu.in>
- Brennan, R. L. (2013). Commentary on Validating the Interpretations and Use of Test Scores. *Journal of Educational Measurement*, 50(1), 74-83.
- Brogden, H. E. (1949), When Testing Pays Off. *Personnel Psychology*, 2: 171–183. doi:10.1111/j.1744-6570.1949.tb01397.x
- Buros Center for Testing. (24 de julio de 2015). <http://buros.org/>. Recuperado de <http://buros.org/history>
- Buros Center (2014) History of The Buros Center for Testing. <http://buros.org/history>. (Recuperado 3 de abril de 2014).
- Buros Center <http://buros.org/test-reviews-information> (Recuperado octubre de 2016).
- Camara, W. J. (1997). Use and Consequences of Assessment in the USA: Professional, Ethical and Legal Issues. *European Journal of Psychological Assessment*, 13, 140-152.
- Camara, W. J., & Lane, S. (2006). A Historical Perspective and Current Views on the Standards for Educational and Psychological Testing. *Educational Measurement: Issues and Practice*, 25(3), 35-41.
- Camargo, S. (2017). *En búsqueda del Consenso sobre el Concepto de Validez: un estudio Delphi*, tesis de maestría Universidad Nacional de Colombia, documento inédito. Bogotá, Colombia.
- Cardinet, J. (1995). Prehistory of the International Test Commission, *European Journal of Psychological Assessment*, Vol. 11, Issue 2, pp. 128–132
- Carlson, J., & Geisinger, K. F. (2012). Test Reviewing at the Buros Center for Testing. *International Journal of Testing*, 12, 122-135.
- Cohen, R.J. & Swerdlik, M.E. (2000) *Pruebas y evaluación psicológicas. Introducción a las pruebas y a la medición*. McGraw Hill. México.
- Cronbach, Lee J. (1951). «Coefficient alpha and the internal structure of tests». *Psychometrika* 16 (3): 297-334. ISSN 0033-3123. doi:10.1007/BF02310555
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. . *Psychological Bulletin*, 52, 281 302.

- DANE Dirección de Regulación, Planeación, Estandarización y Normalización. (2009). *Metodología para la estandarización de conceptos*.
- Eignor, D. R. (1999). Standards for the Development and use of Test: Standards for Educational and Psychological Testing. *European Journal of Psychological Assessment, 17*, págs. 157-163.
- Evers, A. (1996). Regulations Concerning Test Qualifications and Use in the Netherlands. *European Journal of Psychological Assessment, 160-168*.
- Evers, A. (2012). The Internationalization of Test Reviewing: Trends, Differences, and Results. *International Journal of Testing, 136-156*.
- Evers, A., Sijtsma, K., Lucassen, W., Meijer, R. (2010). The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results *International Journal of Testing, 10*: 295-317 DOI: 10.1080/15305058.2010.518325
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingsstelsel voor de kwaliteit van test*. Nederlands Instituut van Psychologen NIP.
- 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, Vol. 25, No. 3*, 283-291.
- Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: history, procedure and results. *International Journal of Testing, 10(4)*, 295-317. DOI: 10.1080/15305058.2010.518325
- Flynn, J. R. (1984). The mean IQ of Americans: massive gains 1932-1978. *Psychological Bulletin, 95*, 29-51.
- Flynn, J.R (2008). El Efecto Flynn. *Mente y Cerebro, 29-35*.
- Flynn, J.R (2009a). Massive IQ Gains in 14 Nations: What IQ Test Really Measure. *Psychological Bulletin, 101*, 171-191.
- Flynn, J.R. (2009b). *¿Qué es la Inteligencia? Más allá del efecto Flynn*. Madrid: TEA ediciones.
- Flynn, J. R. (2007). *What Is Intelligence Beyond The Flynn Effect*. Cambridge: University Press.

- Gempp, R. (2006). Error Estándar de Medida y la Puntuación Verdadera de los Test Psicológicos: Algunas Recomendaciones Prácticas. *Terapia Psicológica*, 24(2), 117-130.
- Hagemeister, C., Kersting, M., & Stemmler, G. (2012). Test Reviewing in Germany. *International Journal of Testing* (12), 185-194.
- Herrera, A. N. (2009). *Evaluación de la Calidad Técnica de las pruebas más frecuentemente Usadas por los psicólogos profesionales en Colombia*. Documento de trabajo Division de medición, evaluación y estadística aplicada.
- Herrera, A. N. (2013). Evaluando la Evaluación estándares técnicos de las pruebas. Conferencia, *III Congreso de Psicología Colpsic Ascofapsi*. Bogotá.
- Herrera, A. N. (2017). Diagnóstico sobre el nivel de formación del pregrado en áreas metodológicas y cuantitativas, juzgadas a partir de los resultados de las pruebas SABER-PRO de psicología. *Congreso colombiano de Psicología Colpsic Ascofapsi*. Medellín.
- Herrera, A.N y León, F.A (2015). Instructivo Instrumento para la Valoración De Calidades Técnicas de las pruebas psicológicas. Documento inédito desarrollado dentro del proyecto: Evaluación de la calidad técnica de las pruebas psicológicas usadas en Colombia. Universidad Nacional de Colombia, Bogotá.Colombia.
- International Test Commission (ITC). (2013). *ITC Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores*.
- Infraestructura Colombiana de Datos Espaciales. (4 de abril de 2015). <http://www.icde.org.co/>.
- Kane, M. T. (2013a). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1-73.
- Kane, M. T. (2013b). Validation as a Pragmatic, Scientific Activity. *Journal Educational Measurement*, 50(1), 115-122.
- Kerting, M. (2006). Zur Beurteilung der Qualität von Test: resüme und Neubeginn. *Psychologische Rundschau*, 57, 243-253.
- Kersting, M., 2008 DIN Screen, Version 2. Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen. In M. Kersting, *Qualitätssicherung in der Diagnostik und Personalauswahl - der DIN*

- Ansatz (S. 141-210). Göttingen: Hogrefe. Weitere Informationen: <http://www.kersting-internet.de/qualitaetssicherung/din-33430-buch/din-screen/>.
- Kersting, M., & Hornke, L. F. (2006). Improving the Quality of Proficiency Assessment the German Standardization Approach. *Psychology Science*, 48, 85-98.
- Koene, C. J. (1997). Test and Professional Ethics and Values in European Psychologist. *European Journal of Psychological Assessment*, 13, 219-228.
- León, F.A. Herrera, A.N.(2013). Retos para el Uso de pruebas en Colombia, ponencia III Congreso Colombiano de Psicología. Bogotá. Colombia
- León, F.A. (2017). Una mirada general sobre el uso de las pruebas en Colombia, *Boletines colpsic no. 24* * Campo De Evaluación, Medición y Estadística Aplicada. ISSN (en línea): 2462-8611
- Ley 1090, *Por la cual se reglamenta el ejercicio de la profesión de Psicología, se dicta el Código Deontológico y Bioético y otras disposiciones* (2006); Congreso de Colombia
- Linacre, John (2008). The Expected Value of a Point-Biserial (or Similar) Correlation. *Rasch Measurement Transactions*. 22 (1): 1154
- Lindley, P.A., Bartram, D., & Kennedy, N. (2008). EFPA Review Model for the description and evaluation of psychological tests: Test review form and notes for reviewers: Version 3.42. Brussels: EFPA Standing Committee on Tests and Testing.
- Lindley, P. A., & Bartram, D. (2012). Use of the EFPA Test Review Model by UK an Issues Relating to the Internationalization of Test Standards. *International Journal of Testing*, 12, 108-121.
- Martínez, R. (1996). *Psicometría: Teoría de los Test Psicológicos y Educativos*. España: Editorial Síntesis.
- Moosbrugger, H., Kelava, A., Hagemester, C., Kersting, M., Lang, F., Reimann, G., et al. (2009, Julio). The German Test Review System (TBS-TK) and first experiences. In D. Bartram (Chair), National approaches to test quality assurance. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway
- Muñiz, J. (1998). La Medición de lo Psicológico. *Psicothema*, 10, 1-21.
- Muñiz, J. (2003). La Validación de los Test. *Metodología de las Ciencias del Comportamiento*, 5(1), 119-139.

- Muñiz, J. (Junio de 2005). La Validez desde una óptica Psicométrica. *Acta Comportamentalia*, 9-20.
- Muñiz, J., & Bartram, D. (2007). Improving International Test and Testing. *European Psychologist*, 17, 206-219.
- Muñiz, J., Bartram *Psicothema*, D., Evers, A., Boben, D., Matesic, K., Glabeke, K. Zaal, J. N. (2001). Testing Practice in European Countries. *European Journal of Psychological Assessment*, 17, 201-2011.
- Muñiz, J., Campillo, Á., Fonseca, E., Fernández, J., & Peña, E. (2011). Evaluación de Test Editados en España. *Papeles del Psicólogo*, 32(2), 113-128.
- Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la Traducción y Adaptación de los Test: segunda edición., 25(2), 151-157.
- Muñiz, J., Fernández, J. R., Fonseca, E., Campillo, Á., & Peña, E. (2012). Test Reviewing in Spain. *International Journal of Testng*(12), 176-184.
- Muñiz, J., & Hambleton , R. K. (1996). Directrices para la Adaptación y Traducción de los Test. *Papeles del Psicólogo*, 66, 63-70.
- Muñiz, J., Hernández, A., & Ponsoda, V. (2015). Nuevas Directrices Sobre el Uso de los Test: Investigación, Control de Calidad y Seguridad. *Papeles del Psicólogo*, 36(3), 161-173.
- Nunnally, J. C., & Bernstein, I. J. (1995). *Teoría Psicométrica*. México, D.F.: McGraw Hill Latinoamericana.
- Oakland, T. (2004). Use of Educational and Psychological Test Internationally. *Applied Psychology: An International Rewiew*, 53(2), 157-172.
- Oaklang, T., Poortinga, Y. H., Schlegel, J., & Hambleton, R. K. (2001). International Test Commission: Its History, Current Status, and Future Directions. *International Journal of Testing*, 1, 3-32.
- Porto Noronha, A. P. (2012). Sistema de Avaliacao dos Test Psicológicos de Conselho Federal de Psicologia. Bogotá: Conferencia Universidad Nacional de Colombia.
- Prieto, G., & Delgado, A. R. (2003) .Análisis de un test mediante el modelo de Rasch. *Psicothema* 2003. Vol. 15, nº 1, pp. 94-100.
- Prieto, G., & Delgado, A. R. (2010). Fiabilidad y Validez. *Papeles del Psicólogo*, 31(1), 67 - 74.

- Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España [A model for the evaluation of test quality in Spain]. *Papeles del Psicólogo*, 77, 65-71.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.
- Resolução CFP N.º 002 (2003); Define e regulamenta o uso, a elaboração e a comercialização de testes psicológicos e revoga a Resolução CFP n.º 025/200. Conselho Federal De Psicologia
- Rossi-Casé, L., Neer, R. y Lopetegui, S. (2001). Test de Matrices Progresivas de Raven: Comparación de baremos. El aumento de los puntajes directos a través del tiempo. *Evaluar*, 2(2), 39-51.
- Rusell, E. W. (2007). Commentary: The Flynn effect revisited. *Applied Neuropsychology*, 14(4), 262-266. Recuperado de <http://dx.doi.org/10.1080/09084280701719211>
- Sarriá, A. (1999). Introducción a la estadística en Psicología. Editions de la Universitat de Barcelona.
- Shepard, L. A. (1993). Evaluating Test Validity. En Review of Research in Education (págs. 405-450). *American Educational Research Association*.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.
- Stevens, S. S. (1951). Mathematics, measurement and Psychophysics. En S. S. Stevens, *Handbook of experimental Psychology*. New York: Jhon Wiley
- Testkuratorium (2010). TBS-TK-Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009. *Psychologische Rundschau*, 61 (1), 52-56. http://www.kersting-internet.de/pdf/tbs_tk_2010_rundschau_61_52-56.pdf (PDF)
- Tristan, A. (1998). Análisis de Rasch para todos. México: CENEVAL
- Tristan, A (2001). Análisis de Rasch para todos: una guía simplificada para evaluadores educativos.
- Tristan y Vidal (2006). Estándares de calidad para pruebas objetivas. Editorial magisterio
- Wechsler, S.M., (2003). Test Development and Use in Brazil: Its History and Current Status. *Testing International*, 13(2), pág. 12.

- Wechsler, S. M.,(2013). El movimiento para el desarrollo de los tests Psicológicos en Brasil. Una experiencia pionera. Simposio Evaluación Psicológica en Iberoamérica Asociación para el Avance de la Ciencia Psicológica. <http://www.psiencia.org/ojs/index.php/psiencia/article/viewFile/116/164>
- Wechsler, S. M., & Lourenconi, M. A. (July de 2011). The Impact of ITC Guidelines on Test User for Test Development: the Brazilian experience. pág 6-7.

ANEXOS