Actualidades en Psicología,
27 (114), 2013, 113-128
ISSN 0258-6444

Referentes conceptuales y metodológicos sobre la noción moderna de validez de instrumentos de medición: implicaciones para el caso de personas con necesidades educativas especiales

Conceptual and methodological referents regarding the modern notion of validity in measurement instruments: implications for the evaluation of people with special educational needs

Eiliana Montero Rojas
Escuela de Estadística, Instituto de Investigaciones Psicológicas
Universidad de Costa Rica
eiliana.montero@ucr.ac.cr / eilianamontero@gmail.com

Correo postal: 11501-2060 San José, Costa Rica

Resumen. Se hace un recorrido en torno al concepto moderno de validez, enfocado en pruebas psicológicas y educativas, con el propósito de presentar una discusión actualizada y brindar herramientas conceptuales y metodológicas a los constructores y usuarios de instrumentos. En cuanto a la validez se indican las importantes contribuciones de Samuel Messick, incluyendo la noción de que se trata de un concepto unitario, referido al grado de propiedad de las interpretaciones e inferencias realizadas a partir de los puntajes del instrumento. Se mencionan los modelos de medición más utilizados, la Teoría Clásica de los Tests (TCT) y Teoría de Respuesta a los ítems (TRI), incluyendo los modelos de Rasch. Se introducen las temáticas de DIF (funcionamiento diferencial de los ítems) y equiparación de puntajes (equating). Finalmente, se analizan implicancias de estos referentes en términos de las adecuaciones que se deben realizar al aplicar los instrumentos a personas con necesidades educativas especiales.
Palabras clave: validez, confiabilidad, instrumentos de medición, pruebas psicológicas, pruebas educativas, necesidades educativas especiales.

Abstract. A general look is taken around the modern concept of validity, with a special focus in psychological and educational tests, with the purpose of presenting an updated discussion and providing conceptual and methodological tools to test developers and users. Regarding validity, the important contributions of Samuel Messick are highlighted, including the notion that establishes that this is a unitary concept, referred to the degree of appropriateness of the interpretations and inferences that are drawn from test scores. The measurement models better known and used currently are mentioned, the Classical Test Theory (CTT), and Item Response Theory (IRT), including the Rasch Model. The topics of DIF (Differential Item Functioning) and equating are also introduced. Finally, some implications of this frame of reference are analyzed in terms of the accommodations that are necessary to implement for persons with special educational needs.
Key Words: Validity, reliability, measurement instruments, psychological tests, educational tests, special educational needs.

Introducción

Necesidad de actualizar el concepto de validez entre los constructores y usuarios de instrumentos

El propósito principal de este artículo es presentar una discusión actualizada a nivel conceptual y metodológico en torno al tema de la validez de instrumentos de medición. Y es que según la experiencia de esta autora, entre los usuarios y constructores de exámenes y escalas, tanto en Educación como en Psicología, existen aún confusiones y vacíos importantes en cuanto a los referentes teóricos para este concepto y también en lo que se refiere a los procesos de validación a que debe someterse un instrumento, particularmente cuando se consideran técnicas analíticas relativamente nuevas como el análisis factorial confirmatorio y los modelos de Rasch.

Incluso textos considerados "clásicos", disponibles en español y todavía muy utilizados al día de hoy, como el de Kerlinger y Lee (2002) definen la validez como una propiedad que nos indica si estamos midiendo lo que creemos que estamos midiendo (p.604). De igual forma, se refieren a la existencia de diferentes tipos de validez.

Las nociones expresadas en el párrafo anterior sintetizan aspectos clave de un marco de referencia que en la comunidad internacional de medición se considera ya anacrónico y que se utilizó con anterioridad a la propuesta de Samuel Messick (1989a, 1989b), que representa, sin duda un referente científicamente más sólido y enfoque más integral que los conceptos existentes previamente y que, además, conlleva todo un conjunto de implicaciones y consecuencias prácticas en torno a los procesos metodológicos involucrados en la construcción y validación de instrumentos.

Aunque desde el año 1999 se incorporó este marco de referencia a una de las publicaciones más influyentes en cuanto a criterios para valorar la calidad de las pruebas en Educación y Psicología, los "Standards for Educational and Psychological Testing" (AERA, APA, NCME, 1999), tal parece que en nuestra región esa actualización no ha permeado suficiente y de manera adecuada, como se explicó arriba. De ahí la intención de escribir este reseña, dirigida a constructores y usuarios de instrumentos que no son especialistas en medición y psicometría, con el objeto de que puedan obtener orientaciones básicas clave a nivel conceptual y metodológico en cuanto al tema de la validación de instrumentos, desde un enfoque moderno y actualizado.

Un test o instrumento de medición es un medio empírico que permite generar puntuaciones en una escala numérica para representar una variable o constructo (Nunnally & Bernstein, 1995). En las ciencias físicas podemos ilustrar este concepto indicando ejemplos como el de la regla y la balanza, los cuales se usan, respectivamente, para aproximar los constructos de longitud y masa. De igual manera, en las ciencias de la salud se utilizan exámenes médicos

para medir y diagnosticar diversas condiciones. En psicología, educación y ciencias sociales, en general, contamos con escalas psicométricas que nos permiten aproximar constructos tan complejos como habilidades intelectuales, conocimientos, rasgos de personalidad y otros atributos no directamente observables (Bond & Fox, 2001; Martínez et al, 2006).

Los puntajes generados por un instrumento pueden estar en un nivel de medición categórico (también llamado nominal), ordinal, intervalo o razón (Hopkins et al, 1997). En este punto es relevante indicar que, a pesar de que usualmente las tratamos como medidas de intervalo, la manera habitual en que generamos los puntajes para instrumentos o pruebas psicométricas, usando, por ejemplo, la suma de las puntuaciones obtenidas en los ítems, indicador derivado de la Teoría Clásica de los Tests (TCT), rinde una medida ordinal del constructo, pues no es posible garantizar que diferencias iguales en la puntuación del instrumento representen diferencias iguales en el constructo (Prieto & Delgado, 2003; Bond & Fox, 2001).

Finalmente, en el caso de las mediciones de razón o proporción existe un cero absoluto y por tanto se pueden realizar interpretaciones multiplicativas, como "A posee el doble que B" en el atributo de interés, tal es el caso de medidas tradicionales de tiempo, longitud y volumen. Parece poco probable que para la mayoría de los constructos con que se trabaja en ciencias sociales se llegue en un futuro cercano a obtener mediciones que permitan este tipo de interpretaciones, teniendo que conformarnos generalmente con niveles ordinales, y de intervalo, en el mejor de los casos (Bond & Fox, 2001).

Medición en educación y psicología

Dado que precisamente los problemas y temas de interés en educación y psicología abarcan la definición y medición de constructos complejos tales como rendimiento académico, habilidad intelectual, conocimientos, actitudes, valores y rasgos de personalidad, es necesario utilizar los conceptos y herramientas de la psicometría y diversos modelos de análisis que permitan generar indicadores empíricos del grado de validez y confiabilidad de las mediciones (Martínez et al, 2006; Nunnally & Bernstein, 1995).

Desde este enfoque se define el constructo como una conceptualización que requiere de un marco teórico explícito para definirse y operacionalizarse (Babbie, 2010). Otros ejemplos de constructos son desarrollo humano, violencia, educación, inteligencia y ciertamente el tema de medición de las pruebas de selección para la Universidad, la aptitud académica.

Por su parte, una variable es también una conceptualización, pero no requiere de un marco teórico explícito para definirse y operacionalizarse (Babbie, 2010). En general, hay más consenso en cuanto a su definición y medición. Ejemplos de variables son la edad, el número de hijos y la zona de residencia (urbana, rural).

Finalmente, generamos indicadores para aproximar empíricamente las variables o los constructos y poder medirlos (Babbie, 2010). Así, el indicador es el resultado de una operación de medición para representar una variable o constructo. El conocido Índice de Desarrollo Humano, definido y calculado por el Programa de las Naciones Unidas para el Desarrollo, es uno de los indicadores de desarrollo humano más utilizados y divulgados.

Parece evidente la necesidad de generar mediciones de constructos en educación y psicología. Para empezar, los problemas de investigación más interesantes y relevantes involucran la medición de constructos o variables latentes. Así, los constructos son la base para la generación de teorías, son el "pan de cada día de la ciencia". Por otra parte, es necesario medir constructos de forma válida para alimentar la toma de decisiones, tal es el caso de seleccionar, entre los aspirantes de primer ingreso a la Universidad, a aquellos que poseen el perfil que permita un adecuado desempeño académico en la Universidad, o bien, como en el caso de pruebas de certificación académica, evidenciar de manera objetiva los logros de aprendizaje para un programa educativo específico.

Los constructos, en general, son difíciles de operacionalizar y de medir. Los procedimientos para lograr mediciones válidas y confiables no son obvios. De ahí que se debieron estructurar enfoques que permitieran desarrollar metodologías para la medición de los constructos. La Psicometría es, sin lugar a dudas, una de las propuestas científicas más exitosas en términos de brindar herramientas útiles para emprender esta tarea de medición de constructos en educación y psicología (Martínez et al, 2006; Nunnally & Bernstein, 1995).

La psicometría es un cuerpo de teoría y métodos para la medición de constructos psicológicos y sociales. Uno de sus propósitos principales es el desarrollo de técnicas de aplicación empírica que permitan construir instrumentos de medición, indicadores, de alta confiabilidad y validez. Estas técnicas y métodos se basan en enfoques cuantitativos y utilizan conceptos, procedimientos y medidas derivadas de la estadística y la matemática. (Martínez et al, 2006; Nunnally & Bernstein, 1995).

Las raíces de los métodos psicométricos se remontan a finales del siglo XIX y principios del XX y a los primeros intentos por definir y aproximar empíricamente la inteligencia. Entre los pioneros podemos mencionar al inglés Charles Spearman, quien hizo sustanciales contribuciones a la psicología y a la estadística (Muñiz, 2003).

Un instrumento psicométrico intenta representar al constructo por medio de un puntaje numérico derivado de la aplicación de un conjunto de reactivos o estímulos a las unidades o elementos de interés. En su forma más usual está compuesto por una serie de ítems o reactivos, cada uno de los cuales es calificado o respondido por el individuo de acuerdo con una cierta escala de medición. El puntaje en el instrumento es una medida compuesta (empírica o estimada mediante un modelo estadístico-matemático) que se genera a partir de las puntuaciones individuales para cada ítem. Ese puntaje es el indicador del

nivel que toma el constructo de interés en cada uno de los elementos estudiados. (Martínez et al, 2006; Muñiz, 2003).

Validez y Confiabilidad

Las dos propiedades fundamentales de una "buena" medición son la validez y la confiabilidad (Nunnally & Bernstein, 1995; AERA et al, 1999; Martínez et al 2006).

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

El artículo seminal de Messick, publicado en la revista Educational Researcher en 1989 se tituló "Meaning and values in test validation: The science and ethics of assessment" (Significado y valores en la validación de pruebas: la ciencia y la ética de la evaluación). Este trabajo provocó la escritura de cientos de obras y textos que discuten, presentan, interpretan o critican a Messick, desde diversas ópticas.

Desde nuestra perspectiva las mayores contribuciones de Messick (1989a, 1989b) incluyen su definición de validez como un concepto unitario, misma que fue adoptada formalmente en los Standards for Educational and Psychological Testing, publicación conjunta de la AERA (American Educational Research Association), APA (American Psychological Associaton) y NCME (National Council on Measurement in Education), y que puede considerarse el "ISO 9000" internacional en cuanto a estándares de calidad de las pruebas educativas y psicológicas.

Así, en vez de hablar de diferentes tipos de validez, Messick (1989a) indica que la idea es recolectar diferentes tipos de evidencias, de acuerdo con los propósitos y usos de los instrumentos, entre ellas evidencias de contenido, predictivas y concurrentes, pero concibiendo todas esas evidencias como contribuyentes a la validez de constructo. Las evidencias de contenido son especialmente relevantes en pruebas educativas de conocimientos que miden resultados de procesos de aprendizaje formales. Las predictivas se refieren a instrumentos que intentan estimar comportamientos futuros, tal es el caso de las pruebas de admisión a la educación superior, en donde se busca que los puntajes se asocien a rendimientos futuros de los estudiantes en la Universidad. Por su parte, las evidencias concurrentes se refieren a las asociaciones que deben presentan entre sí pruebas que intentan medir el mismo constructo.

Otro de los más importantes aportes de Messick (1989a, 1989b) se refiere a su reflexión en torno a que la validez no es una propiedad intrínseca de los instrumentos, sino que se define de acuerdo al propósito de la medición, la población a la que va dirigida y el contexto específico de aplicación. Así, un instrumento puede exhibir un grado aceptable de validez para un propósito específico y para una población particular, pero no para otros.

Además, el proceso de validación no termina, es permanente, dado que, al igual que el resto de actividades de la ciencia moderna, exige comprobaciones empíricas continuas. Igualmente, nos recuerda Messick (1989a; 1989b) que la validez no es un rasgo dicotómico, sino una cuestión de grado, no se puede decir de manera contundente que una prueba es válida, sino más propiamente se puede afirmar que la prueba exhibe un grado aceptable de validez para ciertos usos específicos y con ciertas poblaciones.

Finalmente, Messick hace recapacitar a la comunidad de medición educativa cuando afirma que el constructor(a) del instrumento no solo debe poner atención a lo científico- técnico sino también a lo ético: debe preocuparse por el uso que se da a los instrumentos y por las consecuencias derivadas de la aplicación de los mismos (Messick, 1989a y 1989b; Padilla et al, 2006).

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

En primer lugar las evidencias deben mostrar en qué medida el instrumento, como un todo, y los ítems o reactivos que lo componen, representan adecuadamente al constructo teórico que se pretende medir y a sus componentes. Por esto para lograr un instrumento con alta validez, es indispensable el manejo de los referentes teóricos y su correcta operacionalización.

Sin embargo, lo anterior no es suficiente para generar evidencias sólidas de validez, sino que se debe documentar el grado de propiedad de las diversas inferencias que se generan a partir de los puntajes del instrumento. Como ejemplos, estas inferencias pueden incluir decisiones de promoción en el caso de pruebas de certificación y decisiones de admisión en el caso de pruebas de selección para la Universidad.

Por último, es importante discutir tres conceptos muy útiles en torno a la validez de un instrumento, ellos son variancia relevante al constructo, variancia irrelevante al constructo y sub-representación del constructo, mismos que igualmente fueron precisados por Messick (1989a; 1989b).

Variancia relevante al constructo es, efectivamente, lo que tratamos de maximizar cuando construimos un instrumento, pues buscamos que las puntuaciones reflejen, precisamente, los diferentes niveles que toma el constructo de interés en los sujetos examinados. En otras palabras, se desea

que la variabilidad que se observa entre los puntajes del instrumento sea variabilidad verdadera, debida a las diferencias en el constructo que presentan los examinados.

Por el contrario, variancia irrelevante al constructo está constituida por variaciones en los puntajes del instrumento que no representan variaciones reales en el constructo de interés, sino que son debidas a otros factores, entre ellos podemos mencionar sesgos y errores de medición. Un ejemplo de variancia irrelevante al constructo sería el caso de un instrumento para medir habilidad cuantitativa en la resolución de problemas, en donde los enunciados de los reactivos están cargados de vocabulario poco común y mucha complejidad verbal. Es probable entonces que los puntajes de la prueba no solo reflejen habilidad cuantitativa, sino también conocimiento de vocabulario y comprensión verbal, introduciendo así un sesgo y una fuente de invalidez en la interpretación de los puntajes.

Por su parte, cuando hablamos de sub-representación del constructo nos referimos al hecho de que, en ocasiones, un instrumento particular solamente mide un componente o dimensión de un constructo que es más complejo y que involucra otros aspectos. Se puede mencionar como ilustración el caso de las pruebas de inteligencia tradicionales, donde se podría decir que miden solamente ciertos aspectos específicos del constructo, quedando otras dimensiones sub-representadas en el indicador. Es el mismo caso de la prueba de admisión de la Universidad de Costa Rica, si se afirmara (como se hizo en sus inicios) que mide aptitud académica, cuando en realidad el constructo objeto de la medición son habilidades de razonamiento en contextos verbales y matemáticos, rasgos que ciertamente pueden pensarse como parte del constructo aptitud académica, pero que no lo agotan ni lo representan exhaustivamente.

Antes de concluir esta sección es necesario dedicar nuestra atención al concepto de confiabilidad. Confiabilidad significa precisión, consistencia, estabilidad en repeticiones. Una definición conceptual bastante ilustrativa indica que un instrumento es confiable si aplicado en las mismas condiciones a los mismos sujetos produce los mismos resultados (Nunnally & Bernstein, 1995).

La confiabilidad es condición necesaria pero no suficiente para la validez. Es decir, si el instrumento exhibe un grado aceptable de validez ello implica que también debe poseer un grado aceptable de confiabilidad (como es claro a partir de la definición de esta última), sin embargo, lo opuesto no es cierto, o sea, un instrumento que exhibe un alto nivel de confiabilidad no necesariamente es válido, esto porque puede estar midiendo con alta precisión y consistencia, pero sin garantía de que lo medido sea el constructo de interés (Babbie, 2010). La evidencia de confiabilidad es entonces un requisito necesario pero no suficiente para la validez (Babbie, 2010).

Entre los indicadores de confiabilidad que usamos con más frecuencia en psicometría se incluyen el Alfa de Cronbach que es el resultado más

importante de la Teoría Clásica de los Tests (TCT), el índice de discriminación, calculado en la TCT como la correlación ítem-total, así como la cantidad de error de medición y el tamaño de la función de información en Teoría de Respuesta a los Ítems (TRI) y el modelo de Rasch (Martínez et al, 2006; Muñiz, 2003; Prieto & Delgado, 2003).

Validez psicométrica

El proceso de recolección de evidencias empíricas para la validación de un instrumento implica normalmente y como primer paso, la consulta a jueces expertos, aunque esto no es suficiente para generar evidencia de validez sólida y suficientemente creíble. Hace falta al menos una aplicación piloto del instrumento y un análisis psicométrico básico del instrumento y de los ítems que lo componen. Entre los métodos y modelos de análisis que utilizamos en este proceso se pueden mencionar los siguientes:

- Análisis de factores exploratorio y confirmatorio
- Teoría Clásica de los Tests (TCT)
- Teoría de Respuesta a los Ítems
- Modelo de Rasch
- Teoría G (Generalizabilidad)
- Análisis DIF
- Equiparación de puntajes

Los análisis de factores, tanto exploratorios como confirmatorios, se refieren a técnicas multivariadas que nos permiten explorar la dimensionalidad subyacente en los datos (Martínez et al, 2006; Nunnally & Bernstein, 1995). El análisis factorial exploratorio (AFE) se usa en psicometría para obtener evidencias de las dimensiones subyacentes, factores o componentes que están presentes en el instrumento y que deberían corresponder, en teoría, con los constructos o rasgos latentes que se intenta medir. Se trata de explicar las correlaciones observadas entre los ítems del instrumento a partir de un conjunto más pequeño de componentes o dimensiones, llamados factores, por eso también se lo conoce como una técnica de reducción de datos. A nivel global, las cargas o saturaciones factoriales de los ítems (que estiman la correlación entre cada ítem y cada factor) se consideran óptimas si son iguales o mayores a 0.3, en valor absoluto. En cuanto al factorial confirmatorio, se puede afirmar que es, en la actualidad, la estrategia analítica más apropiada para testear empíricamente la configuración teórica de un instrumento, en términos de los constructos o rasgos latentes que representa, incluidas sus dimensiones o componentes dentro de una posible estructura jerárquica (Brown, 2006; Martínez et al, 2006). El análisis factorial exploratorio puede ser visto como un caso particular de un análisis confirmatorio, y este, a su vez, es un caso particular de un modelo de ecuaciones estructurales, conocidos

también como SEM por sus siglas en inglés (Structural Equations Models) (Mulaik, 2009; Kaplan, 2009).

Por su parte, la Teoría Clásica de los Tests (TCT) es el más antiguo y conocido modelo de medición, que permite generar indicadores empíricos objetivos de la calidad técnica de un instrumento, incluyendo su resultado de mayor importancia práctica, el coeficiente Alfa de Cronbach, indicador que mide la precisión de la prueba en términos del grado de consistencia interna del instrumento y apunta hacia el grado de estabilidad de los puntajes (Muñiz, 2003). Alfa estima qué proporción de la variabilidad observada en los puntajes corresponde a variancia verdadera, es decir variancia debida a diferencias en el constructo que se desea medir. Su valor máximo es 1, y cuanto más se aproxime Alfa a 1 mayor es el nivel de confiabilidad. En general, los programas internacionales de pruebas educativas consideran aceptables valores de Alfa mayores a 0.8. No obstante, autores como Nunnally & Bernstein (1995) son más estrictos cuando se refieren a pruebas de altas consecuencias en donde toman decisiones directas sobre los examinados, e indican que tales exámenes debería exhibir una confiabilidad de al menos 0.9 en la medida Alfa de Cronbach. Por otra parte, si se trata de instrumentos que van a ser utilizados solamente para procesos de investigación se puede ser más flexible en el criterio. En ese caso se consideran aceptables valores de Alfa iguales o mayores a 0,7 (Nunnally & Bernstein (1995).

La fórmula para calcular Alfa se representa a continuación:

$$\alpha = (k / k-1) \ (1 - \Sigma\sigma_i^2 / \sigma_y^2)$$

donde,
k es el número de ítems
$\Sigma\sigma_i^2$ es la sumatoria de las variancias individuales de los ítems
σ_y^2 es la variancia de la suma total de los puntajes

Otra de las medidas más conocidas en la TCT es el índice de discriminación del ítem, que se calcula como la correlación entre el puntaje del ítem y el puntaje total en el instrumento, excluyendo de este último el ítem específico que está siendo analizado. Valores de discriminación superiores a 0.3 se consideran óptimos. Por su parte, la dificultad del ítem se define, tanto para pruebas afectivas como cognitivas, como el promedio de las respuestas obtenidas en el reactivo, hablando de ítems fáciles cuando este promedio es alto e ítems difíciles cuando es bajo. Un caso particular muy conocido es el que define la dificultad como la proporción de respuestas correctas, cuando se trata de prueba cognitivas de calificación dicotómica para cada ítem (1=correcto-0=incorrecto).

Finalmente, en los modelos TRI (Teoría de Respuesta a los ítems) y Rasch se ajusta un modelo matemático al comportamiento del ítem, siendo los parámetros del ítem (dificultad, discriminación y "factor de azar") y los

puntajes del examinado variables latentes que requieren un proceso de estimación matemático-estadístico. Con esto se obtienen parámetros del ítem que son menos dependientes de la muestra de examinados y estimaciones de los niveles del constructo en los evaluados que son menos dependientes de la muestra particular de ítems aplicada. Además, en estos modelos existe una estimación específica del error de medición para cada puntaje en la prueba (a diferencia de la TCT donde se asume que el error es constante) (Martínez et al, 2006; Montero, 2001).

El modelo de Rasch es matemáticamente hablando un caso particular de un modelo TRI donde se asume que la discriminación de los ítems es constante y que el llamado "factor de azar" es igual a cero. De esta manera el único parámetro del ítem a estimar en el modelo de Rasch es la dificultad (parámetro b). Gracias a esto, en este modelo, las estimaciones del constructo en los examinados y la medida de dificultad de los ítems están en las mismas unidades, característica que se denomina propiedad de medición conjunta. Esta propiedad resulta sumamente atractiva a nivel aplicado y de interpretación sustantiva, pues permite evaluar el desempeño del examinado en términos de modelos referidos a criterios, es decir valorando, en términos absolutos, lo que puede o no lograr en el rango de medición del constructo que nos provee el instrumento (Bond & Fox, 2001; Prieto & Delgado, 2003; Wilson, 2005).

La expresión matemática para el modelo de Rasch es la siguiente:

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

$P_i(\theta)$ es la probabilidad de responder exitosamente al ítem i por parte del examinado con nivel θ en el constructo

e es la base de los logaritmos neperianos

b_i es la dificultad del ítem i

A partir de esta fórmula se puede establecer que la probabilidad de acertar (o fallar el ítem) solo depende de la distancia entre la habilidad del sujeto y la dificultad del ítem. Cuando $\theta = b$, el examinado tiene una probabilidad de 50% de acertar el ítem, cuando $\theta > b$, la habilidad del sujeto supera la dificultad del ítem, por tanto su probabilidad de acierto es mayor a 0.5. Por el contrario, cuando $\theta < b$ la dificultad del ítem supera la habilidad del sujeto, por tanto su probabilidad de contestarlo correctamente es menor a 0.5.

Por último, antes de realizar un análisis psicométrico con la TCT, la TRI o Rasch es importante evidenciar, utilizando el análisis factorial exploratorio, que el instrumento mide fundamentalmente solo un rasgo o constructo, pues este es un supuesto que debe cumplirse para que la aplicación de estos modelos de medición sea válida.

Por su parte, la Teoría G (Generalizabilidad) resulta muy útil en los procesos de validación asociados a instrumentos de respuesta abierta o respuesta construida por el examinado, en donde usualmente se debe contar con calificadores y en donde se busca minimizar la influencia de la "idiosincrasia" del evaluador en la asignación de los puntajes (Zúñiga & Montero, 2007).

La Teoría G permite medir la confiabilidad de una prueba por medio de la cuantificación de la importancia de cada una de sus fuentes de variabilidad. Uno de sus aportes más importantes es que redefine y desagrega el error en términos de las diversas condiciones o facetas de medición, utilizando el Coeficiente de Generalizabilidad como medida para estimar la confiabilidad. Se debe notar que este enfoque no contradice los planteamientos fundamentales de la Teoría Clásica de los Tests, sino que puede ser visto como una extensión de ella. Para realizar estudios con Teoría G se utiliza el marco de referencia de los diseños experimentales y el Análisis de Variancia. Uno de sus objetivos fundamentales es cuantificar la variancia debida a las diferentes facetas de medición, y minimizar aquellas fuentes de variabilidad que no se refieren a diferencias de los examinados en el constructo, siendo una de ellas la que mide el "efecto" de los calificadores.

Los temas de Funcionamiento Diferencial del Ítem (DIF, por sus siglas en inglés) y equiparación de puntajes (equating) no son nuevos para la mayoría de los constructores de pruebas e investigadores en medición y psicometría en la comunidad académica internacional. Costa Rica, sin embargo, hasta hace poco se había quedado un tanto rezagada en cuanto a la incorporación de estos procedimientos como parte de los procesos regulares en la construcción y análisis de pruebas estandarizadas. Aunque programas de pruebas similares, tales como el SAT (Scholastic Aptitude Test) de los Estados Unidos, realizan estos análisis "de oficio" desde los años 80 (Dorans & Kulick, 1986), nuestro contexto se ha desarrollado en esta dirección a partir del inicio del siglo XXI.

Ambas temáticas (Funcionamiento Diferencial del Ítem y equiparación de puntajes) revisten gran importancia desde el punto de vista de la validez, entendiendo este concepto en el sentido más moderno propuesto por Messick (1989a; 1989b; 1995) y que fue descrito más arriba, como el grado de propiedad de las interpretaciones derivadas de los puntajes de la prueba. Se trata entonces de tópicos que, desde este punto de vista, están asociados también a la equidad y la justicia en las decisiones que se basan en los resultados de la prueba.

En el caso del Funcionamiento Diferencial del Ítem se busca generar evidencia de comportamiento diferencial de los examinados en las respuestas a los ítems que sea producto de factores espurios y que no refleje diferencias reales en sus niveles del constructo. Así, existen ciertas características de los examinados, que en interacción con el contenido del ítem, pueden provocar desempeños diferenciales (Penfield & Camilli, 2007). La antigua denominación "sesgo del ítem" (item bias) se comenzó a sustituir en los años 90 por la más

neutral Funcionamiento Diferencial del Ítem o Differential Ítem Functioning, en inglés (Angoff, 1993; Embretson & Reise, 2000). Si las diferencias entre los puntajes son producto de otros factores y no de diferencias en los niveles del constructo, entonces se atenta contra la validez de la interpretación y se pone en desventaja a ciertos grupos de examinados. Entonces, según el planteamiento de Messick, decimos que estos factores provocan varianza irrelevante al constructo, son fuente de invalidez y, por tanto, deben ser cuantificados y minimizados (Messick, 1989a; 1989b; 1995).

La equiparación de puntajes, por su parte, se refiere al grado de comparabilidad entre puntuaciones de examinados que han recibido formas diferentes de la misma prueba, es decir, instrumentos que han sido ensamblados con la misma tabla de especificaciones, para medir el mismo constructo y que poseen los mismos niveles de dificultad y confiabilidad (Dorans & Holland, 2000; Pacheco-Villamil, 2007). Dos formas de una prueba están equiparadas cuando es indiferente para el examinado tomar una u otra, es decir cuando se derivan las mismas interpretaciones de los resultados y consecuentemente las mismas decisiones. (Kolen & Brennan, 2004). Obviamente, en la práctica, para lograr aproximarse a esta definición es indispensable contar, en primer lugar, con un banco de ítems de estadísticas conocidas. Sin embargo, aun cuando se trate de ensamblar las formas de la prueba con igual dificultad y confiabilidad, es usual que una vez que se realiza la aplicación operativa de las pruebas, se presenten diferencias relativamente pequeñas de dificultad entre ellas. Aunque estas diferencias no posean significancia estadística sí pueden tener implicaciones diferenciales en las decisiones sobre algunos examinados, especialmente en el caso de pruebas de altas consecuencias. Por eso es que en este contexto se hace necesaria la equiparación. Entonces, la comparación directa entre puntuaciones brutas (no equiparadas) de individuos que han realizado formas diferentes de una misma prueba, debe ser realizada con cautela en la mayoría de las ocasiones, pues esta comparación no controla el efecto de posibles diferencias en el nivel de dificultad. En otras palabras, dos personas pueden tener la misma puntuación bruta, habiendo tomado formas distintas de la prueba y, sin embargo, ambas pueden presentar desempeños diferentes en el constructo objeto de medición cuando se controlan las diferencias en dificultad. Además, cuando estas puntuaciones brutas surgen en el contexto de un modelo de normas, tal como es en el caso de la prueba de admisión de la Universidad de Costa Rica, un elemento adicional que se debe considerar en la comparación es el grupo de referencia para la equiparación, pues el puntaje de un individuo particular se genera a partir de su comparación con el desempeño relativo del resto del grupo, de manera que su calificación solo tiene sentido en términos de su grupo referencial. Entonces, si se desea comparar de manera válida el desempeño de dos examinados que provienen de diferentes grupos de referencia, igualmente se requiere equiparar sus puntajes.

En la actualidad se cuenta con un vasto recurso de software computacional para el análisis psicométrico de instrumentos. Tanto la TCT como el análisis de factores exploratorio se encuentran disponibles en la gran mayoría de paquetes estadísticos de propósito general, incluyendo SPSS, Stata y SAS. El análisis factorial confirmatorio, al ser un caso particular de los modelos de ecuaciones estructurales (o SEM, por sus siglas en ingles) están implementados en diversos paquetes de software especializado, siendo los más comunes, LISREL, EQS, Mplus y Amos (este último actualmente es parte de la corporación SPSS). En cuanto a software para TRI se pueden mencionar el BILOG y el IRTPRO. En lo que toca a modelos de Rasch dos de los paquetes especializados más utilizados son el Winsteps y el ConQuest. Igualmente la mayoría de estos procedimientos están implementados, con diversos niveles de sofistificación, en librerías de R, lenguaje y ambiente gratuito de programación para análisis estadístico, que se ha hecho muy popular en los últimos años.

También existe software disponible para DIF y equiparación de puntajes como parte de los paquetes de software especializado que se mencionaron arriba. El DIFAS es un software especializado para análisis de DIF, construido por el profesor Randall Penfield de la Universidad de Miami, que se puede descargar gratuitamente contactando a su autor. Igualmente, como parte de un proyecto de investigación a cargo de esta autora, se elaboró en el Programa Permanente de la Prueba de Aptitud Académica un software a la medida para detectar DIF con el método de la diferencia p estandarizada propuesto por Neil Dorans (Dorans & Kulick, 1986).

Ahondar más en la fundamentación, teoría y aplicación de estos modelos de análisis escapa el propósito de este documento, basta decir que se trata de temas de cierta complejidad técnica que exigen dedicación para su estudio y cabal comprensión, al igual que sólidas bases estadísticas y matemáticas.

Implicaciones de estos referentes para la aplicación de pruebas a personas con necesidades educativas especiales

El término adecuación (*accommodation*, en inglés) se usa para cualquier acción que se toma una vez que se ha determinado que la discapacidad que posee una persona requiere una desviación del protocolo establecido para la aplicación de una prueba estandarizada (Sireci, 2004; Gordon & Keiser, 2000; AERA, 1999).

El propósito de las adecuaciones debe ser minimizar el impacto de los atributos del examinado que no son relevantes al constructo que es el foco principal de la medición. La adecuación, en principio, no debe implicar un cambio en el constructo de interés. Así, desde el concepto moderno de validez, las adecuaciones tienen su justificación y se realizan con el propósito de eliminar fuentes de varianza irrelevante al constructo y lograr así mayor validez (Sireci, 2004; AERA, 1999).

Por otra parte, en inglés se utiliza el término *modifications* (modificaciones) para designar desvíos en el contenido del test o en los procesos de aplicación que conllevan una modificación al constructo de interés (Sireci, 2004; AERA, 1999).

En este sentido, semánticamente hablando, los términos adecuaciones y modificaciones definidos aquí son consistentes con sus definiciones "de diccionario". Por ejemplo, si el uso de un lápiz para marcar las respuestas es incidental al constructo que se pretende medir con la prueba, una adecuación en el acceso al registro de las respuestas es necesaria para una medición más precisa en el caso de estudiantes que presenten problemas con el manejo de su motora fina. De esta forma se equiparan las condiciones de aplicación del instrumento entre la población con necesidades especiales y los miembros de la población regular, para que tengan la misma oportunidad de mostrar su desempeño en el constructo de interés. Al mantenerse intacto el constructo es posible hacer inferencias válidas acerca de los desempeños comparativos de los grupos con y sin adecuaciones. El significado de los puntajes se mantiene (Sireci, 2004; AERA, 1999).

Un caso muy diferente se presenta cuando la discapacidad, de hecho, es directamente relevante al constructo. Por ejemplo, no se pueden realizar adecuaciones a una persona no vidente si la prueba está diseñada para medir su capacidad visual para distinguir diversos colores. Cualquier intento de ajuste en la prueba o en su aplicación redundaría en una modificación, es decir, en una alteración del constructo objeto de la medición y, por tanto, se invalidaría cualquier inferencia que se intentara en términos de comparar los resultados con y sin modificaciones (Sireci, 2004). A medida que se van introduciendo modificaciones los puntajes comienzan a carecer de significado, por eso es que en los Estados Unidos se reconoce que aunque en ciertos contextos es necesario hacer modificaciones con el propósito de que estudiantes con discapacidades tengan oportunidad de participar en las pruebas y demostrar así sus conocimientos y destrezas, las modificaciones no deberían ser excesivas, y deberían alterar lo menos posible las condiciones estándar de administración de los instrumentos.

Cuando existe discrepancia entre la adecuación solicitada por el examinado y la adecuación otorgada por el programa de pruebas, el criterio fundamental debe regirse por la validez de la inferencia hecha a partir del puntaje en la prueba con adecuación: la adecuación debe proveer una medida más precisa del constructo de interés (AERA, 1999).

Algunas recomendaciones que se pueden derivar desde este enfoque para el proceso de adecuaciones incluyen las siguientes (AERA, 1999):

1- Las personas que toman las decisiones acerca de las adecuaciones para los individuos con discapacidades deberían conocer la investigación existente acerca de los efectos de la discapacidad específica sobre el desempeño en la prueba.

2- Cuando sea posible, el test debería ser probado de manera piloto en individuos con discapacidades similares a los de la población meta, para investigar el grado de propiedad y factibilidad de las adecuaciones.

3- Deberían usarse procedimientos empíricos para establecer los límites de tiempo apropiados para la realización de pruebas con adecuaciones, en el caso de que la prueba regular tenga un límite de tiempo establecido.

4- La validez de las inferencias y la confiabilidad de los puntajes deberían ser investigadas y reportadas para la población con adecuaciones.

5- Aparte de la adecuación formalmente establecida, se deben mantener todas las otras características de aplicación uniforme que forman parte de la prueba estandarizada.

Conclusión

Los profesionales e investigadores de Ciencias Sociales en general, y de Educación y Psicología en particular, por la naturaleza de su objeto de estudio, se ven enfrentados con mucha frecuencia a las tareas de construir, validar o adaptar medidas para aproximar constructos complejos. Con esta actualización teórica y metodológica se pretende brindar a los constructores y usuarios de instrumentos orientaciones claras y renovadas en torno al concepto moderno de validez, así como describir diversas herramientas analíticas disponibles en la actualidad, para generar evidencias científicamente sólidas de validez.

Referencias

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

Angoff, W.H. (1993). Perspectives on Differential Item Function Methodology. En P.W. Holland y H. Wainer (Eds.). *Differential item functioning* (pp. 3-23). New Jersey, Estados Unidos de America: Lawrence Erlbaum Associates.

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

Bond, T. & Fox, C. (2001). *Applying the Rasch model: fundamental measurement in the human Sciences*. Mahwah, New Jersey: LEA.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: the Guilford Press.

Dorans, N. J. & Holland, P. W. (2000). Population invariance and the equability of test: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.

Dorans, N.J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal or Educational Measurement*, 23(4), 355-368.

Embretson, S. E & Reise S.P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associatesm Mahwah, New Jersey.

Gordon, M. & Keiser, S. (Eds.) (2000). *Accommodations in Higher Education under the Americans with Disabilities Act: A No-Nonsense Guide for Clinicians, Educators, Administrators, and Lawyers*. New York: GSI Publications.

Hopkins K.D., Hopkins, B.R. & Glass, G.V. (1997). *Estadística Básica para las Ciencias Sociales y del Comportamiento*. México: Prentice-Hall Hispanoamericana.

Kaplan, D. (2009). *Structural equation modeling: foundations and extensions*. Segunda edición. Thousand Oaks, CA: Sage.

Kerlinger, F.N. & Lee, H.B. (2002). *Investigación del comportamiento*. México: McGraw-Hill.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking* (2nd Ed.). New York: Springer.

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

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, vol. 18, # 2, 5-11.

Messick, S. (1989b). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, Samuel (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and practice*, 14 (2). pp. 5-24. Boston, Estados Unidos de América: Blackwell Publishing.

Montero, E. (2001). La teoría de respuesta a los ítems: una moderna alternativa para el análisis psicométricos de instrumentos de medición. *Revista de Matemática: teoría y aplicaciones*. Centro de Investigaciones en matemática pura y aplicada (CIMPA) y la Escuela de Matemática de la Universidad de Costa Rica. Vol. 7, # 1-2, págs. 217-228.

Mulaik, S.A. (2009). *Linear causal modeling with structural equations*. New York: CRC Press Taylor & Francis Group.

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

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

Pacheco-Villamil, J. (2007). La equipacación de puntuaciones en procesos de comparación de pruebas diferentes. *Avances en Medición, 5*, 153-156

Padilla J.P. et al (2006). La evaluación de las consecuencias del uso de los tests en la teoría de validez. *Psicothema*, vol. 18, nº 2, pp 307-312.

Penfield, R. & Camilli, G. (2007). Differential Item Functioning and Item Bias. En S. Sinharay y C.R. Rao (Eds.). *Handbook of Statistics*. Vol. 26, Elsevier.

Prieto, G. & Delgado A.R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, vol. 15, nº 1, pp. 94-100.

Sireci, S. G. (2004). Validity Issues in Accommodating NAEP Reading Tests. NAGB Conference on Increasing the Participation of SD and LEP Students in NAEP. Commissioned Paper. Center for Educational Assessment Research Report No. 515. Amherst, MA: School of Education, University of Massachusetts Amherst.

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

Zúñiga, M. & Montero, E. (2007). Teoría G: un futuro paradigma para el análisis de pruebas psicométricas. Artículo aceptado para publicación. *Revista Actualidades en Psicología.* San José, Costa Rica: Universidad de Costa Rica, Instituto de Investigaciones Psicológicas.

Recibido: 02 de enero de 2013
Aceptado: 04 de abril de 2013