ISSN - 1659-2921

8

CUADERNOS METODOLÓGICOS

Construcción de pruebas estandarizadas en el ámbito de la medición educativa y psicológica

Maria Paula Villarreal Galera
Lucrecia Alfaro-Rojas
Armel Brizuela Rodríguez

IIP Instituto de Investigaciones Psicológicas

Instituto de Investigaciones Psicológicas

Facultad de Ciencias Sociales

Universidad de Costa Rica

Construcción de pruebas estandarizadas en el ámbito de la medición educativa y psicológica

Maria Paula Villarreal Galera

Lucrecia Alfaro-Rojas

Armel Brizuela Rodríguez

Serie: Cuadernos Metodológicos del IIP

2015

Contenido

El presente documento ofrece una guía sintetizada para la elaboración de pruebas estandarizadas en el ámbito educativo. Su objetivo primordial consiste en ofrecer a investigadores, docentes y estudiantes algunas estrategias para desarrollar pruebas educativas estandarizadas de acuerdo con los estándares actuales en materia de medición y psicometría. Además, se brindan diversas referencias bibliográficas para quienes deseen profundizar en los tópicos esbozados.

Dada la amplitud que caracteriza al campo de la medición educativa, fue necesario delimitar los alcances de este documento a un único tipo de pruebas. Se optó por pruebas cuyo formato de respuesta sea de selección única, debido a su popularidad y fácil calificación. No se aborda en este documento la elaboración de pruebas de respuesta construida (en ocasiones denominadas *pruebas de desempeño*), ya que requieren de consideraciones adicionales, como por ejemplo la elaboración de rúbricas, el control de la subjetividad inherente a la evaluación mediante jueces, el entrenamiento de los calificadores, etc. Se espera poder abordar este tema en el futuro mediante otro documento.

Este cuaderno se divide en tres apartados. En el primero se exponen algunos conceptos fundamentales sobre los temas de validez y confiabilidad. En el segundo, se brindan algunas recomendaciones sobre cómo construir pruebas estandarizadas en el ámbito educativo, incluidas algunas recomendaciones para la provisión de adecuaciones para la aplicación de las mismas. En el tercer apartado se ofrecen algunos comentarios finales que sintetizan los principales aportes realizados con este trabajo.

PARTE I: VALIDEZ Y CONFIABILIDAD

Aun cuando existen diferentes nombres con diversos matices semánticos para referirse a una prueba educativa (test, escala, cuestionario, batería, inventario, examen, etc.), en el presente documento se utilizará la palabra "prueba" para aludir a todos estos mecanismos de evaluación. De este modo, dichos términos serán definidos de la siguiente forma: Una prueba es un dispositivo evaluativo o procedimiento mediante el cual se obtiene una muestra de la conducta de los(as) examinados(as) en un dominio especificado, el cual es posteriormente evaluado y calificado mediante un proceso estandarizado (AERA, APA & NCME, 2014, p. 2).

De acuerdo con la clasificación de Martínez, Hernández y Hernández (2006), las pruebas pueden ser clasificadas en función de diversos criterios, a saber:

1. *Consecuencias para los(as) examinados(as):* Algunas pruebas se consideran de altas consecuencias porque tienen implicaciones importantes sobre la vida de los(as) examinados(as), tales como las que usualmente se aplican como parte de la selección de aspirantes a ingreso a ciertas instituciones educativas. Por otra parte, las pruebas de bajas consecuencias son aquellas que no afectan de manera importante a las personas que las contestan, como las que se utilizan para propósitos de investigación (por ejemplo la medición de actitudes o de rasgos de personalidad).
2. *Formato de respuesta:* Existen fundamentalmente dos tipos de pruebas: De respuesta seleccionada y de respuesta construida. En las primeras, para contestar los ítems los(as) examinados(as) deben elegir una opción dentro de un conjunto que suele oscilar entre dos y siete alternativas, mientras que en las de respuesta construida es requerido crear la respuesta (ensayo, dibujo, modelo tridimensional, etc.).
3. *Área del comportamiento:* Se pueden distinguir las pruebas cognitivas, o de ejecución máxima, en las cuales los ítems tienen opciones con una o varias respuestas correctas y otras incorrectas, de las pruebas no cognitivas (de

ejecución típica) en las que las personas deben mostrar sus opiniones y actitudes respecto de diferentes temas.

4. *Medio de aplicación:* La gran mayoría de pruebas en psicología y educación se aplican en folletos de papel (pruebas de papel y lápiz), mientras que recientemente se han comenzado a aplicar mediante el uso de computadoras. Este último formato incluye las pruebas adaptativas, en las que los ítems se van presentando al examinado (a) de acuerdo con el desempeño mostrado.
5. *Número de sujetos:* Las pruebas de aplicación colectiva se utilizan para evaluar simultáneamente a una gran cantidad de personas, mientras que algunos tipos de pruebas (o en el caso de poblaciones particulares, como niños (as) o algunas personas con discapacidad) solamente pueden ser aplicados de manera individual.
6. *Demandas temporales:* De acuerdo con este criterio, las pruebas pueden ser de potencia o de velocidad. En estas últimas, los ítems suelen ser de baja dificultad y deben ser contestados en un tiempo relativamente corto, de manera que permitan observar la velocidad con la cual las personas resuelven algunos tipos de tareas cognitivas. Por otra parte, las pruebas de potencia suelen aplicarse en un contexto en el que el tiempo es suficiente para intentar contestar correctamente todos los ítems, sin embargo, estos suelen tener diferentes niveles de dificultad (bajos, medios y altos) para identificar distintos niveles de habilidad.
7. *Grado de aculturación:* Esta dimensión es un continuum más que un criterio de clasificación. Algunas pruebas demandan menos conocimientos específicos de una cultura (como las pruebas de razonamiento con figuras, que emplean poco lenguaje ya que en sus ítems la medición se realiza mediante líneas y figuras geométricas, sin embargo no puede afirmarse que están libres de la aculturación en su totalidad, pues estas se desarrollan respetando las características culturales de un país o región, por ejemplo, las que se realizan en lenguaje español siguen un formato de lectura de las figuras de izquierda a derecha, tal como se lee el idioma español), mientras que otras sí asumen un mayor bagaje cultural por parte de los(as) examinados(as), como las pruebas de comprensión

lectora, donde el lenguaje es indispensable para la resolución de los ítems.

8. *Modelo psicométrico:* En el ámbito de la psicometría, existen tres grandes marcos conceptuales (Martínez, 1996), a saber: Teoría Clásica de los Test, Teoría de la Generalizabilidad (considerada una extensión de la Teoría Clásica de los Test) y Teoría de Respuesta al Ítem. Las diferentes pruebas que existen en el mercado actualmente se han desarrollado con base en alguno de estos tres modelos, lo cual debe tomarse en cuenta para utilizarlas e interpretar sus resultados correctamente.
9. *Interpretación de puntuaciones:* De acuerdo con este criterio, existen pruebas referidas a normas y pruebas referidas a criterios. Las puntuaciones obtenidas por los(as) examinados(as) en una prueba referida a normas solamente se utilizan como elemento de comparación entre todas las personas que la contestan, mientras que los resultados obtenidos en una prueba referida a criterios tienen un significado específico con referencia a las habilidades y conocimientos que poseen las personas. Por ejemplo, para medir la inteligencia se utilizan pruebas referidas a normas que permiten establecer el nivel intelectual de una persona en relación con el resto de la población a la que esta pertenece, mientras que las pruebas de diagnóstico clínico (depresión, ansiedad, etc.) son referidas a un criterio con diferentes puntos de corte establecidos por expertos para indicar la presencia o ausencia, o el nivel de ciertos rasgos psicológicos.

I.1 Conceptos básicos sobre la validez

I.1.1 Definición de validez

De acuerdo con los *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014, p. 11, traducción propia),

> La validez se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones hechas a partir de los puntajes obtenidos en una prueba. La validez, por lo tanto, es la consideración más importante en el desarrollo y la evaluación de una prueba. El proceso de validación conlleva acumular evidencia para proveer una base científica sólida a las interpretaciones propuestas de los puntajes […] Cuando los puntajes de una prueba son usados en más de una manera, cada interpretación propuesta debe ser validada.

En este sentido, Cizek (2012) plantea que aun cuando existen diversos enfoques sobre la validez, es posible identificar algunos puntos de consenso, tales como: (a) la validez se refiere a las inferencias e interpretaciones derivadas de los puntajes obtenidos en una prueba y no a la prueba en sí misma, (b) no existen diferentes tipos de validez, (c) los procesos de validación están influidos por los valores y supuestos de quien recaba las evidencias de validez y (d) la validación es un proceso continuo que nunca finaliza de manera definitiva.

En general, se deben tomar en cuenta dos aspectos indispensables en cualquier proceso de validación: la validez de las inferencias basadas en los puntajes obtenidos en una prueba y la justificación de los usos de esta. Ambos aspectos no son compensatorios, es decir, poseer evidencia contundente sobre uno no compensa las debilidades o carencias en cuanto al otro.

De este modo, la validez puede ser adecuada o no en la medida en que los puntajes obtenidos mediante la correcta aplicación de una prueba permiten fundamentar las

inferencias en cuanto a lo que esta evalúa. Consecuentemente, el proceso de validación sería aquel mediante el cual se recaba, resume y evalúa la evidencia requerida para justificar las inferencias a partir de los resultados en una prueba. No obstante, evidenciar la relación entre los puntajes observados y la variable de interés no es suficiente (Cizek, 2012). Aunado a ello, es necesario justificar los usos de la prueba, esto es, la forma en que se utiliza para la toma de decisiones en contextos educativos, de selección de personal, clínicos, etc. En este sentido, se deben considerar todas las posibles consecuencias (tanto las negativas como las positivas) de su uso.

Una perspectiva complementaria a la anterior es que una prueba es válida para medir un atributo psicológico si y solo si (1) dicho atributo existe y (2) las diferencias entre los examinados(as) causan la variabilidad observada de los puntajes obtenidos en la respectiva prueba (Borsboom, Mellenbergh y Heerden, 2004). Por ello, es cuestionable el uso de términos como “validez aparente”, “validez predictiva”, “validez de constructo” y en general cualquier conjunto de categorías que dé a entender la supuesta existencia de diferentes tipos de validez. Por tal razón, la meta de cualquier proceso de validación debería consistir en establecer de qué manera las diferencias que presentan las personas en un constructo (en este documento, el término *constructo* se entenderá como variable no observable que se intenta medir mediante una prueba) medido por una prueba se reflejan en las diferencias observadas en los puntajes, en otras palabras, la preocupación primordial en cualquier esfuerzo de validación debería ser cómo representar adecuadamente mediante los ítems de una prueba los contenidos y habilidades que se pretenden medir.

La validación de las inferencias hechas a partir de una prueba requiere iniciar con una propuesta clara sobre las posibles interpretaciones que se pueden realizar a partir de esta, así como sus posibles usos (Kane, 2013). En este sentido, tanto las interpretaciones como los respectivos usos de una prueba deben basarse en un marco conceptual claro y explícito, ya que no es conveniente asumir que los potenciales usuarios de esta harán las interpretaciones correctas. Lo anterior es de vital

importancia si se toma en cuenta que la validez es una cuestión de grado y puede cambiar con el tiempo en función de los avances teóricos, tecnológicos y metodológicos relacionados con el rasgo o característica psicológica (habilidad, actitud, conocimiento, patología, etc.) que mide una prueba, por lo que siempre es necesario actualizar la validez de las interpretaciones y usos que se realizan con esta, así como también incorporar la evidencia nueva que se vaya generando.

Así las cosas, recabar evidencias de validez no es una tarea puntual que solamente se lleve a cabo una vez, sino que se requiere de un proceso continuo de investigación para fundamentar adecuadamente las inferencias realizadas con base en el uso de una prueba (Kane, 2013). Al igual que en cualquier esfuerzo científico para comprender mejor la realidad, es necesario considerar hipótesis rivales sobre los puntajes obtenidos en una prueba y abordar aquellos aspectos polémicos y cuestionables que puedan poner en entredicho la validez de las inferencias hechas con base en los resultados de una prueba. También es importante señalar que las poblaciones se transforman en el transcurso de los años, por lo cual también es necesario actualizar continuamente los instrumentos de medición para adecuarse de manera óptima a los cambios que experimenten las poblaciones en la variable de interés y en otras características de esta que puedan influir en la medición (léxico, formato de los ítems, etc.).

I.1.2 Evidencias de validez

En el transcurso del siglo XX, han predominado diferentes conceptualizaciones sobre el tema de la validez en el ámbito de la medición psicoeducativa (Markus y Borsboom, 2013). Durante los años 60, 70 y 80, se consideraba que existían distintos tipos de validez, lo cual provocó que los desarrolladores de pruebas eligieran arbitrariamente el tipo de validez que más les convenía para crear un instrumento psicométrico (Kane, 2013) en función, principalmente, de la disponibilidad de los datos. Actualmente, se promueve una concepción unitaria de la validez (Elosua, 2003; Montero, 2013) en virtud de la cual no existen distintos tipos de validez sino tipos de evidencia que se enfocan en diferentes facetas de la validez (AERA, APA & NCME, 2014).

I.1.2.1 Contenidos de la prueba

Para determinar la relevancia de los contenidos de una prueba respecto de lo que se pretende medir, es necesario fundamentar el desarrollo de la prueba en el juicio de expertos(as) que valoren si los ítems son una muestra adecuada de los contenidos en un área de interés, en otras palabras, si los temas, frases y formato de los ítems constituyen una muestra representativa del dominio de interés (Martínez, Hernández y Hernández, 2006). Por ejemplo, si se quisiera construir una prueba de comprensión lectora para estudiantes de secundaria, es necesario que un panel compuesto por expertos en currículo y profesores de lengua den su criterio en cuanto a si la prueba efectivamente evalúa apropiadamente las habilidades de un adolescente para comprender un texto escrito.

Es necesario también recabar evidencia sobre la pertinencia de los contenidos de la prueba, ya que esto fundamenta las inferencias realizadas a partir de una muestra limitada de ítems, dentro de un universo infinito de posibles preguntas (Kane, 2006). La evidencia recabada garantiza que el puntaje obtenido por los(as) examinados(as) en una prueba es generalizable y, por ende, que permite inferir razonablemente bien cuál sería el resultado de cada examinado si estos contestaran todos los posibles ítems en un dominio de interés. Dado que esta última situación es imposible en la mayoría de las ocasiones, es indispensable que los contenidos evaluados por los ítems representen lo mejor posible el conjunto infinito de posibles ítems que podría resolver el (la) examinado(a). Para alcanzar este objetivo de representatividad, es necesario al menos cumplir con los siguientes pasos (Crocker y Algina, 2006):

1. Definir con la mayor precisión posible el dominio de interés.
2. Conformar un panel de especialistas o jueces expertos en los contenidos incluidos en la prueba.
3. Implementar un procedimiento estructurado para que los (as) especialistas clasifiquen los ítems de acuerdo con los contenidos que se evalúan en la prueba.

4. Calcular el grado de concordancia entre los (as) expertos así como la proporción de ítems que fueron clasificados de acuerdo con las expectativas de los (as) desarrolladores de la prueba.

Existen diversas estrategias para recabar evidencias referidas al contenido de una prueba, en las cuales el papel del criterio experto es crucial. Para una explicación pormenorizada de cómo se aplican estas estrategias, se recomienda al lector consultar los artículos de Sireci (1998a, 1998b) y de Sireci y Faulkner-Bond (2014).

I.1.2.2 Estrategias de resolución de los ítems

Este aspecto se refiere a los procesos cognitivos o estrategias utilizados por las personas para contestar los ítems, se vincula directamente a las evidencias de validez de constructo de una prueba (Embretson, 1983; Gorin, 2006). Antes de construir una prueba se deben tener hipótesis específicas sobre la forma en que sujetos con diferentes características van a contestar los ítems que componen la prueba. Para ello, es necesario que la construcción de los ítems se fundamente en un marco teórico claro y sólido que permita operacionalizar en tareas específicas (los ítems) las variables de interés.

En términos metodológicos, existen diferentes técnicas psicométricas para recabar evidencias sobre los procesos de respuesta a los ítems, como los modelos de diagnóstico cognitivo (Roussos, DiBello, Stout, Hartz, Henson y Templin, 2007; Embretson, 2010; Rupp y Templin, 2010). Estas herramientas psicométricas se han difundido en gran medida por la gran disponibilidad de programas informáticos especializados para implementarlas. En términos generales, para utilizarlos es necesario identificar los atributos, procesos o estrategias requeridos para elegir una cierta categoría de respuesta en cada uno de los ítems. Para conocer con mayor detalle cómo aplican estos análisis para recabar evidencias referidas a los procesos de respuesta, se pueden consultar las fuentes citadas anteriormente.

Otra estrategia frecuentemente utilizada para conocer cómo las personas contestan los

ítems es el análisis de reportes verbales, conocidos en la literatura angloparlante como *think-aloud protocols* (Ericsson y Simon, 1993). Esta técnica consiste en solicitar a las personas que se enfrentan a una prueba que reporten en voz alta todo lo que piensan mientras contestan los ítems. Este procedimiento es ampliamente recomendado para obtener información relevante sobre qué habilidades y conocimientos se requieren para elegir una determinada opción de respuesta (Leighton, 2004; Leighton y Gierl, 2007).

Una estrategia similar desarrollada para elaborar encuestas y, en general, pruebas de actitud es la entrevista cognitiva (Smith y Molina, 2013). Además del reporte verbal, en la entrevista cognitiva se requiere que el sujeto lleve a cabo una serie de tareas diseñadas para obtener evidencias empíricas sobre la forma en que las personas interpretan los ítems, el tipo de contenidos que debe recordar, el nivel de las demandas cognitivas exigido, etc. Ambas técnicas son de gran utilidad antes de incursionar en el uso de modelos psicométricos como los de diagnóstico cognitivo. Así las cosas, aunado a la recopilación de las evidencias de validez referidas a los contenidos de la prueba (una labor en la que los expertos en el área cumplen un papel fundamental), para recolectar evidencias de validez también se requiere contar con la participación de los(as) examinados(as) (AERA, APA & NCME, 2014).

I.1.2.3 Estructura interna de la prueba

La identificación de los posibles patrones de asociación entre las respuestas a los ítems ha motivado la creación de métodos de análisis como el Análisis Factorial Exploratorio (Costello y Osborne, 2005; Beavers, Lounsbury, Richards, Huck, Skolits y Esquivel, 2013; Morales, 2013) y el Análisis Factorial Confirmatorio (Brown, 2006; Schreiber, Stage, King, Nora y Barlow, 2006; Jackson, Gillaspy y Purc-Stephenson, 2009). Asimismo, existen métodos especializados para el análisis de la estructura interna de una prueba, la cual suele estar compuesta por ítems cuya distribución no es normal ni continua (Tate, 2003).

Este tipo de evidencias es crucial porque fundamenta las decisiones sobre cómo

estimar la confiabilidad de la prueba, así como también se vincula con las dimensiones sustantivas que se pretenden medir (DeVellis, 2012). Asimismo, evaluar cuántas variables latentes subyacen a la prueba, se relaciona estrechamente con las inferencias que se pueden derivar de la aplicación de esta (Ríos y Wells, 2014). En este sentido, el ejemplo clásico que suele mencionarse es el de una prueba de habilidades matemáticas cuyos ítems presentan una gran cantidad de texto para plantear el problema: aunado a las habilidades cuantitativas, es plausible pensar que se estén evaluando destrezas asociadas a la comprensión lectora. Otro ejemplo es el de las pruebas de actitud en las que, además de las características psicológicas de interés, intervienen estilos de respuesta que no se relacionan con este, pero que afectan la medición (von Davier y Khorramdel, 2013).

A partir de las correlaciones (o varianzas y covarianzas) entre las respuestas de todas las personas, estos métodos permiten identificar (o confirmar la existencia) de conjuntos homogéneos de ítems. Dichas agrupaciones, en principio, deberían conformarse de acuerdo con los planteamientos teóricos que motivaron la creación de la prueba.

En relación con los procedimientos requeridos para evaluar la estructura interna de los test, se han publicado diversos trabajos en los que se exponen recomendaciones de suma importancia para utilizar adecuadamente los análisis factoriales en lo que respecta al análisis de ítems. A fin de evitar errores comunes en el uso de estas técnicas psicométricas, se recomienda al lector los trabajos de Ferrando y Lorenzo (2014); Izquierdo, Olea y Abad (2014); Lloret, Ferreres, Hernández y Tomás (2014).

I.1.2.4 Relación de la prueba con otras variables

Uno de los aspectos fundamentales respecto del tema de la validez es el grado en el que las puntuaciones de una prueba reflejan la variable (actitudes, habilidades, rasgos de personalidad, estados emocionales, conocimientos adquiridos mediante la educación formal, etc.) que se desea evaluar. Para indagar sobre este aspecto, Cronbach y Meehl (1955) plantearon que desde un punto de vista científico, para establecer con claridad cuál variable se mide mediante una prueba, es necesario enmarcarlo en una red de relaciones teóricas con otras variables. Este conjunto de relaciones se estructura por los vínculos teóricamente justificados entre diversas variables latentes (i.e., constructos) y sus respectivos indicadores (i.e., variables observadas). Así pues, recabar evidencias de validez referidas a la relación de la prueba con otras variables implica establecer un modelo sobre cómo se relacionan y cuantificar el ajuste de este a los datos observados (Bollen y Hoyle, 2012).

En la actualidad existen técnicas sumamente sofisticadas para poner a prueba diferentes hipótesis sobre las relaciones entre una prueba y otras variables, como los Modelos de Ecuaciones Estructurales (Kline, 2013) y el Análisis Factorial Confirmatorio (Brown, 2006), el cual es un caso particular de los modelos estructurales. Otra estrategia que ha sido utilizada para proporcionar evidencias de validez (discriminante y convergente) son las Matrices Multirrasgo-Multimétodo, tanto en su formulación original (Campbell y Fiske, 1959) como en su implementación mediante el Análisis Factorial Confirmatorio (Kenny, 1979; Brown, 2006; Raykov y Marcoulides, 2013) o mediante la Teoría de la Generalizabilidad (Woehr, Putka y Bowler, 2012).

Finalmente, otra técnica de uso extendido fundamentalmente en el ámbito de la medición educativa es la de los modelos de regresión múltiple, cuya formulación gira alrededor de los conceptos de validez predictiva (cuando la variable o criterio por predecir se mide posteriormente a la aplicación de la prueba), validez concurrente (cuando el criterio y la prueba se aplican de manera simultánea) y validez incremental (Smith, Fischer y Fister, 2003; Pedhazur y Pedhazur, 1991). Esta última noción se

refiere a la cantidad de información adicional que proporciona la prueba respecto de otros predictores como el género, nivel socioeconómico, institución educativa del que proviene el (la) examinado(a), entre otros.

I.1.2.5 Consecuencias del uso de la prueba

Las consecuencias (positivas o negativas) para los(as) examinados(as) derivadas del uso de una prueba fueron incorporadas de manera explícita y sistemática dentro del concepto de validez a mediados de los años 70 (Messick, 1975, 1980,1995). Desde esta perspectiva, el desarrollo de una prueba exige recabar evidencias sobre la presencia de varianza irrelevante al constructo así como los posibles efectos de la subrrepresentación de la variable de interés en el desempeño mostrado por los(as) examinados(as) (Haladyna y Downing, 2004). Este tipo de evidencias ha sido objeto de mucha polémica entre los (as) especialistas en medición y psicometría (Padilla, Gómez, Hidalgo y Muñiz, 2006). Lo anterior porque, aunque identificar posibles sesgos en contra de grupos poblaciones específicas es parte de los controles de calidad de una prueba estandarizada, se ha discutido ampliamente su conceptualización como un asunto de validez, así como también el papel que debería jugar quien crea la prueba.

Como se ha señalado, esta faceta de la validación está directamente vinculada al posible sesgo en contra de ciertos grupos sociales evaluados mediante una prueba. En este sentido, el *sesgo* en contra de ciertas poblaciones es causado por todos aquellos componentes irrelevantes que resultan en menores puntajes para ciertos subgrupos de examinados(as) (AERA, APA & NCME, 2014). Por tal motivo, es necesario identificar las posibles diferencias en cuanto a la capacidad predictiva de una prueba entre grupos sociodemográficos, tasas diferenciales de selección de examinados(as) con características irrelevantes al objetivo de una prueba, entre otros.

Aunado a lo anterior, también es importante determinar si personas con el mismo nivel en el rasgo evaluado presentan diferentes probabilidades de contestar correctamente los ítems que componen la prueba. Para ello, existen técnicas como el análisis del

funcionamiento diferencial del ítem (Hidalgo, Gómez y Padilla, 2005), las pruebas de invarianza factorial en el contexto del Análisis Factorial Confirmatorio (Millsap y Olivera, 2012), entre otras. Finalmente, también es importante cuantificar el impacto de la prueba, esto es, identificar las posibles diferencias entre los puntajes promedio de los distintos grupos sociodemográficos de examinados(as).

I.2 Conceptos básicos sobre la confiabilidad

Además de la validez, una propiedad importante de las inferencias realizadas a partir de los resultados de la aplicación de una prueba es que estos últimos sean confiables. En este sentido, la confiabilidad de los puntajes obtenidos en una prueba se refiere al grado en que estos están libres de error de medición, por lo cual un instrumento es considerado como confiable si arroja resultados similares cuando es aplicado en diversas ocasiones a un mismo conjunto de examinados(as) (Kumar, 2009). En este sentido, la confiabilidad de un instrumento se ve reducida en la medida en que las mediciones realizadas con este se ven afectadas por errores aleatorios de medición debidos a diversas circunstancias de los(as) examinados(as) (cansancio, motivación, etc.) y del ambiente (temperatura, ruido, etc.) en el que responden los ítems (Ross y Rowley, 1991).

Cabe señalar que la confiabilidad de los puntajes obtenidos en una prueba recibe un tratamiento diferente de acuerdo con la teoría psicométrica que subyazca a la prueba (Teoría Clásica de los Test, Teoría de la Generalizabilidad o Teoría de Respuesta al Ítem). Inclusive, dentro de la misma Teoría Clásica de los Test la confiabilidad se interpreta de distintas maneras (concordancia entre evaluadores, estabilidad temporal, equivalencia y consistencia interna u homogeneidad) en función de las técnicas utilizadas para estimarla (Tornimbeni, Pérez y Olaz, 2008; Muñiz, 2001). Aun cuando la estimación del error de medición ha sido una preocupación de la psicometría desde sus inicios, actualmente existe un gran dinamismo en cuanto a diferentes alternativas para estimarlo mediante técnicas de análisis sumamente sofisticadas, como los Modelos de Ecuaciones Estructurales y los Análisis Factoriales Confirmatorios (Raykov, 2012;

McDonald, 1999) o los modelos de respuesta al ítem (Furr y Bacharach, 2014).

Un alto nivel de consistencia de un instrumento para evaluar a los(as) examinados(as) no garantiza la validez de las interpretaciones que se realizan con base en este (Martínez, 1996; Kane, 2013). Como regla general, en el ámbito de la medición psicoeducativa se plantea que la confiabilidad es una condición necesaria pero no suficiente para concluir que las inferencias a partir de una prueba son válidas. Esta situación se da porque las diferentes técnicas estadísticas y psicométricas utilizadas para generar evidencias de validez (regresión múltiple, prueba t, ANOVA, etc.) se ven afectadas de diversas formas por el error de medición (sistemático y aleatorio) presente en las puntuaciones de las personas. Por otra parte, las estimaciones tradicionales de la confiabilidad no incluyen dentro de su concepción un manejo explícito del error sistemático (no aleatorio) que afecta los resultados de la aplicación de una prueba. Algunos de los métodos tradicionales para estimar el coeficiente de confiabilidad en el contexto de la Teoría Clásica de los Test son la correlación test-retest (Urbina, 2004) o entre pruebas paralelas y la correlación promedio entre los ítems de una prueba (Coaley, 2010).

Con el objetivo de aclarar la diferencia entre confiabilidad y validez, así como los de varianza irrelevante y subrrepresentación de la variable de interés, en la figura 1 se pueden apreciar cuatro cuadrantes en los cuales los círculos representan los contenidos o características psicológicas por medir y los cuadrados simbolizan los ítems de una prueba; cabe señalar que cuanto más cerca están entre sí los ítems (cuanto más alta es la correlación entre estos) mayor es su confiabilidad.

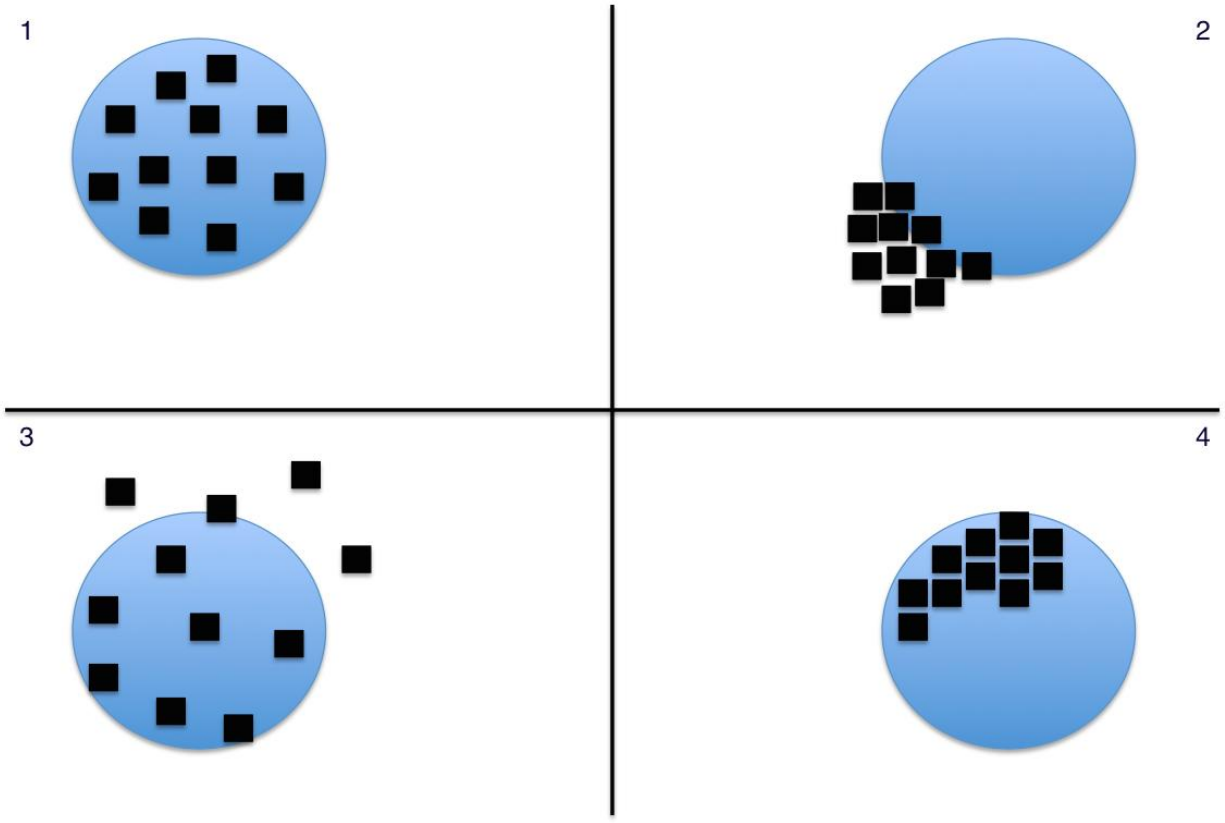


Figura 1. Medición de variables a través de indicadores. **Fuente:** Elaboración propia.

En el primer cuadrante se presenta la situación ideal en la que todos los ítems miden únicamente la variable de interés y tienen una alta correlación entre sí. Además, los contenidos de interés están siendo adecuadamente representados por los ítems, dado que toda el área del círculo está cubierta por ítems. Como se puede observar, el cuadrante 4 también contiene ítems altamente correlacionados entre sí, sin embargo, hay una gran parte de la variable que no está siendo medida por los ítems; de esta manera, el cuadrante cuatro es un ejemplo de subrrepresentación.

Por otra parte, en el cuadrante 2 se observa nuevamente que los ítems presentan un alto nivel de confiabilidad (porque están altamente correlacionados entre sí) pero hay una gran cantidad de estos que introducen varianza irrelevante. Esto se observa en el cuadrante 2 puesto que varios ítems no están dentro del área del círculo. Un caso similar sería el del cuadrante 3, en el cual la confiabilidad es mucho menor (los ítems

se alejan unos de otros) y algunos ítems no están dentro de la zona circular; no obstante, la gran mayoría sí están dentro del círculo y, además, representan de una mejor manera lo que se pretende medir con la prueba.

Hasta aquí se ha esbozado un panorama general sobre los conceptos de validez y confiabilidad. En el siguiente apartado, se expondrá cómo desarrollar una prueba psicoeducativa desde las perspectivas esbozadas anteriormente. Para ello, la información se estructura alrededor de las etapas que deberían seguirse para la construcción de una prueba estandarizada de altas consecuencias. En este sentido, adecuar la creación de una prueba a las etapas que se presentan a continuación dará como resultado un instrumento psicométrico de alta calidad en términos de validez y confiabilidad. Cabe señalar que estas etapas constituyen una guía de propósito general, la cual deberá ser adaptada según el propósito de la prueba que se desee construir y el contexto en el que este proceso se implemente. Es necesario recalcar que lo expuesto a continuación no es exhaustivo, por lo cual se recomienda profundizar en este tema mediante lecturas especializadas (Wilson, 2005; Downing y Haladyna, 2006; Schmeiser & Welch, 2006; Muñiz & Fonseca, 2008).

PARTE II: ETAPAS EN EL DESARROLLO DE UNA PRUEBA ESTANDARIZADA

Los estándares para la medición psicológica y educativa se crearon con el objetivo de promover el uso racional y ético de las pruebas estandarizadas, ya que estos permiten evaluar la rigurosidad científica y metodológica con que estas se construyen, aplican y utilizan (AERA, APA & NCME, 2014). En general, ofrecen una guía para medir la calidad de las pruebas y brindan un hilo conductor de las labores asociadas a la elaboración y uso de pruebas psicoeducativas.

A continuación se describen algunos lineamientos generales que los autores del presente documento sugieren implementar en las distintas etapas que componen el desarrollo de pruebas estandarizadas y particularmente las de selección única:

planificación, contenido o constructo, especificaciones de la prueba, construcción de ítems, diseño y ensamblaje, aplicación, calificación y reporte de resultados. Cabe señalar que estas recomendaciones se alimentan de los estándares internacionales de la AERA, la APA y el NCME (2014) y los propuestos por el *Educational Testing Service* (2002) que se consideraron pertinentes, así como otros aspectos que según los autores suponen relevantes. Los criterios presentados no son traducciones literales de los estándares mencionados, sino insumos para adaptar al caso particular de la medición educativa costarricense en cuanto a las pruebas estandarizadas, así como los estándares que los autores consideran pertinentes incorporar desde su experticia. Para adaptarlos al contexto costarricense los autores se basan en la bibliografía citada y en su experiencia profesional como desarrolladores de pruebas estandarizadas en lo que fue el Programa de Pruebas Específicas para Ingreso a Carrera.

La figura 2 es un esquema secuencial de las etapas y los recursos indispensables (banco de ítems, manual técnico y protección de la información) para el desarrollo de una prueba estandarizada. Para generar nuevas versiones de la prueba, es esencial contar con un banco o archivo en el que se incluyan los ítems y toda la información referente a estos (más adelante se especifica qué información debe incorporarse). Por otro lado, es de gran relevancia proteger la información a lo largo de todo el proceso, con el objetivo de resguardar la confidencialidad de la información de la prueba y sus resultados. El manual técnico es un documento que contiene un desglose estructurado de todas las tareas inherentes a la implementación de la prueba.

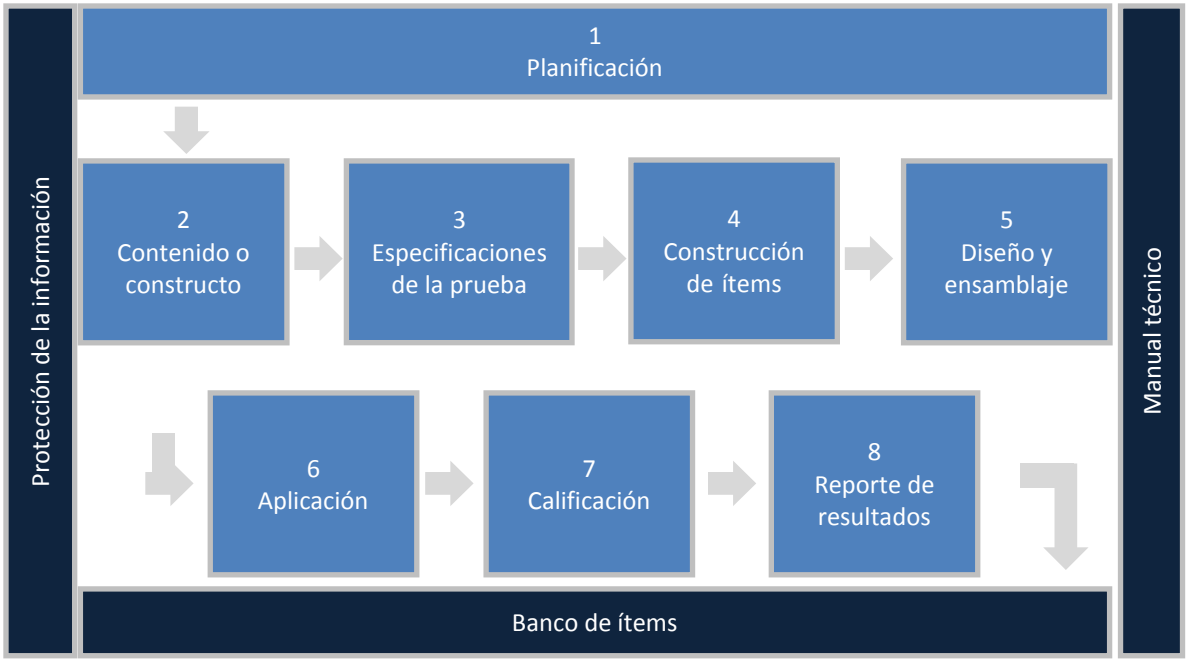


Figura 2. Desarrollo de una prueba estandarizada. **Fuente:** Elaboración propia.

II.1 Planificación

Antes de iniciar con la construcción de una prueba es necesario definir lo que se espera obtener de sus resultados. A continuación se exponen las actividades más relevantes en la etapa inicial de planificación:

1. Describir el propósito de la prueba, de tal manera que se pueda comprender la finalidad de esta. Para ello se debe contar con suficiente evidencia que respalde este objetivo, es decir, suficiente información que sustente la necesidad de su creación.
2. Definir las características psicológicas o contenidos que se intenta medir mediante la prueba (ver etapa II.2).
3. Describir el público meta (*stakeholders*), es decir, los usuarios de la prueba; personas evaluadas (examinados(as) o población meta); y autoridades o instituciones que podrían emplear los resultados de las pruebas para tomar decisiones.

4. Obtener realimentación de fuentes internas (equipo de trabajo) y externas (personas evaluadas y autoridades que emplean la prueba). En este sentido, es fundamental incorporar el aporte de proyectos de investigación relacionados con la prueba, otras investigaciones relacionadas con la variable que se intenta medir, las opiniones de la población meta y el criterio de las instituciones y *stakeholders* involucrados.
5. Obtener y documentar evidencias lógicas y empíricas sobre el propósito previsto de la prueba. Se debe recoger los resultados de las pruebas piloto realizadas, los prototipos y las observaciones realizadas por los(as) expertos(as) en los contenidos que se desea medir. La evidencia puede incluir los atestados de quienes diseñan, construyen y revisan la prueba. Esto es relevante en aras de generar evidencias que indiquen la idoneidad de la prueba para el propósito establecido y el público meta.
6. Elaborar un sustento sólido, mediante argumentos basados en investigaciones empíricas previas, sobre la forma elegida para medir los contenidos o constructos de interés. También es necesario clarificar la relación que tiene el constructo o contenidos evaluados con otros constructos o contenidos relacionados (ver apartado I.1.2.4).
7. Escoger un modelo de medida: Teoría Clásica de los Test, Teoría de Respuesta al Ítem, Modelo de Rasch, Teoría de la Generalizabilidad, Modelos de Diagnóstico Cognitivo, etc.
8. Presentar las limitaciones previstas con respecto al uso de la prueba y las posibles interpretaciones (ver apartado I.1.1) derivadas de su uso (selección, certificación, diagnóstico, etc.).

II.2 Contenidos o constructo

En esta etapa es imprescindible velar por la medición objetiva de los constructos o contenidos definidos en la etapa de planificación. A continuación se presentan algunas recomendaciones sobre esta etapa en el contexto del desarrollo de pruebas estandarizadas:

1. Definir el constructo o contenidos que se desea medir mediante la prueba.
2. Describir la teoría y la evidencia que sustenta el constructo, así como el uso y las inferencias que se realizarán a partir de los resultados de la aplicación de la prueba.
3. Detallar el tipo de conocimientos o habilidades por medir en la prueba.
4. Definir y documentar el grado en que la prueba representa el constructo que se intenta medir (ver apartado I.1.2.5).
5. Obtener y documentar la evidencia que respalde si la evaluación puede cumplir el propósito previsto, esto es, indicadores empíricos que respalden las inferencias derivadas del uso de la prueba:

    - Listado o tabla de los procesos cognitivos, contenidos, habilidades, estrategias, tareas o atributos que los desarrolladores determinan para construir los ítems.
    - Listado o tabla de los procesos cognitivos, habilidades, estrategias o atributos que emplean los(as) examinados(as) para resolver los ítems.
    - Fundamento que respalde la interpretación de las respuestas de los ítems.
    - Relación entre los puntajes obtenidos en la prueba y otras variables externas.
    - En general, evidencias que sustenten la relación entre los resultados de la prueba y los de otras evaluaciones o variables pertinentes.

6. Con base en lo expuesto, es necesario generar una tabla de especificaciones que les permita a quienes construyen ítems conocer en detalle los temas y contenidos, habilidades cognitivas, tareas, entre otros, que se evaluarán mediante la prueba. Dicha tabla también será de gran utilidad a quienes deban juzgar la calidad técnica de los ítems, los cuales deberán determinar si estos se ajustan al propósito de la prueba y a los contenidos establecidos en la primera etapa. Asimismo, la tabla de especificaciones les servirá a los(as) examinados(as) para conocer los contenidos o habilidades que serán evaluados,

para así prepararse cuando deban tomar la prueba. Finalmente, la tabla será de la mayor relevancia para quienes emplean los resultados de las pruebas en la toma de decisiones, ya que esta les permitirá fundamentar adecuadamente las inferencias derivadas de los puntajes obtenidos en la prueba.

II.3 Especificaciones de la prueba

Esta etapa debe reflejar detalladamente las características de la prueba, junto con su justificación y el proceso mediante el cual fueron desarrollados cada uno de los siguientes aspectos:

1. Tipo de prueba: Es conveniente clasificar la prueba dentro de alguna tipología (ver los criterios expuestos en la primera parte) para delimitar las inferencias que se pueden derivar del uso de la prueba.
2. Instrucciones: Debe aclararse qué está permitido y qué está prohibido durante la aplicación de la prueba, así como también todas las directrices que deben cumplirse para contestar los ítems. (Ejemplo: No se permite usar calculadora, deberá emplear una hoja de lectora óptica para seleccionar las respuestas, etc).
3. Requisitos para realizar la prueba, como por ejemplo el poseer un grado mínimo de bachiller universitario, haber aprobado un determinado bloque de materias, etc.
4. Extensión de la prueba: Cantidad de ítems y tiempo con que se cuenta para contestarlos.
5. Cantidad de ítems para cada área de contenido o habilidad.
6. Explicitar, de acuerdo con el tiempo de administración, si la prueba es de potencia, velocidad o una combinación de ambas.
7. Procedimiento utilizado para calificar e interpretar los puntajes de la prueba. Cuando sea pertinente, se deben describir las normas o muestras de estandarización (muestra de examinados(as) en la que se aplica una prueba para obtener ciertos estadísticos de referencia para interpretar los puntajes). En este sentido, se debe aclarar si el puntaje es absoluto, relativo o una

combinación de ambos, así como también si los puntajes se interpretan con base en normas o en criterios.

8. Protocolos sobre cada procedimiento por realizar para quienes desarrollan la prueba, para quienes deben administrarla y para los(as) examinados(as).

II.4 Construcción de los ítems

Una vez que en las etapas anteriores se realizó la selección de los contenidos o constructo de interés, la población meta, el tipo de ítems, los formatos de respuesta, los procedimientos de calificación y los protocolos de aplicación, etc., se procede a abordar la etapa de construcción de los ítems. Seguidamente se enlistan algunas recomendaciones básicas que deben tomarse en cuenta en esta etapa. Para un abordaje más detallado sobre cómo construir ítems, puede consultar los trabajos de Moreno, Martínez y Muñiz (2006) y Haladyna y Rodriguez (2013).

1. Los ítems construidos deben medir el constructo o contenidos establecidos en la etapa de planificación y desglosados en la tabla de especificaciones, de modo que quienes construyan los ítems tengan claro cómo elaborarlos.
2. Los procedimientos para desarrollar, revisar, probar y seleccionar ítems deben ser documentados. En el contexto del juzgamiento de ítems por parte de jueces expertos(as), es conveniente registrar si los ítems fueron clasificados en diferentes categorías de acuerdo con las especificaciones de la prueba, los procedimientos usados para la clasificación, así como la adecuación y exactitud de la clasificación (ver apartado I.1.2.1).
3. Si en investigaciones previas se determina que la varianza irrelevante al constructo (ver apartado I.1.2.5) podría confundir el dominio de interés de la prueba, en la medida de lo posible, el (la) desarrollador(a) debe investigar las fuentes de este tipo de varianza, es decir, indagar sobre aquellos aspectos que se alejen del contenido de interés. Un ejemplo sería el de una prueba para evaluar conocimientos sobre operaciones con números fraccionarios en la que los ítems incluyan textos muy largos, de modo que algunos examinados(as)

podrían contestarlo incorrectamente por problemas de comprensión lectora y no por el desconocimiento de las operaciones con este tipo de números. Cuando sea factible, esas fuentes deberían reducirse y, preferiblemente, eliminarse. En este sentido, si además se cuenta con evidencia empírica de diferencias en los efectos de este tipo de varianza irrelevante al constructo en diferentes subgrupos de examinados(as), la prueba debe utilizarse para la toma de decisiones solamente en aquellos casos en los que se puedan realizar inferencias válidas a partir de los puntajes obtenidos en la prueba.

4. Si en investigaciones preliminares se ha documentado que pruebas similares subrrepresentan el constructo (ver apartado I.1.2.5), en la medida de lo posible, el (la) desarrollador(as) de la prueba debe detallar cuáles son aquellas áreas que se excluyen de la prueba y una justificación que respalde esta decisión. Cabe señalar que una prueba debe estar compuesta por ítems que representen lo mejor posible los contenidos o el constructo de interés.
5. Los desarrolladores de la prueba deben juzgar los ítems que hayan construido otros colegas, y viceversa. Para ello se empleará una rúbrica (derivada de la tabla de especificaciones mencionada en el punto 6 del apartado II.2) que les permita valorar de manera adecuada la calidad técnica de cada ítem. Es conveniente incluir los criterios de otros profesionales que puedan servir para mejorar los ítems.
6. Debe haber un balance en cuanto a la cantidad de ítems que se construyen para cada área de contenido o faceta del constructo.
7. Antes de ser utilizada para la toma de decisiones, la prueba debe ser aplicada en un estudio piloto para conocer las propiedades psicométricas (dificultad, discriminación, funcionamiento diferencial, impacto, entre otros) de los ítems. El estudio piloto también permitirá identificar la presencia de posibles errores de redacción o inconsistencias respecto al propósito de la prueba.
8. La evidencia de naturaleza cualitativa (ver apartado I.1.2.2) es de gran importancia para determinar si los ítems realmente cumplen el propósito para el que fueron construidos. Por ello es importante analizar los ítems por medio de las distintas técnicas esbozadas en la primera parte del documento para recabar

información sobre las habilidades y los conocimientos empleados por los(as) examinados(as) para escoger una respuesta, tales como los *think-aloud protocols* (Ericsson y Simon, 1993) o la entrevista cognitiva (Smith y Molina, 2013). En la aplicación de estas técnicas es importante que participen personas similares (en edad, estatus socioeconómico, nivel educativo, etc.) a las que tomarán la prueba cuando esta se utilice para la toma de decisiones. De este modo, será posible determinar si los(as) examinados(as) comprenden los ítems, si interpretan correctamente las instrucciones, si la redacción es clara, etc.

9. Mediante distintas técnicas psicométricas, como los modelos de diagnóstico cognitivo, es conveniente recolectar evidencias sobre los procesos de respuesta a los ítems (Roussos, DiBello, Stout, Hartz, Henson y Templin, 2007; Embretson, 2010; Rupp y Templin, 2010).
10. Finalmente, un paso crucial es el de documentar el adecuado cumplimiento de estos lineamientos, así como aquellos que no fueron implementados en el desarrollo de la prueba.

II.5 Diseño y ensamblaje

Para el diseño y ensamblaje de la prueba es conveniente elaborar una guía que contenga los procedimientos para seleccionar los ítems con mejores propiedades psicométricas, respetando lo establecido en la tabla de especificaciones. A continuación se exponen algunas recomendaciones para implementar esta etapa.

1. Con base en los índices estadísticos de cada ítem (fundamentalmente la dificultad y la discriminación), seleccionar los ítems que conformarán la prueba.
2. Determinar la cantidad de versiones que tendrá la prueba. En el contexto de la medición educativa, con el objetivo de evitar que entre los(as) examinados(as) se copien las respuestas, se suelen crear dos o más folletos de la prueba. La diferencia entre los formularios puede ser en el orden de los ítems o inclusive pueden estar compuestos por un subconjunto de ítems que es el mismo en todos los formularios (usualmente denominados ítems de anclaje) y otro

subconjunto con ítems diferentes para cada folleto.

3. Establecer el orden de las partes o secciones de la prueba, de modo que los ítems de mayor dificultad aparezcan al final de la prueba (o de cada sección en caso de estar compuesta por varias partes).
4. Distribuir los ítems de acuerdo con el área de contenido o faceta del constructo medido.
5. Decidir cuántos ítems servirán de anclaje en cada versión de la prueba. Cuando se cuenta con dos o más versiones de una misma prueba, se suele utilizar un conjunto de ítems en común para todas las versiones o folletos de examen. Dicho conjunto de ítems se conoce como *de anclaje*.
6. En el caso de que se incluyan ítems experimentales (ítems con propiedades psicométricas desconocidas) dentro de las versiones de la prueba, determinar cuántos serán y dónde serán colocados dentro de cada folleto. En algunas pruebas se le advierte al examinado cuáles son los ítems experimentales (normalmente al final de la prueba) para que colabore con la resolución del ítem y así obtener estadísticas del mismo; sin embargo no se le califica. En otros casos los ítems experimentales pueden estar distribuidos en la prueba sin que el examinado pueda determinar cuáles son, en este caso lo recomendable es calificar estos ítems como correctos a todos los sujetos, pues hasta ese momento es que serán experimentados y se desconoce su calidad psicométrica.
7. Una vez ensamblada la prueba, debe proporcionarse una versión borrador a los desarrolladores de la prueba para que estos realicen una última revisión de la prueba.

II.6 Aplicación

Una prueba se considera estandarizada cuando se construye y administra respetando normas preestablecidas y bajo condiciones sistematizadas y equiparables. Es necesario cumplir con ciertos requerimientos para lograr una consecución exitosa en su fase de administración, es decir, cuando los(as) examinados(as) reciben la prueba para que contesten los ítems.

1. Se debe capacitar de manera uniforme a quienes conformen el equipo de aplicadores. Como consecuencia de ello se espera que los lineamientos de aplicación sean llevados a cabo de la misma manera en cada aplicación de la prueba.
2. Las instrucciones para aplicar la prueba deben presentarse con suficiente claridad, de modo que sea posible replicar las condiciones de aplicación en las que se obtuvieron las evidencias de validez y confiabilidad iniciales.
3. Las instrucciones de aplicación presentadas a los(as) examinados(as) deben ser lo suficientemente específicas y claras como para que estos puedan responder a la prueba de la manera prevista por el desarrollador. Cuando sea apropiado, se debe brindar a los(as) examinados(as) algunos ejemplos de ítems de cada parte de la prueba y de los criterios utilizados para calificar las respuestas. Es recomendable elaborar un folleto de práctica con esta información que sea similar a la prueba, de modo que los(as) examinados(as) puedan familiarizarse con el tipo de ítems que deberán resolver en la prueba.
4. Los(as) aplicadores(as) de la prueba deben contar con información sobre el propósito de la evaluación, de modo que puedan responder a eventuales dudas o consultas de los(as) examinados(as). Asimismo, conviene que los aplicadores conozcan las fechas de entrega de resultados, tiempo de aplicación de la prueba, entre otros. En general, es necesario brindar a los aplicadores materiales y herramientas aptas para poder cumplir con su tarea (listas de examinados(as), lápices, documentos con información sobre la prueba, instrucciones de aplicación, etc).
5. Entregar y explicar a los(as) aplicadores(as) los procedimientos y protocolos de seguridad y mantenimiento de la información confidencial (ver apartado II.10).
6. En caso de que los aplicadores tengan inquietudes o inconvenientes durante la administración de la prueba, estos deben tener acceso a un contacto inmediato de los encargados principales de la prueba.
7. Es recomendable que los(as) aplicadores(as) dispongan de un instrumento para evaluar el proceso de aplicación en aras de recabar información sobre fallos o

debilidades imprevistas en la aplicación de la prueba.

8. Cada aplicador(a) debe tratar de garantizar las mejores condiciones posibles para la aplicación: ambiente silencioso, temperatura e iluminación adecuadas, mobiliario en buen estado, respeto y consideración hacia los(as) examinados(as), entre otros.
9. Cuando se designe al equipo de aplicadores, se debe exponer todo lo anterior mediante capacitaciones planificadas que utilicen manuales técnicos y apoyos audiovisuales para ilustrar la aplicación de la prueba.

II.7 Calificación de las respuestas

En la etapa de calificación de las respuestas, quienes tengan bajo su responsabilidad dicha tarea deberán realizar este proceso de la forma más sistemática posible, con el fin de evitar sesgos y manipulaciones incorrectas de los datos. A continuación se especifican algunos aspectos que constituyen una parte esencial de esta fase.

1. Es importante especificar los procedimientos y criterios para calificar la prueba, con la especificidad y claridad necesaria.
2. Explicitar los modelos psicométricos utilizados (Teoría Clásica de los Test, Teoría de Respuesta al Ítem, Modelo de Rasch, Teoría de la Generalizabilidad, Modelos de Diagnóstico Cognitivo, entre otros) para evaluar las propiedades psicométricas de los ítems. Según el modelo empleado, deben registrarse los parámetros pertinentes de los ítems (ajuste del ítem, dificultad, discriminación, etc.).
3. En el caso de tener varias versiones o fórmulas de una misma prueba, es necesario implementar y documentar los procesos de escalamiento o equiparación entre las distintas versiones.
4. Se debe contar con una plantilla de “claves” o respuestas correctas para que el programa informatizado proceda a calificar las respuestas de los(as) examinados(as) en cada versión de la prueba.
5. Se debe contar con procesos de revisión y análisis de inconsistencias en la

calificación de las pruebas que permitan reducir tanto como sea posible los errores en el procesamiento de las respuestas de los(as) examinados(as).

II.8 Reporte de resultados

Los(as) encargados de las pruebas deben reportar resultados correctos, exactos (tanto como se pueda), y comprensibles para los(as) examinados(as). Por ello, quienes se encargan de construir pruebas estandarizadas deben asegurar que los resultados sean confiables (ver apartado I.2) y comunicar cuando sea necesario las fuentes de error que afecten dichos resultados.

1. Es de vital importancia documentar la justificación y los procedimientos empleados para establecer determinados puntos de corte (como por ejemplo las notas mínimas de aprobación de un examen) cuando las interpretaciones hechas a partir de los puntajes de una prueba así lo demanden.
2. Siempre que sea posible, se deben establecer puntos de corte para definir categorías y realizar interpretaciones sustantivas. Esto se debe generar con base en datos empíricos adecuados sobre la relación entre el desempeño en la prueba y otros criterios relevantes. Este proceso debe permitir que los especialistas en los contenidos o constructo medidos por la prueba utilicen su conocimiento y su experiencia para establecer puntos de corte útiles la selección de examinados(as).
3. En lo que respecta a las pruebas educativas y de certificación, los(as) examinados(as) tienen derecho a recibir respuestas razonables sobre las apelaciones que realicen.
4. Los(as) encargados(as) de la prueba tienen el deber de reportar a los(as) examinados(as) los resultados en un periodo razonable de tiempo y de manera comprensible.
5. Quienes desarrollan la prueba deben estar preparados para corregir interpretaciones erróneas de los puntajes de la prueba, así como para lidiar con consecuencias inesperadas del uso de la prueba (ver apartado I.1.2.5).

6. En contextos educativos, cuando se divulguen los puntajes promedio para grupos, se debe reportar información acerca del tamaño de la muestra y la forma o dispersión de la distribución de los puntajes.
7. Informar a los *stakeholders* de la evaluación sobre los propósitos de la prueba, cómo se aplica, de qué modo se califica y durante cuánto tiempo se archivan los resultados.

II.9 Banco de ítems

El banco de ítems es un archivo (físico o digital) en el que se almacenan los ítems que cumplen con todos los requerimientos psicométricos para evaluar los contenidos o constructo de interés. En este archivo se debe disponer de la información sobre las propiedades psicométricas de cada ítem. Tiene la ventaja de permitirle al desarrollador observar de una manera fácil y rápida las características de los ítems disponibles, lo cual agiliza el diseño y ensamblaje de las diversas versiones de la prueba. Las siguientes figuras permiten ejemplificar las tarjetas de un ítem perteneciente a un banco de preguntas de una prueba de razonamiento con figuras.

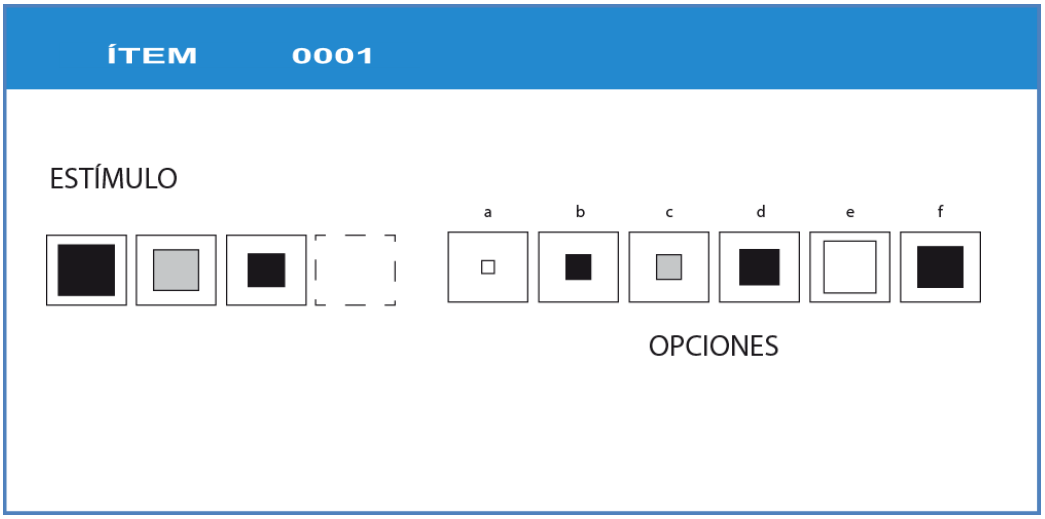


Figura 3. Tarjeta del ítem 0001 de una Prueba**. Fuente:** Elaboración propia**.**

En la figura 3 puede observarse el enunciado o estímulo del ítem y sus respectivas opciones de respuesta (en este caso, son seis opciones). Esto permite identificar el ítem independientemente de la numeración que asuma en distintas fórmulas de examen aplicadas posiblemente en diferentes momentos. Por otro lado, en la figura 4 se presentan las propiedades psicométricas del ítem en una aplicación específica (en este ejemplo, el ítem fue aplicado en el mes de julio del año 2014). Cabe señalar que el número de tarjetas para cada ítem dependerá de cuántas veces haya sido aplicado. Dicho registro histórico permite analizar el comportamiento de este ítem a lo largo del tiempo, lo cual es información de gran utilidad para determinar su calidad psicométrica.

ÍTEM 0001

Fecha de aplicación	julio, 2014
Población	Meta
Fórmula	5
Posición del ítem	1
N	25000

Infit	0.92
Outfit	0.83

DIF Género: Ausencia

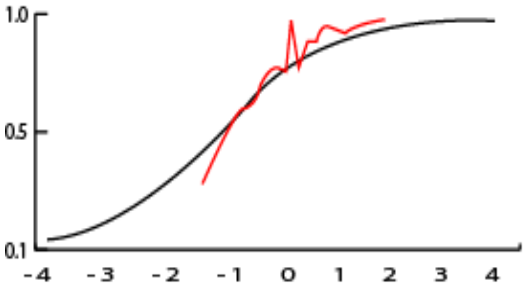
Rasch	0.495	b Fem.	-1.180
MH	0.646	b Masc.	-1.360

Parámetros	TCT	Rasch
Discriminación	0.423	0.44
Dificultad	0.817	-1.26

DIF Colegio: Ausencia

Rasch	0.095	b Públic.	-1.17
MH	0.459	b Privad.	-1.88

Análisis de Distractores	Prop. Total	Corr.bis. Puntaje T
a	0.008	-0.174
b	0.061	-0.188
c	0.817	0.423
d	0.014	-0.190
e	0.006	-0.097
f	0.090	-0.227
No respuesta	0.004	-0.126



Respuesta c

Figura 4. Tarjeta del ítem 0001 con la información psicométrica**. Fuente:** Elaboración propia.

1. Los ítems deben ser archivados tanto en un banco digital como en uno físico, los cuales deberán ser actualizadas constantemente con sus respectivas estadísticas.
2. Este archivo de ítems debe organizarse de acuerdo con la tabla de especificaciones de la prueba.
3. En el banco físico, cada ítem debe corresponder a una ficha o tarjeta. En el anverso de esta deberá aparecer el ítem así como la respuesta correcta señalada y en el reverso de la tarjeta se podrá consignar un historial con datos estadísticos del ítem organizado cronológicamente de acuerdo con el año en el que se aplicó el ítem.
4. Dentro del conjunto de ítems que componen el banco, es conveniente señalar aquellos ítems que hayan sido aplicados recientemente para así evitar que un mismo ítem se exponga demasiado a la población meta. De manera indirecta, esto permitirá conocer con toda certeza cuántos ítems hay disponibles para una determinada aplicación.
5. De acuerdo con las necesidades específicas de cada caso, es necesario y deseable incrementar el banco anualmente ingresando los ítems experimentales que presenten propiedades psicométricas adecuadas.
6. Deben ser documentados los ítems que se desechan del banco. Es importante contar con una justificación (ya sea de tipo psicométrico, teórico o de otro tipo) que respalde dicha decisión.
7. El lugar en donde se ubique el banco de ítems deberá ser seguro y estar acondicionado para el almacenamiento de documentos. Será necesario contar con una bitácora de las personas que ingresan a este espacio así como también del propósito de haber accedido al banco. Cabe resaltar que el resguardo de los ítems en el banco digital es de la mayor importancia, dada la gran cantidad de herramientas informáticas existentes para irrumpir en bases de datos. En este sentido, sería conveniente contar con la colaboración de un especialista en seguridad informática para establecer la forma en que deben archivarse los ítems en el banco digital.

II.10 Protección de la información

Los programas de pruebas deben garantizar la protección de la información confidencial relacionada con las pruebas y con los(as) examinados(as). Deben promover una cultura organizacional que apunte a la internalización del resguardo de las pruebas (en algunos casos, las pruebas son divulgables una vez que se aplican en la población meta), los resultados derivados de ellas y las consecuencias directas sobre los(as) examinados(as). Este componente es transversal a las etapas anteriormente descritas, por consiguiente, debe evidenciarse en ellas.

1. Debe existir un protocolo de seguridad para aquellos(as) que tienen contacto con los ítems, en el cual se expongan detalladamente las normas de protección y seguridad de la información confidencial relacionada con la prueba.
2. Quienes tengan acceso a los ítems (desarrolladores, revisores, jueces expertos, analistas de datos, equipo logístico y demás personal) deben firmar un compromiso o acuerdo de confidencialidad en el que certifiquen que conocen las normas de seguridad de la información confidencial y que se comprometen a cumplir con estas disposiciones.
3. La edición digital de la prueba debe realizarse en computadoras que contengan diversos dispositivos de seguridad, como el acceso restringido de usuario mediante contraseñas de seguridad para quienes estén autorizados a acceder a los ítems. De ser posible, estas computadoras no deben estar conectadas a ningún tipo de red local ni internacional.
4. La impresión de las pruebas debe llevarse a cabo en un espacio seguro que permita una debida monitorización de la cantidad de ejemplares impresos así como de quienes se encargan de realiza tal labor.
5. Los datos personales de los(as) examinados(as) constituyen información confidencial, por lo cual debe resguardarse con la misma rigurosidad con la que se protegen los ítems.
6. Es conveniente archivar la información relacionada con las respuestas de los(as) examinados(as) por un tiempo prudencial en caso de apelaciones o solicitudes

de revisión. Por ejemplo, en algunos casos se guardan los cuadernillos de prueba durante un año y las respectivas hojas de respuesta por dos años. Después de dicho periodo, lo recomendable es destruir dicho material de manera que la información sea ilegible.

7. Asegurar la información relacionada con la prueba y sus datos derivados (la que se considere pertinente) en caso de un eventual desastre (incendios, terremotos, inundaciones, entre otros). En este sentido, es importante respaldar periódicamente la información digital y guardarla en dispositivos de seguridad como cajas de seguridad o archivos anti-fuego.
8. Los archivos y hojas con las claves (respuestas correctas) deben permanecer en un lugar seguro.

II.11 Manual técnico

El manual técnico es un documento que debe ser transversal a las etapas expuestas, pues recopila todo aquello que se realiza en cada una de las etapas del proceso. A continuación unos lineamientos para que el lector conozca lo que este manual debe especificar.

1. Quienes elaboran pruebas no solo deben fundamentarse en bases científicamente sólidas para realizar su labor de una manera óptima, sino que también deben documentar adecuadamente las evidencias que fundamentan la creación, el desarrollo y la aplicación de la prueba. En este sentido, se debe brindar información actualizada y detallada de la prueba.
2. En el manual se debe describir la población meta y las especificaciones de la prueba. Si es pertinente, también conviene presentar los procedimientos para crear los ítems. Si se utilizaron muestras normativas, se debe describir la población con la cual se establecieron las normas y el año en el que dicho proceso ocurrió.
3. En dicho manual aparecerán las estadísticas sobre la confiabilidad de los puntajes y la validez de las inferencias hechas a partir de la prueba, cuando exista dicha información. Cuando sea relevante para interpretar la prueba, se

debe incluir información sobre los ítems, puntos de corte, puntajes brutos y derivados, datos normativos, errores estándar y una descripción de los procedimientos usados en la equiparación de los puntajes obtenidos en distintas versiones de la prueba.

4. Este documento debe incluir información referente a la evidencia recolectada que respalda la prueba y su utilidad para los potenciales usuarios(as), examinados(as) y autoridades o entes que emplean los resultados de la prueba.
5. Es conveniente también documentar los procedimientos empleados en la selección de la muestra de los estudios piloto, así como de las variables sociodemográficas de los(as) examinados(as) y algunas otras variables consideradas relevantes.
6. Si se estima una muestra representativa de la población meta, se debe describir a detalle la población y cualquier elemento que pudiera limitar la representatividad de la muestra. En general, la composición de cualquier muestra de la que se obtenga evidencias de validez debe describirse de manera detallada, así como la representatividad de esta (AERA, APA y NCME, 2014).
7. En el manual técnico debe presentarse una compilación de todo lo que se presenta en cada una de las etapas descritas a lo largo del presente documento.

II.12 Las adecuaciones en pruebas estandarizadas

Un tema ineludible a considerar para la creación de una prueba estandarizada es el de las adecuaciones que garantizan el acceso a dicha prueba por parte de personas con discapacidad o, con las así llamadas en contextos educativos, necesidades educativas especiales. En este sub-apartado se provee un marco de referencia general para la medición apropiada de un constructo dado, en el marco del principio de inclusión y el respeto por la diversidad intrínseca a cualquier grupo humano.

Como ya se discutió en apartados anteriores, cuando se habla de medición a través de la utilización de pruebas el concepto de la estandarización es fundamental para garantizar la confiabilidad de la prueba, así como la validez de las inferencias a realizar

a partir de los resultados para cada persona examinada y la equidad en el proceso. Sin embargo, a pesar de su importancia, pueden encontrarse ocasiones en las que la estandarización, en lugar de garantizar la rigurosa medición del nivel de una persona en el constructo de interés, interviene o atenta contra esta. Lo anterior se refiere a situaciones donde las condiciones específicas de las personas evaluadas requieran que se modifiquen en alguna medida las condiciones de aplicación, siempre que el formato de la prueba o las condiciones de aplicación no sean relevantes para la medición del constructo (AERA, APA & NCME, 2014).

Un ejemplo de lo expuesto en el párrafo anterior es que para una persona a quien se le dificulta marcar en una hoja para respuestas, debido a alguna deficiencia en la motricidad fina, gran parte del tiempo de aplicación se vería consumido por la actividad de hacer las marcas correspondientes a las respuestas elegidas, por lo que puede que no logre terminar la prueba en el tiempo "estándar". Esta persona, debido a que no alcanzó a terminar un número determinado de ítems, tendría un desempeño en la prueba que no reflejaría su nivel de habilidad real (ver apartado I.1.2.5).

Ketterlin-Geller y Johnstone (2006) definen las adecuaciones en contextos de medición como cambios en las instrucciones o en las prácticas de medición que reducen el impacto de la deficiencia de una persona en su interacción con el material de la prueba utilizada[1]. Las adecuaciones pueden incluir cambios en el contexto en el que las instrucciones son presentadas, en la cantidad de tiempo que se concede para completar una tarea, en el método de respuesta y en los materiales o el equipo que apoya la interacción con el material de la prueba (AERA, APA & NCME, 2014; Ketterlin-Geller y Johnstone, 2006), es decir, se consideran únicamente adecuaciones de acceso a la prueba.

Para que sean consideradas efectivas las adecuaciones deben reducir la varianza

[1] El término "adecuaciones curriculares" debe evitarse en el contexto de medición mediante pruebas estandarizadas, ya que, generalmente, la medición no se hace en función ni responde a una adaptación del curriculum. Se debe utilizar un término más apropiados como "adecuaciones de acceso" o "adecuaciones en las condiciones de aplicación" o, como se usa en este texto "adecuaciones".

irrelevante al constructo asociada a la deficiencia presente en la persona, sin afectar la medición de los contenidos o constructo (Georgia Department of Education, 2008; Johnson y Monroe, 2004; AERA, APA & NCME, 2014) y sin caer en la subrrepresentación de éste (medir menos de lo que se requiere medir). Así, las adecuaciones vienen a ser intentos por "nivelar el terreno de juego", de manera que todas las personas examinadas tengan las mismas posibilidades de mostrar su nivel de habilidad. Estas apuntan hacia la equidad, no hacia la ventaja para el grupo que las recibe, de manera que la medición de habilidades se realice con la debida precisión (Thurlow y Wiener, 2000; Abedi et al., 2012).

Cuando se trata de pruebas estandarizadas, para garantizar la validez de las inferencias derivadas de su uso se debe asegurar que ninguna de las adecuaciones ofrecidas comprometa la calidad técnica de la medición del constructo medido, es decir, que estas no afecten la dificultad de la prueba y que no se generen ventajas ni desventajas para las personas que las reciben, en comparación con el resto de examinados(as). Esta condición permite que los puntajes de la totalidad de la población examinada sean comparables entre sí y que se puedan hacer interpretaciones válidas acerca del nivel de cada persona examinada respecto de los contenidos o constructo que se desea medir (Messick, 1995; Cox, Herner, Demczyk y Nieberding, 2006). Un caso contrario es cuando a una persona examinada se le aplica una prueba a la que se le han eliminado contenidos o que tiene un nivel inferior de dificultad; esta persona obtendría una puntuación que no reflejaría su verdadero nivel de habilidad o de conocimientos.

El concepto de adecuaciones en las condiciones de aplicación de una prueba se puede confundir con otro que sí tiene implicaciones para la validez de las inferencias realizadas a partir de los resultados obtenidos por una persona: las modificaciones. Estas consisten en la incorporación de cambios que alteran o reducen las expectativas de lo que se mide mediante la prueba aplicada. Estas pueden aumentar la brecha entre el logro obtenido por personas con discapacidad y las expectativas sobre el desempeño del grupo total de personas examinadas.

Las modificaciones necesariamente implican algún impacto en la medición del constructo y, a diferencia de las adecuaciones, implican más que un cambio en el escenario de la medición (Stone, Cook, Laitusis y Cline, 2010), es por esto que las modificaciones deben evitarse en un contexto de medición mediante pruebas estandarizadas. Algunos ejemplos de modificaciones en el contexto de una evaluación incluyen bajar el nivel de los objetivos de evaluación, presentar una prueba con menos ítems, permitir a una persona completar solo los ítems más fáciles, presentar una prueba más fácil, reducir las opciones de respuesta en pruebas de selección única o dar pistas de las respuestas correctas (Georgia Department of Education, 2008).

De acuerdo con el *Georgia Department of Education* (2008) para cada persona evaluada deben otorgarse solamente las adecuaciones estrictamente necesarias para asegurar su acceso al especio o material de evaluación, ya que proveer adecuaciones no requeridas puede interferir e impactar de forma negativa en el desempeño y la medición. Además de lo anterior, las adecuaciones brindadas en procesos de evaluación deben adherirse a los siguientes principios:

- Las adecuaciones deben permitir la participación completa de la persona examinada en el proceso de evaluación, de manera que esta pueda demostrar de mejor manera su conocimiento o habilidades.
- Las adecuaciones deben basarse en las necesidades individuales de cada persona a examinar.
- Las adecuaciones deben justificarse y documentarse en cada caso, así como para cada persona a examinar.
- Las adecuaciones para la aplicación de la prueba deben estar en consonancia con las adecuaciones que la persona ha recibido en su proceso educativo previo, esto es, no deben introducirse por primera vez en el contexto de la evaluación.
- Las adecuaciones deben facilitar la independencia de las personas examinadas.

Además de lo anterior, la decisión sobre cuáles adecuaciones deben otorgarse para cada persona debe tomarse con base en una evaluación rigurosa de los requerimientos de acceso de cada una, en la cual se justifiquen objetivamente la oferta de adecuaciones aprobadas para una prueba determinada.

Por su parte, la *American Educational Research Association*, en conjunto con la *American Psychological Association* y el *National Council on Measurement in Educaction* han definido algunos estándares para garantizar la justicia y equidad a través de la aplicación de pruebas. Algunos de estos estándares, considerados como los más substanciales por los autores de este cuaderno metodológico, se resumen a continuación:

- Quienes tienen a cargo el desarrollo de una prueba deben asegurar que este instrumento minimice las posibilidades de que los resultados obtenidos por las personas examinadas sean afectados por características irrelevantes al constructo, sean estas de tipo lingüístico, comunicativo, cognitivo, cultural, físico o de otro tipo.
- Las personas a cargo del desarrollo de una prueba deben incluir en la muestra utilizada para los estudios piloto previos a la aplicación de la misma, la representación de subgrupos considerados relevantes, de manera que se descarte el funcionamiento diferencial de los ítems para estos conglomerados.
- La totalidad de personas examinadas deben recibir un tratamiento equivalente durante el proceso de aplicación y calificación de la prueba.
- Las personas a cargo del desarrollo de la prueba deben especificar y documentar las medidas que se han tomado para reducir las barreras irrelevantes al constructo para todos los subgrupos relevantes dentro de la población examinada.
- Las personas encargadas de desarrollar o aplicar la prueba deben ser responsables de examinar la evidencia que garantice la validez de las interpretaciones hechas a partir de los puntajes para cada subgrupo relevante dentro de la población examinada.

- Las personas a cargo del desarrollo o aplicación de una prueba son las responsables de desarrollar y proveer adecuaciones, siempre que estas sean apropiadas y factibles, de manera que estas remuevan las barreras irrelevantes al constructo que de otra manera hubiesen interferido con la correcta medición del mismo.
- Cuando se permitan las adecuaciones para una prueba dada, las personas a cargo del desarrollo o aplicación de la misma son los responsables de documentar las provisiones para el uso de las adecuaciones y de monitorear su correcta implementación.
- Siempre que una prueba sea modificada para remover barreras a la accesibilidad en la medición del constructo, las personas a cargo del desarrollo o aplicación de la prueba deben obtener y documentar la evidencia de la validez de las interpretaciones de los puntajes, siempre que los tamaños de muestra así lo permitan.
- Una prueba debe ser administrada utilizando el lenguaje que sea más relevante y apropiado a los propósitos de la misma.
- Los intérpretes de lengua de señas que medien las aplicaciones de pruebas a personas sordas, deben seguir procedimientos estandarizados.
- Al aplicar una prueba para propósitos de diagnóstico, los resultados obtenidos en esta no deben ser usados como un único indicador del funcionamiento del sujeto, sino que deben usarse múltiples fuentes de información para la toma de decisiones.

Una vez analizadas estas recomendaciones para la provisión de adecuaciones para la aplicación de una prueba, se debe valorar la oferta de adecuaciones que se va a proveer para la misma. En el siguiente apartado se presenta un compendio de posibles adecuaciones a ofrecer en pruebas estandarizadas.

II.12.1 Algunas adecuaciones ofrecidas para pruebas estandarizadas

Antes de definir las adecuaciones pertinentes para una prueba estandarizada, se debe evaluar en qué medida el formato de presentación y de respuesta de la prueba es crítico para las inferencias que se realizarán a raíz de los resultados obtenidos en la prueba por parte de una persona determinada (Cawthon, Winton, Garberoglio y Gobble, 2013). Por ejemplo, si en una prueba dada el componente de velocidad en el proceso de resolución de los ítems es parte del constructo o contenidos por medir, la concesión de tiempo adicional afectaría seriamente la medición del constructo, por lo que no sería una adecuación pertinente en ese contexto. Por el contrario, si la velocidad de respuestas no es parte de lo que se desea evaluar, sino una convención definida a partir del tiempo utilizado por la mayoría de personas que resuelven la prueba, el otorgamiento de tiempo adicional no debería afectar la medición.

La oferta de adecuaciones debe ajustarse a las características de cada prueba y justificarse a partir de la investigación sobre su utilidad y, como ya se mencionó, la evaluación de su impacto en la medición. También se deben tomar en cuenta los recursos tecnológicos disponibles que permitan contar con recursos novedosos para el acceso a la información contenida en la prueba.

Algunos ejemplos de adecuaciones usualmente otorgadas en la aplicación de pruebas estandarizadas incluyen:

- concesión de tiempo adicional para finalizar la prueba.
- adaptación de los folletos de prueba a formatos accesibles (por ejemplo en letra ampliada o impresión en Braille).
- cambios en los formatos de respuesta (tales como responder en el folleto o contar con el apoyo de un escribiente que recoja las respuestas de la persona examinada).
- periodos de descanso a lo largo de la prueba.
- ubicación específica en el recinto de aplicación (por ejemplo: primeros asientos o cerca de la puerta).

- apoyo de intérpretes de lenguaje de señas.
- aplicación de la prueba en espacios accesibles.
- mobiliario adaptado de acuerdo con su necesidad corporal.
- permitir el uso de apoyos técnicos (lámparas, lupas, entre otros).
- apoyo de una persona que funja como lector o uso de lectores de pantalla.
- aplicaciones individuales o en grupo más pequeños de lo usual.

El uso de la calculadora como adecuación durante la aplicación de una prueba puede considerarse en el caso de personas con discalculia[2] o a personas con ceguera que, por su condición, se les dificulta hacer las anotaciones necesarias para realizar cálculos extensos. Sin embargo, la pertinencia de esta adecuación debe valorarse especialmente en el caso de pruebas de corte matemático, ya que de acuerdo con Scarpati, Wells y Lewis (2013) en ciertos casos la utilización de una calculadora para resolver problemas matemáticos puede ser confundida con la habilidad por medir. Además, estos autores indican que en la resolución de ítems de contenido matemático el uso de calculadora como una adecuación beneficia al aspirante con niveles bajos de habilidad, ya que les facilita la resolución de los ítems más fáciles, mientras que no necesariamente ejerce el mismo efecto en aspirantes con niveles altos de habilidad, lo cual se interpreta como una interferencia en la medición.

Otra adecuación que debe ser valorada de acuerdo con la finalidad de la prueba y la posible afectación de la medición del constructo es el uso del diccionario en el caso de pruebas con ítems compuestos por texto escrito. En el caso de que el uso del diccionario afecte la correcta medición del constructo, este debe permitirse únicamente para aspirantes con condición de sordera que no tengan un uso instrumental adecuado de la lengua oficial oral y escrita, debido a una condición de sordera prelocutiva[3].

Esta adecuación se justifica para personas con sordera prelocutiva debido a la

[2] Trastorno de aprendizaje caracterizado por una dificultad para asimilar y recordar datos numéricos y aritméticos, para realizar procedimientos de cálculo y crear estrategias para la solución de problemas (Roselli y Matute, 2011).

[3] Personas que presentaron sordera antes de la adquisición del lenguaje oral.

presencia de dificultades en el desarrollo de la capacidad de comprensión lectora y expresión escrita. Cabe señalar que dicha adecuación no se suele justificar para personas con otros tipos de deficiencias que no afectan el acceso al lenguaje ya que el uso de un diccionario otorgaría una ventaja en comparación con personas que no reciben esta adecuación y que pueden tener un vocabulario limitado o una baja habilidad para la comprensión lectora, tal como lo explica Phillips (1994) en relación con el otorgamiento de adecuaciones no específicas al área de afectación de la persona.

II.12.2 Actividades esenciales para la provisión de adecuaciones para una prueba estandarizada

Una vez definidas las adecuaciones por ofrecer para la prueba, se debe iniciar con el proceso que llevará a la implementación de estas. Dicha tarea involucra una serie de actividades que se describen grosso modo en los párrafos siguientes.

El proceso de provisión de adecuaciones para una prueba estandarizada debe iniciarse con la adaptación de todos los materiales relativos a la prueba. Esto incluye la elección de los ítems que contendrán los materiales (tanto de práctica como definitivos) de acuerdo con una clasificación de estos como aptos para personas con deficiencia visual o con condición de sordera, y la adaptación de estos a formatos accesibles.

Para aquellas personas que requieren letra ampliada o lector de pantalla, es necesario eliminar de los documentos los ítems con imágenes, tablas o gráficos. Si lo que se pretende es una “conversión” de los ítems con material gráfico a prosa, una vez hecho esto debe revisarse el texto final del ítem para asegurar que sea comprensible para la población meta. Lo más recomendable es hacer una aplicación piloto de estos ítems con personas que presentan ceguera o muy baja visión o pedir a una persona con esta condición que haga un “juzgamiento” de la accesibilidad del ítem.

Para personas que requieran de materiales en formato braille, los originales deben transcribirse a dicho formato. Esto se puede hacer mediante paquetes informáticos como Duxbury®. Una vez transcritos los materiales se deben revisar todas las expresiones matemáticas y corregirlas para que sean equivalentes a las del texto original.

Para personas con otros tipos de deficiencia visual, el texto de la prueba debe ser ampliado a cada uno de los tamaños de letra ofrecidos como adecuación. Luego de esto, el original de cada una de las versiones ampliadas debe ser revisado para verificar que no haya palabras cortadas al final de los renglones y que las expresiones matemáticas queden de forma correcta en el nuevo tamaño. Asimismo, se debe verificar que las versiones ampliadas contengan las mismas palabras “en negrita” o subrayadas presentes en la fórmula original. En el caso de las fórmulas adaptadas con la totalidad del texto en “negrita", se sugiere verificar que las palabras que se encontraban en “negrita” en el texto original, se encuentren subrayadas en la nueva versión del texto y que el subrayado original se mantenga.

Por otra parte, para personas que requieren lector de pantalla, las expresiones matemáticas deben plantearse en palabras (sin símbolos matemáticos ni numerales) de manera que el lector de pantalla “verbalice” de forma correcta dichas expresiones.

En el caso de las instrucciones de aplicación que se comunican a las personas examinadas al inicio de la aplicación, debe asegurarse también su idoneidad y accesibilidad para todas las personas. Así, al igual que con los folletos de práctica y los folletos de prueba, se deben adaptar las instrucciones regulares e incluir los ajustes necesarios para que se adapten a los requerimientos específicos de las personas con adecuaciones aprobadas, así como a las particularidades del grupo de aplicación. Estos ajustes referentes a las particularidades grupales se deben hacer en lo referente a tiempos de aplicación, los recursos para acceso del texto (por ejemplo: lector, letra ampliada y braille) y el modo de respuesta, entre otros.

Antes de la aplicación de las pruebas a poblaciones con adecuaciones, es imprescindible asegurar la capacitación de las personas que fungirán como aplicadores. La capacitación debe incluir la atención aspectos que diferencian estas aplicaciones de las regulares y la aclaración de dudas.

En caso de pruebas que se apliquen en periodos definidos (anualmente, semestralmente, etc.) una vez concluido un periodo de aplicaciones de pruebas, es importante realizar análisis descriptivos de los datos referidos a tiempos de aplicación empleados por los(as) examinados(as), de frecuencias de las adecuaciones otorgadas, entre otros, para fundamentar decisiones futuras.

No debe omitirse la realización de un análisis cualitativo del proceso para generar recomendaciones que mejoren las próximas aplicaciones. Asimismo, es conveniente llevar a cabo análisis estadísticos para estimar si el tiempo adicional asignado es adecuado de acuerdo con los resultados obtenidos y con la utilización del tiempo por parte de los(as) examinados(as).

Finalmente, deben revisarse a la luz de los hallazgos de los análisis, las adecuaciones asignadas durante el proceso anterior. Se debe realizar una propuesta de adecuaciones para el periodo de aplicación siguiente, que incluya la modificación o ajuste de las adecuaciones ya ofrecidas o la implementación de nuevas adecuaciones que sean pertinentes, de acuerdo con los avances tecnológicos o con los requerimientos de la población.

PARTE III: COMENTARIOS FINALES

En este trabajo se han expuestos algunas recomendaciones sobre cómo construir pruebas estandarizadas en el ámbito educativo en el marco de las nociones contemporáneas sobre la validez y la confiabilidad. Cabe señalar que desde esta perspectiva, la sección correspondiente al tema de las adecuaciones es parte consustancial de la validez. La implementación de las adecuaciones en las pruebas

estandarizadas, no debe considerarse como un aspecto separado del proceso de desarrollo de un instrumento de medición. Al contrario, es imprescindible plantear este tema desde el inicio del mismo de pruebas estandarizadas.

Asimismo, una vez expuestas las etapas para desarrollar un test estandarizado, es importante realizar una evaluación general de su calidad técnica. Para ello, existen diferentes modelos que podrían utilizarse, sin embargo, se recomienda al lector consultar el trabajo de Evers et al. (2013), dado que en este se expone de una manera muy sintética y clara el modelo revisado de la Federación Europea de Asociaciones de Psicólogos (EFPA) para evaluar pruebas.

Finalmente, es importante enfatizar la naturaleza panorámica de este documento. Por ello, a lo largo del documento se refiere al lector a distintas fuentes bibliográficas para profundizar en los aspectos esbozados. Asimismo, es importante mencionar que este trabajo se enfoca en los aspectos prácticos del desarrollo de pruebas, dejando por fuera la explicación de los modelos psicométricos que subyacen a la tarea de elaborar este tipo de pruebas.

REFERENCIAS BIBLIOGRÁFICAS

Abedi, J., Bayley, R., Ewers, N., Mundhenk, K., Leon, S., Kao, J. & Herman, J. (2012). Accessible Reading Assessments for Students with Disabilities. *International Journal of Disability, Development and Education*, *59* (1), 81-95.

AERA, APA & NCME (2014). *Standards for Educational and Psychological Testing*. Washington, Estados Unidos: American Educational Research Association.

Albertini, J., Kelly, R. & Matchett, M. (2012). Personal factors that influence deaf college students. *Journal of Deaf Studies and Deaf Education*, *17*(1), 85-101.

Asenjo, M. (2010). El alumnado con discapacidad auditiva en el aula ordinaria. *Revista Digital Eduinnova*, *27*, 2-8.

Beavers, A., Lounsbury, J., Richards, J., Huck, S., Skolits, G. & Esquivel, S. (2013). Practical Considerations for using Exploratory Factor Analysis in Educational Research. *Practi cal Assessment, Research & Evaluation*, *18*(6). Recuperado de http://pareonline.net/pdf/v18n6.pdf

Bollen, K. y Hoyle, R. (2012). Latent Variables in Structural Equation Modeling. En R. Hoyle (Ed.), *Handbook of Structural Equation Modeling,* (pp. 56-67). Estados Unidos: The Guilford Press.

Borsboom, D., Mellenbergh, G. & Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061-1071.

Brown, T. (2006). *Confirmatory Factor Analysis for Applied Research.* New York, Estados Unidos: The Guilford Press.

Cahalan, C., King, J., Cline, F. & Bridgeman, B. (2006). *Observational timing study on the SAT Reasoning Test for test-takers with learning disabilities and/or ADHD* (College Board Research Report No. 2006-4). Estados Unidos: College Board.

Cahalan, C., Morgan, D. L., Bridgeman, B., Zanna, J., & Stone, E. (2007). *Examination of fatigue effects from extended-time accommodations on the SAT Reasoning Test* (College Board Research Report No. 2007-1). New York, NY: The College Board.

Calhoon, M., Fuchs, L. & Hamlett, C. (2000). Effects of Computer-Based Test Accommodations on Mathematics Performance Assessmentsfor Secondary Students with Learning Disabilities. *Learning Disability Quarterly*, *23*(4), 271-282.

Campbell, D. & Fiske, D. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, *56*(2), 81-105.

Cawthon, S., Winton, S., Garberoglio, C. & Gobble, M. (2013). The Effects of American Sign Language as an Assessment Accommodation for Students Who Are Deaf or Hard of Hearing. *Journal of Deaf Studies and Deaf Education*, *16*(2), 198-211.

Cizek, G. (2012). Defining and Distinguishing Validity: Interpretations of Score Meaning and Justifications of Test Use. *Psychological Methods*, *17*(1), 31-43.

Coaley, K. (2010). *An Introduction to Psychological Assessment and Psychometrics*. Reino Unido: SAGE.

Cohen, A., Gregg, N. & Deng, M. (2005). The Role of Extended Time and Item Content on a High-Stakes Mathematics Test. *Learning Disabilities Research & Practice*, *20*(4), 225-233.

Costello, A. & Osborne, J. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation*, *10*(7). Recuperado de http://pareonline.net/pdf/v10n7.pdf

Cox, M., Herner, J., Demczyk, M. & Nieberding, J. (2006). Provision of Testing Accommodations for Students With Disabilities on Statewide Assessments: Statistical Links with Participation and Disciplines Rates. *Remedial and Special Education*, *27*(6), 346-354.

Craddock, D. & Mathias, H. (2009). Assessment Options in Higher Education. *Assessment & Evaluation in Higher Education*, *34*(2), 127-140.

Crocker, L. & Algina, J. (2006). *Introduction to Classical and Modern Test Theory*. Ohio, Estados Unidos: CENGAGE Learning.

Cronbach, L. & Meehl, P. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*(4), 281-302.

DeVellis, R. (2013). *Scale Development*. California, Estados Unidos: SAGE.

Downing, S. & Haladyna, T. (2006). *Handbook of Test Development*. Estados Unidos: Lawrence Erlbaum Associates, Publishers.

Dwyer, C., Gallagher, A., Levin, J. & Morley, M. (2003). What is Quantitative Reasoning? Defining the Construct for Assessment *Purposes* (Research Reports). Princeton, NJ: Educational Testing Service.

Elosua, P. (2003). Sobre la validez de los test. *Psicothema*, 15(2), 315-321.

Embretson, S. (1983). Construct Validity: Construct Representation Versus nomothetic Span. *Psychological Bulletin*, *93*(1), 179-197.

Embretson, S. (2010). *Measuring Psychological Constructs: Advances in Model-Based Approches*. Estados Unidos: American Psychological Association.

Ericsson, K., & Simon, H. (1993). *Protocol Analysis: Verbal Reports as Data*. Estados Unidos: MIT Press.

Educational Testing Service. (2002). *ETS Standards for Quality and Fairness*. Estados Unidos: Educational Testing Service.

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

Ferrando, P. & Lorenzeo-Seva, U. (2014). El análisis factorial exploratorio de los ítems: algunas consideraciones adicionales. *Anales de psicología*, *30*(3), 1170-1175.

Furr, M. & Bacharach, V. (2013). *Psychometrics: An Introduction*. California, Estados Unidos: SAGE.

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

Gorin, J. (2006). Test Design with Cognition in Mind. *Educational Measurement: Issues and Practice*, *25*(4), 21-35.

Haladyna, T. & Rodriguez, M. (2013). *Developing and Validating Test Items*. New York, Estados Unidos: Routledge.

Haladyna, T. & Downing, S. (2004). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hidalgo, M., Gómez, J. & Padilla, J. (2005). Regresión logística: alternativa de análisis en la detección del funcionamiento diferencial del ítem. *Psicothema*, *17*(3), 509-515.

Izquierdo, I., Olea, J. & Abad, F. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema*, *26*(3), 394-400.

Jackson, D., Gillaspy, J. & Purc-Stephenson, R. (2009). Reporting Practices in Confirmatory Factor Analysis: An Overview and Some Recommendations. *Psychological Methods, 14*(1), 6-23.

Johnson, E. & Monroe, B. (2004). Simplified Language as an Accommodation on Math Tests. *Assessment for Effective Intervention*, *29*(3), 35-45.

Kane, M. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, *38*(4), 319-342.

Kane, M. (2006). Content-Related Validity Evidence in Test Development. En S. Downing y T. Haladyna (Eds.), *Handbook of Test Development* (pp. 131-153). New Jersey, Estados Unidos: Lawrence Erlbaum Associates, Inc.

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kenny, D. (1979). *Correlation and Causality.* Nueva York, Estados Unidos: John Wiley & Sons Inc.

Ketterlin-Geller, L. & Johnstone, C. (2006). Accommodations and Universal Design: Supporting Access to Assessments in Higher Education. *Journal of Postsecondary Education and Disability*, *19*(2), 163-172.

Kline, R. (2013). *Principles and Practice of Structural Equation Modeling*. Estados Unidos: The Guilford Press.

Kumar, N. (2009). *Applied Psychometry*. India: SAGE.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.

Laitusis, C. (2010). Examining the Impact of Audio Presentation on Tests of Reading Comprehension. *Applied Measurement in Education*, *23*(2), 153-167. DOI: 10.1080/08957341003673815

Lane, S. (2014). Validity evidence based on testing consequences. *Psichotema*, *26*(1), 127-135.

Lee, K., Osborne, R. & Carpenter, D. (2010). Testing Accommodations For University Students With AD/HD: Computerized vs. Paper-Pencil/Regular vs. Extended Time. *Journal of Educational Computing Research*, *42*(4), 443-458.

Leighton, J. (2004). Avoiding Misconception, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice*, *23*(4), 6-15.

Leighton, J. & Gierl, M. (2007). Verbal Reports as Data for Cognitive Diagnostic Assessment. En J. Leighton y M. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 146-172). New York, Estados Unidos: Cambridge University Press.

Lesaux, N., Pearson, M. & Siegel, L. (2006). The Effects of Timed and Untimed Testing Conditions on The Reading Comprehension Performance of Adults with Reading Disabilities. *Reading and Writing*, *19*, 21-48.

Lewandowski, L., Cohen, J. & Lovett, B. (2013). Effects of Extended Time Allotments on Reading Comprehension Performance of College Students with and Without Learning Disabilities. *Journal of Psychoeducational Assessment*, *31*(3), 326-336.

Lloret, S., Ferreres, A., Hernández, A. & Tomás, I. (2014). El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. *Anales de Psicología*, *30*(3), 1151-1169.

Mandinach, E., Bridgeman, B., Cahalan, C. & Trapani, C. (2005*). The Impact of Extended Time on SAT® Test Performance* (College Board Research Report No. 2005-8 ETS RR-05-20). New York, NY: The College Board.

Markus, K. & Borsboom, D. (2013). Frontiers of Test Validity Theory: Measurement, Causation, and Meaning. New York, Estados Unidos: Routledge.

Martínez, M., Hernández, M. & Hernández, M. (2006). *Psicometría*. España: Alianza Editorial.

Martínez, R. (1996). *Psicometría: Teoría de los tests psicológicos y educativos.* España: Editorial Síntesis.

McDonald, R. (1999). *Test Theory: A Unified Treatment*. Estados Unidos: Lawrence Erlbaum.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*(10), 955-966.

Messick, S. (1980). Test Validity and the Ethics of Assessment. *American Psychologist*, *35*(11), 1012-1027.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. American Psychologist, 50 (9), 741-749.

Milsap, R. & Olivera, M. (2012). Investigating Measurement Invariance Using Confirmatory Factor Analysis. En R. Hoyle (Ed.), *Handbook of Structural Equation Modeling,* (pp. 380-391). Estados Unidos: The Guilford Press.

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

Morales, P. (2013). El Análisis Factorial Exploratorio en la construcción e interpretación de tests, escalas y cuestionarios. Recuperado de http://web.upcomillas.es/personal/peter/investigacion/AnalisisFactorial.pdf

Moreno, R., Martínez, R. & Muñiz, J. (2006). New Guidelines for Developing Multiple-Choice Items. Methodology, 2(2), 65-72.

Muñiz, J. (2001). *Teoría clásica de los test*. España: Ediciones Pirámide.

Muñiz, J. & Fonseca, E. (2008). Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación Educativa*, 5, 13-25.

Padilla, J., Gómez, J., Hidalgo, M. & Muñiz, J. (2006). La evaluación de las consecuencias del uso de los tests en la teoría de la validez. *Psicothema*, 18(2), 307-312.

Pedhazur, E. & Pedhazur, L. (1991). *Measurement, Design, and Analysis.* New Jersey, Estados Unidos: Lawrence Erlbaum Associates, Publishers.

Phillips, S. (1994). High-Stakes testing Accommodations: Validity Versus Disabled Rights. *Applied Measurement in Education*, *7*(2), 93-120.

Raykov, T. (2012). Scale Construction and Development Using Structural Equation Modeling. En R. Hoyle (Ed.), *Handbook of Structural Equation Modeling,* (pp. 472-492). Estados Unidos: The Guilford Press.

Raykov, T. & Marcoulides, G. (2013). *Introduction to Psychometric Theory.* Estados Unidos: Routledge.

Rios, J. & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116.

Ross, T. & Rowley, G. (1991). Understanding Reliability. *ITEMS: Instructional Topics in Educational Measurement*. Recuperado de http://ncme.org/publications/items/

Rosselli, M. & Matute, E. (2013). La neuropsicología del desarrollo típico y atípico de las habilidades numéricas*. Revista neuropsicología, neuropsiquiatría y neurociencias*, *11* (1), 123-140.

Roussos, L., DiBello, L., Stout, W., Hartz, S. Henson, R. & Templin, J. (2007). The Fussion Model Skills Diagnosis System. En J. Leighton y M. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 146-172). New York, Estados Unidos: Cambridge University Press.

Rubio, F. (2009). Los alumnos/as con discapacidad auditiva. *Revista Innovación y Experiencias*, *24*, 1-13.

Rupp, A., Templin, J. & Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Estados Unidos: The Guilford Press.

Scarpati, S., Wells, C. & Lewis, C. (2013). Accommodations and Item-Level Analyses Using Mixture Differential Item Functioning Models. *Journal of Special Education*, *45*(1), 54-62.

Schmeiser, C. & Welch, C. (2006). Test development. En R. Brennan (Ed.), *Educational measurement* (pp. 307-353). Estados Unidos: Rowman & Littlefield Publishers.

Schreiber, J., Stage, F., King, J., Nora, A. & Barlow, E. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research, 99*(6), 323-337.

Sireci, S. (1998a). The construct of content validity. *Social Indicators Research*, *45*, 83-117.

Sireci, S. (1998b). Gathering and Analyzing Content Validity Data. *Educational Assessment*, *5*(4), 299-321.

Sireci, S. & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psichotema*, *26*(1), 100-107.

Smith, G., Fischer, S. & Fister, S. (2003). Incremental Validity Principles in Test Construction. *Psychological Assessment*, *15*(4), 467-477.

Smith, V. & Molina, M. (2013). *La entrevista cognitiva: Guía para su aplicación en la evaluación y mejoramiento de instrumentos de papel y lápiz*. San José, Costa Rica: Instituto de Investigaciones Psicológicas.

Stone, E., Cook, L., Laitusis, C. & Cline, F. (2010). Using Differential Item Functioning to Investigate the Impact of Testing Accommodations on an English-Language Arts Assessment for Students who are Blind or Visually Impaired. *Applied Measurement in Education*, *23*(2), 132-152.

Suárez, J. & Castellano, J. (2013). Atención educativa a personas ciega y con baja visión en la Sierra de Cádiz. *Clave XXI Reflexiones y Experiencias en Educación*, *6*, 1-16.

Tate, R. (2003). A Comparison of Selected Empirical Methods for Assessing the Structure of Response to Test Items. *Applied Psychological Measurement*, *27*(3), 159-203.

Thurlow, M. (2010). Steps Toward Creating Fully Accesible Reading Assessments. *Applied Measurement in Education*, 23(2), 121-131. DOI: 10.1080/08957341003673765

Thurlow, M. y Wiener, D. (2000). *Non-approved accommodations: Recommendations for use and reporting* (Policy Direction N0. 11). Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes.

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

Urbina, S. (2004). *Essentials of Psychological Testing*. New Jersey, Estados Unidos: John Wiley & Sons, Inc.

Van Hoogmoed, A., Knoors, H., Schreuder, R., Verhoeven, L. (2013). Complex word reading in Dutch deaf children and adults. *Research in Developmental Disabilities*, 34(3), 1083-1089. DOI: 10.1016/j.ridd.2012.12.010

Von Davier, M. & Khorramdel, L. (2013). Differentiating Response Styles and Construct-Related Responses: A New IRT Approach Using Bifactor and Second-Order Models. En R. Millsap, L. van der Ark, D. Bolt y C. Woods. (Eds.), *New Developments in Quantitative Psychology* (pp. 463-487). Estados Unidos: Springer.

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

Wise, L. (2010). Accessible Reading Assessments for Students with Disabilities: Summary and Conclusions. *Applied Measurement in Education*, 23(2), 209–214. DOI: 10.1080/08957341003673849

Woehr, D., Putka, D. & Bowler, M. (2012). An Examination of G-Theory Methods for Modeling Multitrait-Multimethod Data: Clarifying Links to Construct Validity and Confirmatory Factor Analysis. *Organizational Research Methods*, *15*(1), 134-161.