Validez en test adaptativos informatizados: alternativa para evaluar población con limitaciones visuales[1,2]

Aura Nidia Herrera Rojas[3], *Rocío Barajas Sierra, Gillen Javier Jiménez López*
Universidad Nacional de Colombia, Bogotá, Colombia

RESUMEN

Este artículo centra su interés en la importancia de la validez dentro de los procesos de evaluación psicológica cuando se pretende evaluar a personas invidentes o con baja visión, y en el desarrollo de los Test Adaptativos Informatizados (TAI) como una alternativa para la evaluación de esta población. Se presenta una revisión sobre el concepto de validez a partir de las aproximaciones contemporáneas dominantes, y se habla acerca del desarrollo de los TAI, el problema de su validez y los alcances de esta tecnología como una alternativa para evaluar personas con baja visión o invidentes. Finalmente, este trabajo deja abierta la posibilidad para futuros desarrollos e investigaciones en torno a otras alternativas de evaluación con equidad para poblaciones con discapacidad física.

Palabras clave: validez; test adaptativos informatizados; limitación visual

RESUMO – Validade em testes adaptativos computadorizados: alternativa para avaliar população com deficiência visual

Este artigo centra o seu interesse na importância da validade dentro dos processos de avaliação psicológica de pessoas cegas ou com baixa visão e no desenvolvimento dos Testes Adaptativos Computadorizados (TAC) como uma alternativa para a avaliação dessa população. Apresenta-se uma revisão sobre o conceito de validade a partir das abordagens contemporâneas dominantes e sobre o desenvolvimento dos TAC, o problema da sua validade e os alcances dessa tecnologia como uma alternativa para avaliar pessoas cegas. Finalmente, o trabalho deixa em aberto a possibilidade de futuros desenvolvimentos e pesquisas sobre alternativas de avaliação com equidade para a população com deficiência física.

Palavras-chave: validade; testes adaptativos computadorizados; limitação visual.

ABSTRACT – Validity in computerized adaptive testing: An alternative to assess visually impaired individuals

This article focuses on the importance of validity in the process of psychological evaluation, especially when the evaluation is intended for people with visual impairment, and about the development of the Computerized Adaptive Testing (CAT) as an alternative for the assessment of this population. First, a review of the main controversies about the concept of validity and the development of Computerized Adaptive Test is presented. Then, topics related to validity in CAT and scope, technological limitations, and the challenges of its use as an alternative to evaluate the visual impairment population are discussed. Finally, possibilities for future developments and research on high quality alternative assessment for populations with physical disabilities are set forth.

Keywords: validity; computerized adaptive testing; visual impairment.

El uso de pruebas psicológicas en procesos evaluativos busca apoyar la toma de decisiones respecto a un grupo de individuos, por ello, se busca que estas herramientas sean válidas y generen confianza en la diada evaluador-evaluado, y en las interpretaciones y conclusiones que se deriven de su uso. La *American Psychological Association* (APA), junto con la *American Educational Research Association* (AERA) y el *National Council on Measurement in Education* (NCME), en sus dos últimas versiones de los *Standards for Educational and Psychological Testing* (*American Psychological Association*, *American Educational Research Association* & *National Council on Measurement in Education*, 1999, 2014) reconocen explícitamente la validez como la característica más importante en la evaluación de las pruebas objetivas; sin embargo, la complejidad del concepto y la constante intervención de nuevas variables producto de los avances en el diseño y la teorización de las pruebas objetivas, han dificultado el consenso alrededor de su definición. Cizec (2012) señala que uno de los puntos de acuerdo

[1] Nota: Esta investigación fue parcialmente financiada por la Universidad Nacional de Colombia – Proyecto 12907 – DIB – y por el Instituto para Evaluación de la Calidad de la Educación ICFES y el Departamento Administrativo de Ciencia, Tecnología e Innovación, COLCIENCIAS. Proyecto 255-2011.

[2] Agradecimientos: A los integrantes del grupo de investigación "Métodos e Instrumentos para investigación en Ciencias del Comportamiento" por sus observaciones y sugerencias durante el desarrollo de la investigación.

[3] Endereço para correspondência: Laboratorio de Psicometría, Oficina 227, Edificio 212, Ciudad Universitaria, Universidad Nacional de Colombia, Bogotá, Colombia. Tel.: 571-3165000 (ext. 16347) / 571-3165304. *E-mail:* anherrerar@unal.edu.co

en la conceptualización contemporánea de validez es considerar que esta se refiere a las inferencias y los usos de las puntuaciones de las pruebas, lo cual implica que la población a la que va dirigida la prueba así como los demás aspectos contextuales son importantes a la hora de establecer qué tan válido es un instrumento de medida. Esto representa un reto si se pretende evaluar mediante los mismos instrumentos a personas con y sin algún tipo de limitación.

Con el fin de favorecer la inclusión de poblaciones con limitaciones visuales (LV), es decir, de personas con baja visión o invidentes, en los procesos evaluativos que involucran pruebas de lápiz y papel, frecuentemente se cuenta con lectores capacitados, quienes leen las preguntas y las opciones de respuesta, y diligencian la hoja de respuestas con la opción que el examinado señale como correcta. Esta o cualquier otra estrategia de aplicación introduce variables en el proceso de evaluación, lo que puede afectar su validez, y en consecuencia, reducir el grado de certeza de las inferencias realizadas acerca del atributo medido.

En general, se ha encontrado que las personas con LV obtienen resultados comparativamente inferiores a los de la población sin este tipo de limitación. El Instituto Nacional para Ciegos (INCI, 2010) reporta que la media de las calificaciones obtenidas por los estudiantes con LV que finalizaron su ciclo de educación media en 2009 fue menor que la media nacional. Si bien la falta de acceso a herramientas adecuadas para el proceso de enseñanza de personas con LV puede ser uno de los factores de gran incidencia para tener estos resultados, se plantean algunas cuestiones relacionadas con el nivel de habilidad de estas personas, la precisión de la evaluación para esta población, el posible efecto de variables asociadas con la participación del lector, y la utilización de otras metodologías de evaluación que favorezcan la autonomía de respuesta para la población con LV, con un adecuado nivel de validez y confiabilidad.

La Teoría de Respuesta al Ítem (IRT, por sus siglas en inglés) trajo consigo la promesa de independizar las estimaciones del atributo, del instrumento utilizado y de las características de la población (Wright, 1968), además de una serie de aplicaciones impensables desde la TCT, como la calibración de bancos de ítems, algunos procedimientos para la equiparación de puntuaciones y para la detección del Funcionamiento Diferencial de los Ítems (DIF, por sus siglas en inglés), y el diseño de Test Adaptativos Informatizados (TAI).

Los TAI son pruebas computarizadas que teniendo como base un banco de ítems calibrado, adaptan la presentación de los ítems a los examinados, basándose en las estimaciones de su nivel de atributo (Gómez & Hidalgo, 2003), de manera que con el menor número de ítems posible se logre tener una medida precisa de la característica evaluada. Los TAI presentan grandes ventajas sobre las pruebas de lápiz y papel convencionales, como la ruptura de las concepciones tradicionales de que es necesario contar con formas paralelas de una prueba para poder realizar comparaciones de puntuaciones entre individuos, y que cuantos más ítems tiene una prueba mayor confiabilidad se puede predicar de ella (Gómez & Hidalgo, 2003), entre otras.

Sin embargo, al involucrar la mediación de herramientas computarizadas, vale la pena preguntarse si al estimar el nivel de un atributo o una característica determinada a través de un TAI, existen variables que afecten la validez del instrumento. Esto resulta de gran importancia ya que los TAI se postulan como una de las estrategias de evaluación más prometedoras dada la inclusión de la tecnología y las ventajas que representa debido a su flexibilidad en la presentación de las preguntas, el aumento de la precisión y la eficiencia en la estimación de la habilidad de los individuos, y su autonomía durante la aplicación. Teniendo en cuenta las características propias de la población con LV y la escasa literatura aplicable al presente caso, es necesario adelantar estudios en la adecuación de pruebas para estas personas.

En el presente artículo se realizará una revisión en torno al problema de la validez en los TAI y la aplicación de esta tecnología como una alternativa para evaluar personas con LV. Se abordará el concepto de validez a partir de las aproximaciones contemporáneas dominantes, y se hablará acerca de los TAI, en particular, acerca de su desarrollo y los potenciales alcances en el campo de la medición en personas con LV.

Consideraciones en torno al concepto de validez

Para Elosua (2003) la historia de la conceptualización de validez se puede dividir en tres etapas evidenciadas a partir de las distintas definiciones de los estándares de APA, AERA y NCME. La primera etapa, "dominada por una visión pragmática en la que prima la validez externa" (p. 315) abarca todas las aproximaciones al concepto hasta la publicación de los estándares de 1966, cuando la definen como el grado en el que la prueba mide lo que pretende medir, y la clasifican en tres formas: de criterio, de contenido y de constructo. La segunda etapa, caracterizada por la "visión unificada" de validez, se condensa en los estándares de APA, AERA y NCME de 1974 y 1985 y tiene sus orígenes en el trabajo de Cronbach y Meehl (1955) en el que la definen como el "análisis de la significación de las puntuaciones de los instrumentos de medida expresado en términos de los conceptos psicológicos asumidos en su medición" (p. 442).

De acuerdo con la visión unificada de Messick (1995) el propósito de la validación del uso de las pruebas así como de las interpretaciones de sus puntuaciones requiere tanto bases evidenciales (aproximaciones prácticas de las relaciones con otros constructos) como consecuenciales (valoraciones de las implicaciones reales

y potenciales de las interpretaciones dadas a las puntuaciones obtenidas). Para Elosua (2003) este énfasis sobre el contexto en el que se desarrollan y aplican las pruebas, sus usos potenciales y las posibles inferencias hechas a partir de sus puntuaciones, caracteriza la última etapa de la historia de la conceptualización de validez y se condensa en la versión de los estándares de APA, AERA y NCME de 1999.

En esta versión de estándares y en la posterior de 2014, se define validez como "el grado en el cual la evidencia y la teoría soportan las interpretaciones de las puntuaciones de una prueba para los usos propuestos" (*American Psychological Association*, *American Educational Research Association* & *National Council on Measurement in Education*, 2014 p. 11. Traducción de los autores), y se proponen cinco fuentes de evidencia de validez relacionadas con: los procesos de respuesta, el contenido de las pruebas, su estructura interna, su relación con variables externas, y sus consecuencias.

Esta definición de validez generó fuertes debates debido a la poca claridad del concepto y la ausencia de ejemplificación (Sireci, 2009), la dificultad para definir el criterio que determine que se cuenta con suficiente evidencia para validar un uso o inferencia (Sireci, 2009), la paulatina disociación entre la teoría y la práctica de la validez (Kane, 2009), y la inclusión de las consecuencias sociales como fuente de validez (Linn, 1997; Mehrens, 1997; Popham, 1997). Algunas de estas críticas pueden verse al menos parcialmente superadas en los 25 criterios de validez descritos en la última versión de los estándares. (*American Psychological Association*, *American Educational Research Association* & *National Council on Measurement in Education*, 2014). Estos criterios se organizan en tres grupos: sobre los usos e interpretaciones de la prueba, sobre la muestra y el contexto de validación y sobre formas específicas de evidencia de validez; este último incluye 10 criterios distribuidos en las cinco fuentes de evidencia de validez ya mencionadas.

Sea que se enfatice en las propiedades de la prueba *per-se* y del constructo evaluado y sus relaciones con otros constructos dentro de la teoría psicológica o que se acepte que también hacen parte de la validez las inferencias y consecuencias de la interpretación de sus resultados, un aspecto importante que ha estado presente en diferentes concepciones de validez, es el relacionado con la invarianza de la medida (Vandenberg & Lance, 2000) y los consecuentes estudios sobre sesgo en las pruebas, y funcionamiento diferencial de los ítems (DIF) y de los test (DTF), dentro de los procesos de validación (Gómez & Hidalgo, 2003; Vandenberg & Lance, 2000). La intervención de variables irrelevantes en la evaluación suele ser un problema al que con frecuencia se ven enfrentados los validadores sobre todo cuando va dirigida a poblaciones compuestas por subpoblaciones con características particulares como idioma, cultura o algún tipo de limitación intelectual o física.

Recientemente, Kane (2013) y Rios y Wells (2014), centrados en los procesos de validación, han propuesto algunas opciones que contribuyan a generar lineamientos para facilitar la labor de los validadores. Sin embargo, esta tarea se hace aún más compleja cuando se contempla la inclusión de nuevas tecnologías en la evaluación, y cuando se buscan medidas equivalentes a partir de los mismos instrumentos a población con características diferenciales. Al observar los mecanismos de evaluación generalmente utilizados para la población con LV, es evidente que en el proceso intervienen variables diferentes que pueden introducir sesgos de la medida. Los TAI como estrategia alternativa para evaluar a esta población, buscan disminuir la probabilidad de que variables extrañas intervengan en el proceso de medida y tratar de mejorar los niveles de precisión en la misma.

Test adaptativos informatizados (TAI)

De acuerdo con Weiss (2004) las características más relevantes de la IRT en el desarrollo de los TAI son la descripción de los ítems en función de sus parámetros, la introducción de la función de información como medio que identifica el nivel de atributo que está siendo mejor medido, y el método de estimación del atributo que supone la existencia de un error de medida específico para cada nivel del mismo.

El algoritmo de los TAI es iterativo, parte de la presentación de un ítem seleccionado mediante una regla de inicio previamente definida, realiza una estimación del atributo con base en las respuestas del individuo, y con ello selecciona el ítem que se presentará a continuación. Este proceso finaliza cuando se ha alcanzado un criterio estipulado para garantizar que se cuenta con una estimación suficientemente precisa del nivel de atributo del individuo. En consecuencia, desarrollar un TAI implica contar con tres requisitos básicos: un banco de ítems calibrado, procedimientos de selección de ítems intermedios, de inicio y de terminación, y un método estadístico de estimación del atributo (Olea & Ponsoda, 2004).

El TAI empieza con una determinada estrategia de arranque. El objetivo de este procedimiento es establecer de alguna forma el nivel inicial de habilidad del individuo, lo cual depende de si se posee o no información previa del mismo (Abad, Olea, Aguado, Ponsoda, & Barrada, 2010). Las estrategias más comunes son: un criterio estándar sobre el nivel de habilidad para cada examinado; la selección del mismo ítem inicial para todos los examinados, por lo general ubicado en una dificultad media; y la escogencia del ítem inicial de manera aleatoria.

Una vez respondido el primer ítem, es necesario establecer el procedimiento para obtener una estimación provisional del nivel de habilidad. Este punto es esencial para todo el proceso debido a que a partir de esta predicción inicial, se seleccionará el ítem siguiente; el estadístico deberá realizar la estimación tras la respuesta a cada ítem hasta determinar con una alta

probabilidad y precisión el nivel de habilidad del examinado. Los métodos estadísticos más utilizados para la estimación de la habilidad son: el Método de Máxima Verosimilitud (ML) y dos procedimientos Bayesianos: el de Estimación Máxima a Posteriori (MAP) y el de Estimación Esperada a Posteriori (EAP). El primero se fundamenta en los datos empíricos y mientras que los métodos bayesianos incorporan información sobre la distribución a priori de los niveles de habilidad de la población o del contexto (Abad et al., 2010).

El método ML no proporciona estimaciones finitas cuando un individuo tiene un patrón regular de respuestas, es decir cuando todos son aciertos o desaciertos (Abad et al., 2010); para solucionar este problema se han desarrollado algunas estrategias que incluyen el uso de métodos bayesianos, y el método propuesto por Dodd (1990) que busca obtener sucesivas estimaciones del nivel del rasgo hasta que sea posible estimar mediante el procedimiento de ML, es decir hasta que se tenga un vector de respuestas de aciertos y fallos. Olea y Ponsoda (2004) modifican parcialmente esta técnica, con el fin de considerar la distribución probable de los niveles de la habilidad de la población proponiendo que el procedimiento se fije en la media o la mediana de una distribución normal y no en el punto medio entre el último valor del rasgo y el parámetro de dificultad.

Por otra parte, las dificultades que presentan los métodos bayesianos incluyen el hecho de que el nivel de habilidad estimado no depende exclusivamente del desempeño de la persona, sino de la distribución a priori del nivel de habilidad en la población, es decir, de los valores de la media y la varianza de la misma (Olea & Ponsoda, 2004); cuando la longitud de una prueba es pequeña el sesgo en las estimaciones aumenta, lo cual representa un problema en los TAI teniendo en cuenta que se aplica un número reducido de ítems. Sin embargo, Abad et al. (2010) señalan que con más de 30 ítems existirán pocas diferencias.

Para seleccionar los ítems que se le presentar al individuo a lo largo de un TAI se suele utilizar el método de Máxima Información (MMI) o el de Máxima Precisión Esperada (MPE). El primero se basa en la cantidad de información que cada ítem ofrece en el nivel de atributo del individuo mientras que el segundo es un método bayesiano que consiste en elegir los ítems que proporcionan una menor varianza de la distribución posterior del nivel de rasgo (Olea & Ponsoda, 2004). El MMI tiende a sobreexponer los ítems y a utilizar los más discriminativos aun cuando estén alejados en la estimación del nivel de habilidad verdadero del individuo.

Finalmente, el criterio de parada o terminación de la prueba puede definirse a través de: (a) un número de ítems determinado, que incluye los procedimientos de longitud fija (LF); (b) cierto valor de precisión de la medida, que involucra el establecimiento de un error estándar de medida (SEM) preespecificado; (c) un criterio que permita clasificar al individuo ubicándolo dentro de un grupo (Muñiz, 1997); o (d) la administración de todos los ítems.

La elección de alguno de los criterios de parada depende de los objetivos de la evaluación y de la distribución de los parámetros de los ítems. Cuando el objetivo de la evaluación tiene implicaciones relevantes para el evaluado, se suele optar por la longitud fija para que el evaluado tenga la sensación de ser examinado en igualdad de condiciones al responder el mismo número de ítems y no menos. Sin embargo, este criterio es problemático ya que con él no se logra un nivel óptimo de precisión para todas las estimaciones de habilidad (Olea & Ponsoda, 2004). En la segunda estrategia, con un SEM preespecificado se estima la desviación estándar de las diferencias entre las puntuaciones verdaderas y las puntuaciones observadas y la prueba finaliza cuando se obtiene el valor establecido o cuando todos los ítems han sido administrados. No obstante, si se fija un SEM muy pequeño, se incrementa el número de ítems a aplicar, y viceversa (Babcock & Weiss, 2012).

Revuelta, Ponsoda, y Olea (1998) señalan que aun cuando se tienen en cuenta todos estos criterios de selección, a causa de que en cada nivel de atributo hay ítems más informativos que otros, resulta probable que individuos con el mismo nivel de habilidad, se les presenten los mismos ítems, por lo que es necesario desarrollar mecanismos de control de las tasas de exposición, y de este modo evitar que entre individuos se dé a conocer el contenido de los ítems. Chen y Liou (2003) y Stocking y Swanson (1998), entre otros, han propuesto diversos métodos para el control de la tasa de exposición de los ítems.

Los TAI presentan muchas ventajas sobre las pruebas de lápiz y papel, como: (a) mayor precisión de la estimación del nivel de atributo al reducir el error de medida (Muñiz & Hambleton, 1999) y no hacer uso de ítems poco informativos (Embretson, 1992); (b) mayor discriminación en los extremos del continuo del atributo valorado (Brown & Weiss, 1977); (c) mayor eficiencia referida al menor tiempo de aplicación, calificación e interpretación de los resultados (Arribas, 2004; Weiss & Betz, 1973), reduciendo la fatiga (Shermis & Lombard, 1998) y favoreciendo una buena actitud del evaluado frente a la prueba (Brown & Weiss, 1977); (d) mayor variedad de tareas que se le pueden presentar al individuo (Embretson, 1992); (e) posibilidad de registrar información relacionada con el proceso de respuesta como los tiempos de exposición de los ítems (Arribas, 2004); (f) mayor seguridad de las pruebas dada la variabilidad de los ítems presentados a cada individuo (Olea & Ponsoda, 2004); y (g) mayor flexibilidad en cuanto a la escogencia del lugar de aplicación de la prueba (Fritts & Marszalek, 2010).

A pesar de estas ventajas, en la literatura revisada no se encuentra información acerca de TAI dirigidos a evaluar población con LV, los trabajos que más se acercan al tema se refieren básicamente a adecuaciones de páginas

web y de equipos como ordenadores con lectores y magnificadores de pantalla, navegadores de internet parlantes, herramientas de reconocimiento de textos impresos OCR parlantes y conversores de Braille; que le permiten a las personas que se encuentran en esta situación de discapacidad, acceder fácilmente a la información.

TAI como alternativa de evaluación para población con limitación visual

El interés por garantizar evaluaciones equitativas cuando se utilizan los mismos instrumentos para poblaciones con características disímiles, ha estado presente a lo largo de la historia de la psicometría, como se ha propuesto recientemente a través del diseño universal para el desarrollo de pruebas (Melo, Nuernberg, & Sancineto, 2013) y se ha plasmado particularmente en las discusiones antes presentadas sobre el concepto de validez. El desarrollo de pruebas psicológicas basadas en Diseño Universal busca el ajuste de los procesos, estrategias y materiales que garanticen su accesibilidad por cualquier evaluado independientemente de su condición mental o física. No obstante, proveer esta garantía para todos los grupos poblacionales evaluados mediante pruebas de aplicación masiva, aún no ha sido posible en muchos contextos, en particular cuando se deben evaluar con adecuados niveles de validez y precisión, de manera autónoma, y en condiciones de equidad, personas con algún tipo de LV para tomar decisiones relacionadas con su acceso a beneficios en el área educativa o laboral, entre otras.

En la actualidad, se ha buscado alcanzar este objetivo mediante la incorporación de tecnologías en los procesos evaluativos, lo cual, sumado a los importantes avances conceptuales y metodológicos en medición psicológica en las últimas décadas, ha generado en esta área de estudio un amplio campo de acción que necesita ser abordado de manera inmediata. Resulta incuestionable la importancia de desarrollar investigaciones dirigidas a mejorar las condiciones de evaluación de la población con LV, generando estrategias que faciliten al evaluado tener mayor autonomía en la forma de abordar la prueba, y al evaluador, obtener niveles adecuados de precisión y validez.

Como se mencionó antes, es amplia la literatura en la que se resaltan las ventajas y beneficios de los TAI en comparación con las pruebas tradicionales (Gómez & Hidalgo, 2003; Olea & Ponsoda, 2004) pero no ocurre lo mismo cuando se busca evidencia acerca de las ventajas que puede representar a la hora de evaluar población con LV. Algunas de ellas, pueden ser: (a) mayor precisión en la estimación de la habilidad del evaluado, (b) control de la interacción de otras variables asociadas a las acomodaciones, (c) obtención de resultados de calidad similar en términos de invarianza de la medida, y (d) posibilidad de registrar información para investigación sobre las diferencias en las dos poblaciones.

Algunos de estos beneficios se derivan del formato computarizado en el que se desarrollan los TAI, motivo por el cual pueden ser considerados además para las pruebas informatizadas no adaptativas, sin embargo, al comparar estos dos formatos, el TAI ofrece ventajas que no se obtienen con las pruebas informatizadas, y que son útiles principalmente cuando se trata de evaluar personas con limitación visual. Referido a la capacidad de ofrecer niveles óptimos de precisión y validez en todo el continuo de la habilidad, aun en niveles extremos del mismo; así como la posibilidad de disminuir factores psicológicos no deseados al momento de la evaluación como la ansiedad. Algunas de estas son: a) mayor precisión en la estimación de la habilidad en niveles extremos del continuo, b) presentación de preguntas acorde al nivel de habilidad demostrado por el examinado, c) menor número de preguntas que componen un test, d) menor tiempo de evaluación.

En primer lugar, el INCI (2010) había hecho notar las diferencias de puntajes en pruebas de aplicación masiva entre videntes y no videntes, y más recientemente el estudio de Herrera, Espinosa, y Soler (2014) encontró que las personas con LV tienden a ubicarse en el extremo inferior del continuo de habilidad cuando se evalúa comprensión de texto con la acomodación del lector; extremos en los cuales las pruebas de lápiz y papel suelen ser menos precisas, como efectivamente se evidenció en el mencionado estudio. Enfocándose en el proceso de evaluación, sea que se acepte la hipótesis de que efectivamente las personas con LV tienen menor nivel de habilidad que quienes no tienen tal limitación, o que se suponga que ese resultado es el producto de la interacción de algunas variables asociadas a la acomodación utilizada en la evaluación que desfavorece a la población con LV, lo que resulta evidente es que las dos poblaciones están siendo evaluadas con diferentes niveles de precisión, lo que pone en riesgo la validez de las inferencias e interpretación de los resultados de la evaluación.

En general, las características de los ítems definidas a partir de adecuadas estimaciones de los parámetros de la IRT, mejoran la eficiencia en la estimación de la habilidad del evaluado con un mayor nivel de precisión en todo el continuo de habilidad y con un número menor de preguntas. Mediante un adecuado algoritmo adaptativo se consigue un mejor equilibrio entre la dificultad del ítem y el nivel estimado de habilidad del individuo, lo que permite una estimación precisa del nivel de rasgo con la exposición de pocos ítems. Efectivamente, Kingsbury & Hauser (2004) encontraron que los TAI son más precisos en los niveles extremos de la distribución de la habilidad, en comparación con las pruebas no adaptativas; en consecuencia, es de esperarse que a través de los TAI se obtengan medidas más precisas para la población con LV sobre todo si se utiliza un criterio de parada variable que permita definir el nivel de error estándar óptimo para el proceso de evaluación.

En segundo lugar, cuando se evalúan personas con LV mediante pruebas de lápiz y papel se suele hacer uso de algunas estrategias para que puedan tener acceso a la información presentada. La acomodación más frecuente es la participación de un lector especializado, que lee cada una de las preguntas y registra las respuestas del examinado. Esta y otras acomodaciones como el uso del sistema Braille, generan fuentes significativas de varianza irrelevante para el constructo que se pretende evaluar, afectando la validez de la evaluación. Algunas de estas variables son: la mayor demanda de recursos atencionales y de memoria, el efecto de fatiga y ansiedad y la poca autonomía del evaluado frente a la prueba.

Se ha encontrado que el acceso de la información por vía auditiva demanda más tiempo para decodificar y reconocer las palabras, e implica un mayor uso de algunos procesos cognitivos como la atención y la memoria de trabajo (Mohammed & Omar, 2011); el acceso a la información por vía auditiva, da lugar a la intervención de variables adicionales como las claves prosódicas (entonación y ritmo, entre otras) que pueden dificultar la fluidez del mensaje transmitido, y por tanto, tener efectos negativos en la calidad de la evaluación (Crowder, 1985). Algo similar ocurre con la evaluación realizada en formato Braille, puesto que para que los individuos puedan diferenciar los grupos de puntos, y asociarlos con sus grafemas y fonemas correspondientes, requieren más de estos procesos cognitivos que cuando se accede a la información por vía visual (Ochaíta, 1988). Sin embargo, esta estrategia es poco utilizada por la poca formación recibida para interpretar esta escritura y los altos costos que conlleva en pruebas masivas.

El efecto de variables extrañas sobre la medición del constructo puede verse significativamente reducido mediante el uso de los TAI ya que dado su carácter computarizado, esta herramienta permitiría la inclusión de software que favorece *la autonomía* del evaluado para interactuar con los ítems y demás material de prueba, brindando la posibilidad de controlar el tiempo que requiere para responder cada tarea o pregunta, hacer lectura y relectura de textos o instrucciones de manera similar a los examinados videntes; y de apropiarse de la situación de evaluación sin depender de un tercero. Estrategias como el software lector de pantalla permitirían que mediante el uso del teclado o el ratón, el evaluado pueda moverse a lo largo del texto de manera similar a la utilizada con personas videntes (Douglas, Kellami, Long, & Hodgetts, 2001). Además, el uso de esta estrategia computarizada podría mitigar el efecto de fatiga y ansiedad que provoca en los examinados la necesidad de solicitarle al lector que repita información, cuando sus recursos atencionales comienzan a escasear.

En tercer lugar, y en estrecha relación con este tema, el estudio de Herrera et al. (2014) mostró que algunos ítems de comprensión de texto presentaron DIF cuando se compararon examinados con y sin LV, explicado desde el juicio de expertos, por variables como la longitud del texto, la presencia de señales como signos de admiración, cursivas o comillas en el mismo, la referencia a frases específicas del texto o a términos particulares del mismo en la pregunta. Tanto las variables relacionadas con la acomodación empleada en la evaluación cuyo efecto puede ser mitigado mediante el uso de TAI, como estas últimas que se refieren a características propias de la construcción de la prueba, tienen sin duda un efecto importante sobre la validez de las pruebas desde la perspectiva de la invarianza de la medida.

El formato computacional en el que se basa el TAI permite el registro fácil y preciso de información relacionada con el proceso de respuesta, que puede soportar investigaciones futuras que permitan comparar la calidad de la evaluación en poblaciones diferentes y proponer nuevas estrategias para garantizarla. Algunas variables de interés pueden ser, por ejemplo, el tiempo empleado en cada pregunta, el número de preguntas que aborda el examinado, el número de lecturas y relecturas de un mismo texto o pregunta, los diferentes intentos de respuesta cuando estos se permiten, e incluso, la posibilidad de tener grabaciones de la verbalización del examinado durante el examen.

El uso del TAI, diseñado para evaluar una población con limitación visual, en preferencia a una prueba informatizada no adaptativa, permite obtener estimaciones del nivel de habilidad del examinado con mayor precisión con un número menor de preguntas, a través del conocimiento previo del error de medida del ítem (función de información del ítem – FFI), así como de la evaluación a través del criterio variable de parada (error estándar de medida – ESM), lo cual garantizaría una evaluación equitativa al aumentar la precisión en los niveles inferiores de la habilidad, en la que habitualmente se ubican las personas con limitación visual; aumenta la confianza en la evaluación; al ser adaptativo puede generar menor ansiedad ya que el examinado responde preguntas con una dificultad acorde a su habilidad; por último, al emplear un número menor de preguntas, se dedica menor tiempo a la evaluación, lo cual permite al examinado mantener niveles adecuados de atención al disminuir bajas de rendimiento asociados a la fatiga.

A pesar del optimismo en el empleo de la herramienta, es necesario considerar algunas exigencias y retos que representa su desarrollo y uso. Si bien es cierto que representa algunas ventajas para el control de algunas variables que afecten la validez de las pruebas, también es cierto que otras variables asociadas al uso del TAI pueden entrar en juego, una de ellas que ha sido la más mencionada es la familiaridad de los evaluados con esta estrategia evaluativa y su habilidad en el uso de computadores y demás componentes asociados. Será necesario entonces diseñar estrategias de capacitación breve que le permita

a los usuarios interactuar con el mecanismo evaluativo, y por tanto, solventar esta dificultad.

Por otra parte, es importante la percepción que tengan tanto examinados como familias, instituciones educativas y demás tomadores de decisiones sobre la "confianza" que pueden tener en los resultados de pruebas más cortas que las tradicionales y con diferentes ítems para cada evaluado. La larga tradición de la TCT y su énfasis en la longitud de las pruebas y la estandarización, y el rigor en el control de las condiciones de su aplicación, constituye para esta nueva perspectiva una barrera que puede ser difícil de superar.

Los argumentos más frecuentemente utilizados en contra del uso de los TAI incluyen el alto costo de desarrollo e implementación, y la exigencia de contar con tecnología apropiada para su aplicación, sobre todo cuando se trata de evaluaciones masivas. Si bien tales costos son considerables, en el caso de evaluaciones de población con necesidades especiales podría considerarse la aplicación de TAI solo para personas con baja visión o invidentes – tal como se usan actualmente las acomodaciones – lo cual reduciría costos de aplicación, al mantener los beneficios de ensamblar para los demás examinados pruebas de lápiz y papel. Un reto muy interesante que se deriva de un esquema como este, consiste en garantizar la equivalencia entre estas medidas con poblaciones diferentes.

En síntesis, se evidencia que a pesar de que los requisitos para desarrollar un TAI están bien definidos en la literatura, "la puesta en funcionamiento y el mantenimiento de un programa de test adaptativos es bastante más complejo" (Wise & Kingsbury, 2000, p. 135), lo que plantea a los desarrolladores de TAI una serie de retos de diferente tipo.

Conclusiones

La evaluación de personas con LV, sobre todo cuando deben participar en procesos masivos de aplicación de pruebas, presenta retos importantes para la psicometría toda vez que los métodos tradicionales de evaluación se caracterizan por el uso de acomodaciones que pueden involucrar el efecto de variables que afectan la calidad de la medida, y que no son consideradas a la hora de interpretar los resultados en las pruebas y de tomar decisiones con base en los mismos. Aunque es reducida la literatura sobre alternativas de evaluación con esta población, algunas investigaciones han mostrado que los resultados en pruebas de lápiz y papel suelen ser inferiores a los de la población sin limitación visual y que estas pruebas suelen tener menor precisión en la evaluación en estos niveles de habilidad.

Estos últimos han mostrado también el efecto de variables asociadas a las acomodaciones frecuentemente utilizadas y a las características de las preguntas mismas, con lo cual se pone en entredicho la validez de los instrumentos y sus resultados. El efecto de algunas de estas variables sobre la invarianza de la medida ataca tanto la evidencia relacionada con el constructo que evalúa la prueba y sus relaciones con otros constructos, como la calidad de las interpretaciones de los resultados y sus consecuencias, sobre todo si se considera la asignación de privilegios y oportunidades con base en ellos.

Existen elementos para considerar que el desarrollo e implementación de TAI pueden ser una alternativa apropiada para evaluar población con LV superando las limitaciones mencionadas. Algunos de ellos son la posibilidad de garantizar precisión similar en todo el continuo de la magnitud de habilidad, la posibilidad de controlar algunas variables asociadas con las acomodaciones frecuentemente utilizadas como la participación de un lector y que afectan la validez de las pruebas y la facilidad en el registro de información sobre el proceso de respuesta. Además, de acuerdo con Huff y Sireci (2001) el TAI brinda la posibilidad de hacer uso de nuevos formatos de ítems, proporcionando "medidas más amplias de un dominio de constructo así como mayor eficiencia en la medición de habilidades cognoscitivas de alto nivel" (p. 17).

Un aspecto que ha sido poco tratado es la identificación, como evidencia de validez, de las consecuencias de la evaluación que no están estrechamente relacionadas con las interpretaciones de las puntuaciones (*American Psychological Association*, *American Educational Research Association* & *National Council on Measurement in Education*, 2014). Puede esperarse que además del control de algunas variables o la facilidad de registro de información para estimar su efecto en la evaluación, una estrategia como un TAI al brindar mayor autonomía y comodidad, pueda aumentar la motivación y el interés de los examinados por el proceso de lectura.

Sin embargo, la adopción de esta estrategia representa algunos retos que deben valorarse adecuadamente antes de emprender la empresa; algunos de ellos son la posible intervención de otras variables asociadas al uso de tecnologías y entornos de evaluación que no son igualmente conocidos para los evaluados, la percepción de los examinados y usuarios de los test y sus resultados ante las nuevas condiciones de aplicación de las pruebas, y los altos costos económicos y exigencias tecnológicas para su implementación.

En síntesis, si se dispone de la infraestructura apropiada para superar estas dificultades y para diseñar bancos de preguntas bien calibrados para el continuo del nivel de habilidad, el desarrollo y uso de TAI para la evaluación de población con LV constituye una estrategia que permitiría superar algunas de las limitaciones que presentan las evaluaciones con lápiz y papel y ayuda de lectores. Sin duda, las mayores ganancias se verán en la calidad de la evaluación en términos de la precisión y la validez al garantizar evaluación con igual calidad en diferentes poblaciones.

Referencias

Abad, F. J., Olea, J., Aguado, D., Ponsoda, V., & Barrada, J. R. (2010). Deterioro de parámetros de los ítems en test adaptativos informatizados: estudio con eCAT. *Psicothema, 22*(2), 340-347.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1966). *Standards for Educational and Psychological Test and Manuals.* Washington: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1974). *Standards for Educational and Psychological Testing.* Washington: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1985). *Standards for Educational and Psychological Testing.* Washington: American Educational Research Association.

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

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

Arribas, D. (2004). Diferencias entre los test informatizados de primera generación y los test en papel y lápiz: Influencia de la velocidad y el nivel de destreza informática. *Acción Psicológica, 3*(2), 91-100.

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length cats provide efficient and effective measurement? *The Journal of Computerized Adaptive Testing, 1*(1), 1-18.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries. Research Report.* 77-6. Psychometric Methods Program, Minneapolis: University of Minnesota.

Chen, P. E., & Liou, M. (2003). Computerized adaptive testing using the Nearest-Neighbours criterion. *Applied Psychological Measurement, 27*(3), 204-216.

Cizec, G. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological test. *Psychological Bulletin, 52*, 281-302.

Crowder, R. (1985). *Psicología de la Lectura.* Madrid: Alianza.

Dodd, B. G. (1990). The effect of item selection procedure and step size on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*(4), 355-366.

Douglas, G., Kellami, E., Long, R., & Hodgetts, I. (2001). A comparison between reading from paper and computer screen by children with a visual impairment. *British Journal of Visual Impairment, 19*(1), 29-34.

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

Embretson, S. E. (1992). Computerized adaptive testing: Its potential substantive contributions to psychological research and assessment. *Current Directions in Psychological Science, 1*(4), 129-131.

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education, 13*(3), 441-458.

Gómez, J., & Hidalgo, M. D. (2003). Desarrollos recientes en psicometría. *Avances en Medición, 1*(1), 17-36.

Herrera, A. N., Espinosa, A. M., & Soler, M. P. (2014). *Análisis bajo modelo de Rasch de una prueba de comprensión lectora en personas con y sin limitación visual.* Artículo en proceso de publicación.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16-25.

Instituto Nacional para Ciegos [INC] (2010). *Análisis de la inclusión social de la población con limitación visual en Colombia. Educación, salud e inserción laboral.* Bogotá: Autor.

Kane, M. (2009). Validating the interpretations and uses of test scores. Em R. W. Lissitz, (Eds.), *The Concept of Validity: Revisions, new directions, and applications* (pp. 39-64). EEUU: Information age Publishing, INC.

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

Kingsbury, G. G., & Hauser, C. (2004), *Computer adaptive testing and the no child left behind act.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California. Recuperado de http://www.nwea.org/sites/www.nwea.org/files/Computerized_Adaptive_Testing_ and_NCLB_0.pdf

Linn, R. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14-16.

Mehrens, W. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-19.

Messick, S. (1995). Standards of validity and the validity of the standards in performance assessment. *Educational Measurement: Issues and practice, 14*(4), 5-12.

Melo, C. Nuernberg, A. H., & Sancineto, C. H. (2013). Desenho universal e avaliação psicológica na perspectiva dos direitos humanos. *Avaliação Psicológica, 12*(3), 421-428.

Mohammed, Z., & Omar, R. (2011).Comparison of reading performance between visually impaired and normally sighted students in Malaysia. *British Journal of Visual Impairment, 29*(3), 196-207.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems.* Madrid: Pirámide.

Muñiz, J., & Hambleton, R. (1999). Evaluación psicométrica de los test informatizados. Em J. Olea, V. Ponsoda & G. Prieto (Eds.), *Test Informatizados: Fundamentos y Aplicaciones* (pp. 23-52). Madrid: Pirámide.

Ochaíta, E. (1988). *Aspectos Cognitivos del Desarrollo Psicológico de los Ciegos (II).* Madrid: Centro de Publicaciones del Ministerio de Educación y Ciencia: CIDE.

Olea, J., & Ponsoda, V. (2004). *Test Adaptativos Informatizados.* España: UNED.

Popham, W. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-14.

Revuelta, J., Ponsoda, V., & Olea, J. (1998). Métodos para el control de las tasas de exposición en test adaptativos informatizados. *Relieve, 4*(2). Recuperado de http://www.uv.es/relieve/v4n2/RELIEVEv4n2_4.htm

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

Shermis, M. D., & Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior, 14*(1), 111-123.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. Em R. W. Lissitz (Ed.), *The Concept of Validity: Revisions, new directions, and applications* (pp. 19-37). EEUU: Information age Publishing, INC.

Stocking, M. L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. *Applied Psychological Measurement, 22*(3), 271-280.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods. 3*(1), 4-70.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70-84.

Weiss, D. J., & Betz, N. E. (1973). *Ability Measurement: Conventional or adaptive? Research Report.* 73-1. Minneapolis: University of Minnesota.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*(1), 135-155.

Wright, B. D. (1968, febrero). *Sample-free test calibration and person measurement.* Paper presented at the National Seminar on Adult Education Research, Chicago, Illinois.

recebido em maio de 2014
reformulado em junho de 2015
aprovado em junho de 2015

Sobre os autores

Aura Nidia Herrera Rojas es Ph.D en Evaluación y tecnología informática en Ciencias del Comportamiento, Universidad de Barcelona, España. Profesora asociada de la Universidad Naciona de Colombia.

Gillen Javier Jiménez es profesional con formación en psicología, candidato a Magister en Psicología de la Universidad Nacional de Colombia en la línea de profundización de psicometría.

Rocio Barajas Sierra es psicóloga egresada de la Universidad Nacional de Colombia, aspirante a Magíster en Psicología de la misma institución. Actualmente, se encuentra vinculada al Instituto Colombiano para la Evaluación de la Educación.