Validez en Test Adaptativos Informatizados (TAI): Evidencia en un TAI diseñado para evaluar comprensión lectora en personas con y sin limitación visual[1]

Rocío Barajas Sierra

Universidad Nacional de Colombia

Facultad de Ciencias Humanas

Departamento de Psicología

Bogotá, 2017

[1] Proyecto enmarcado en el macroproyecto *Diseño de una estrategia integral de evaluación alternativa en personas con y sin limitación visual* —financiado por el Instituto Colombiano para la Evaluación de la Educación (Icfes) y la Dirección de Investigación sede Bogotá (DIB), de la Universidad Nacional de Colombia—.

Validez en Test Adaptativos Informatizados (TAI): Evidencia en un TAI diseñado para evaluar comprensión lectora en personas con y sin limitación visual

Rocío Barajas Sierra

Tesis para optar al título de Magíster en Psicología con énfasis en investigación

Dirigido por:

Aura Nidia Herrera Rojas

Línea de investigación:

Métodos e instrumentos para la investigación en ciencias del comportamiento

Universidad Nacional de Colombia

Facultad de Ciencias Humanas

Departamento de Psicología

Bogotá, 2017

Agradecimientos

A la Dirección de Investigación sede Bogotá (DIB) de la Universidad Nacional de Colombia por el apoyo económico brindado para el desarrollo de este proyecto.

Al Instituto Colombiano para la Evaluación de la Educación (Icfes) por la financiación del macroproyecto *Diseño de una estrategia integral de evaluación alternativa en personas con y sin limitación visual* que enmarca este proyecto.

A la profesora Aura Nidia Herrera Rojas por su apoyo y sus orientaciones en el desarrollo de este proceso.

A Sandra Camargo Salamanca por su acompañamiento en la construcción de este documento.

A los integrantes del grupo de investigación *Métodos e instrumentos para la investigación en ciencias del comportamiento* por sus sugerencias.

Resumen

El presente estudio tuvo como propósito principal brindar evidencia de validez de un Test Adaptativo Informatizado (TAI) que evalúa comprensión de lectura en personas con y sin limitación visual (LV), con el fin de aportar elementos que sirvieran para proponer un modelo general de validación de TAI. Este objetivo se definió con base en dos problemas esenciales: la ausencia de lineamientos prácticos para llevar a cabo la validación de los instrumentos de medida que se enmarcan en la evaluación psicológica, y los retos que los TAI representan al concepto de validez ya que, potencialmente, pueden llegar incorporar fuentes de varianza irrelevante dada la mediación de herramientas computarizadas. Para cumplir este objetivo se diseñó un TAI que evalúa comprensión de lectura, se le aplicó a una muestra de 128 individuos —40 con LV y 88 sin LV— y se recolectaron evidencias de su confiabilidad, mediante la estimación de la Función de Información, de su estructura interna, mediante análisis factoriales de ejes principales, y de los procesos de respuesta de los evaluados, y la relación entre su habilidad y algunas variables de interés, mediante exploraciones bivariadas. Los resultados de este estudio señalan que los elementos explorados pueden ser útiles para el desarrollo de un modelo de procedimiento de validación de TAI y que el uso de TAI para evaluar personas con LV es apropiado ya que genera medidas más equitativas.

Palabras clave: Validez, validación, Test Adaptativo Informatizado (TAI), limitación visual

Abstract

In this study, we provide evidence of the validity of a Computerized Adaptive Test (CAT) for a reading comprehension test in students with and without visual impairment (VI). The goal was to identify some key issues for developing a general model for CAT validation. This goal seeks to satisfy two needs of practitioners. On one hand, there are no guidelines for validating measurement instruments for psychological assessment. On the other hand, CATs challenge the concept of validity since they are potentially affected by irrelevant variance sources due to technological tools. In order to achieve this goal, a reading comprehension CAT was developed and applied to 128 participants —40 of them with VI and 88 without—, and the following pieces of evidence were gathered: CATs reliability estimated through Information functions, internal structure of the test by fitting a principal axes factor analysis, and response processes, and association between estimates ability with several criteria variables. The results show that the considered issues may be useful to develop a general validation model for CATs and may guide other researchers who wish to validate a CAT. Also, the CAT analyzed here is appropriate for assessing individuals with VI because it provides fairer measurements.

Key words: Validity, validation, Computerized Adaptive Testing (CAT), visual impairment

Tabla de contenido

Lista de tablas

Lista de figuras

Introducción

De manera general y desde un enfoque cognitivo, la compresión de lectura se refiere al proceso mediante el cual un lector construye representaciones mentales acerca del mensaje planteado en un texto, por lo que el objetivo principal al leer es extraer el significado del texto como un todo en lugar de quedarse solo con el entendimiento de frases aisladas (Woolley, 2011). Esta definición, que en principio puede parecer simple, esconde una gran complejidad que se hace evidente cuando se indaga sobre los procesos cognitivos involucrados en esta actividad, la forma en la que se lleva a cabo el procesamiento de la información y, derivado de ello, los mecanismos que se han diseñado para evaluar este constructo.

Los procesos cognitivos que intervienen en la comprensión de lectura han sido clasificados en inferiores y superiores. Los inferiores incluyen la descodificación, el reconocimiento de palabras y la segmentación, entre otros; mientras que los superiores incluyen el análisis, la síntesis y, por último, los diferentes niveles de comprensión (Perfetti y Adolf, 2012). La demanda de estos procesos varía en intensidad dependiendo del canal de entrada de la información, la cual puede darse por vía visual, en el caso de personas videntes, o táctil o auditiva, en personas con ceguera o baja visión (Lorenzo, 2001; González, 2004).

Dentro de los modelos que se han propuesto para dar cuenta de cómo se lleva a cabo la comprensión de lectura han predominado los que se fundamentan en la psicología cognitiva. De acuerdo con Barnes (2015), estos modelos se pueden clasificar en dos, en aquellos que tratan de definir cuáles son las habilidades que componen la comprensión de lectura y aquellos que buscan describir el proceso de la comprensión tomando como base procesos cognitivos. Para Pérez (2005), este segundo tipo de modelos se clasifica en: ascendentes, descendentes e interactivos. Los modelos ascendentes consideran que la comprensión de lectura inicia con la ejecución de procesos cognitivos básicos y finaliza con procesos complejos (Pérez, 2005); los modelos descendentes destacan la influencia que tienen los procesos de alto nivel sobre los de bajo nivel (Jiménez, 2004); y los modelos interactivos proponen un procesamiento en paralelo que contempla de manera simultánea la

interacción entre la información proporcionada por el texto y la información previa con la que cuenta el lector (Jiménez, 2004).

Los mecanismos utilizados para evaluar la comprensión de lectura varían dependiendo de si esta es abordada como proceso o como producto (Pérez, 2005). En el primer caso, prima el uso de entrevistas cognitivas, entre otros; y en el segundo, el uso de pruebas de lápiz y papel de respuesta cerrada. Más allá de los argumentos que existan alrededor de la pertinencia o no del uso de estos instrumentos para evaluar atributos tan complejos, se han aunado esfuerzos por generar evaluaciones de alta calidad en estos contextos, evidencia de ello son las pruebas Saber, diseñadas por el Instituto Colombiano para la Evaluación de la Educación (Icfes) con base en el enfoque de competencias propuesto por el Ministerio de Educación Nacional (MEN, 2006); y analizadas bajo la Teoría de Respuesta al Ítem (IRT, sigla del ingl. *Ítem Response Theory*).

No obstante, la mayoría de pruebas de lápiz y papel se fundamenta en la Teoría Clásica de los Test (TCT) y, por lo tanto, representa grandes limitaciones para la evaluación y medición psicológica. En primer lugar, esta teoría supedita las propiedades psicométricas de la prueba y de los ítems a las características de la población evaluada, y las estimaciones del atributo a la prueba utilizada (López y Sánchez, 2005), es decir, que, por ejemplo, una prueba puede llegar a tener valores de confiabilidad tan diversos como grupos poblacionales a los que sea aplicada, y la magnitud de atributo medida en los evaluados puede variar considerablemente entre distintas pruebas aunque compartan el mismo sustento teórico. Por otra parte, dado que el criterio de selección de ítems en las pruebas de lápiz y papel se basa en la maximización de la consistencia interna de la prueba, los ítems seleccionados suelen tener proporciones de acierto cercanas a 0,5, lo cual trae como consecuencia que los individuos que presentan niveles medios de atributo sean evaluados de manera más precisa que aquellos que se ubican en los extremos del continuo (Weiss, 2004). Finalmente, los análisis basados en la TCT implican contar con formas paralelas para realizar comparaciones de puntuaciones entre individuos y con un gran número de ítems para poder asegurar la confiabilidad de la prueba (Gómez e Hidalgo, 2003).

Por lo anterior, se han desarrollado los Test Adaptativos Informatizados (TAI) como estrategia de evaluación alternativa; estos instrumentos adaptan la presentación de los

ítems en función del nivel de atributo manifestado por el individuo a través de sus respuestas (Olea y Ponsoda, 2004), y, al fundamentarse en los planteamientos de la IRT, logran solventar las limitaciones mencionadas, permitiendo que individuos con equivalente nivel de atributo obtengan la misma puntuación a pesar de ser evaluados mediante diferentes ítems, reduciendo los niveles de fatiga experimentados por los individuos y garantizando que todos los niveles de atributo puedan ser evaluados con alta precisión, entre otras ventajas (Weiss, 2004). Sin embargo, los TAI representan grandes retos al concepto de validez.

La *American Educational Research Association* (AERA), la *American Psychological Association* (APA) y el *National Council on Measurement in Education* (NCME), en la versión de sus estándares de calidad para pruebas psicológicas y educativas de 1999 y 2014, definieron validez como «el grado en el cual la evidencia y la teoría soportan las interpretaciones de las puntuaciones de una prueba asociadas con los usos propuestos» (2014; p. 11), y propusieron cinco fuentes de evidencia que se encuentran relacionados con diferentes aspectos de las pruebas: contenido, procesos de respuesta, estructura interna, relación con otras variables y consecuencias.

Esta definición, a pesar de hacer más específicos algunos factores que intervienen en el proceso de validación, aún presenta grandes vacíos como la poca claridad del concepto y la falta de lineamientos prácticos que guíen a los validadores (Kane, 2009); vacíos que han estado presentes a lo largo de la historia de la conceptualización de validez, naciendo en la tradicional trilogía planteada a partir de los preceptos de la TCT, pasando por el enfoque consecuencial fuertemente defendido por Samuel Messick, y desembocando en esta definición dada por la AERA et ál. en 1999 y 2014.

Los estudios que se enfocan en el problema de la validez de los TAI son escasos, sin embargo, es claro que estas herramientas de evaluación presentan retos significativos para sus constructores. De acuerdo con Huff y Sireci (2001), los TAI pueden propiciar la subrepresentación del constructo, generar errores en la puntuación, originar consecuencias sociales imprevistas y, lo más preocupante, introducir fuentes de varianza irrelevante para el constructo dada la mediación de herramientas computarizadas. Se hace entonces imprescindible la realización de estudios que aborden el problema de validez de estas

herramientas y que ayuden a generar lineamientos para sus validadores, puesto que desde hace años se postula como una de las estrategias de evaluación más prometedoras dada la inclusión de la tecnología y las ventajas que representa en términos de movilidad y reducción de tiempo en las aplicaciones, entre otras.

Este trabajo se enmarca dentro del proyecto *Diseño de una estrategia integral de evaluación alternativa en personas con y sin limitación visual (LV)* que tuvo como objetivo principal diseñar una estrategia integral para evaluar comprensión de lectura en poblaciones con y sin LV, y dar cuenta de una primera evidencia empírica acerca de su validez. Asimismo, tiene como insumo los resultados del proyecto *Procedimiento para establecer equivalencia en las puntuaciones de pruebas de aplicación masiva, en personas con y sin LV* llevado a cabo entre 2010 y 2012 en el que se identificaron y sometieron a revisión de contenido los ítems que funcionaban diferencialmente entre videntes y no videntes en la subprueba de Lenguaje de la prueba Saber 11.º, y se propuso el uso de procedimientos de equiparación para establecer equivalencias entre las puntuaciones obtenidas por individuos de las dos poblaciones en la misma prueba.

Teniendo en cuenta estos antecedentes, el presente trabajo pretende brindar evidencia de validez de un TAI que evalúa comprensión de lectura en personas con y sin LV, con el fin de aportar elementos que sirvan para proponer un modelo general de validación de TAI; y tiene como objetivos específicos: (a) Brindar evidencia de la confiabilidad de un TAI que evalúa comprensión de lectura en personas con y sin LV, en términos de Función de Información; (b) brindar evidencia empírica de la validez de un TAI que evalúa comprensión de lectura en personas con y sin LV, a través de la evaluación de la estructura factorial de la prueba; (c) proponer una relación entre variables que describa el desempeño de personas con y sin LV en un TAI que evalúa comprensión de lectura; y (d) evaluar la percepción de las personas con y sin LV sobre la validez de un TAI que evalúa comprensión de lectura.

Revisión bibliográfica

Comprensión de lectura

La diversidad de factores que intervienen en la comprensión de lectura y que hacen de ella una actividad superior de gran complejidad ha conducido a los investigadores a formular diversas teorías en torno a cómo se lleva a cabo, cuáles son los procesos psicológicos involucrados, y, derivado de ello, cuáles serían los mejores mecanismos para evaluarla. Como en todos los contextos académicos, estas teorías explicativas difieren entre sí según en el enfoque que se haya adoptado para formularlas, así como el momento histórico en el que fueron desarrolladas.

Hasta mediados de la década de los noventa prevalecía la noción de comprensión de lectura como proceso que se reducía al «reconocimiento, la memorización y la interpretación de palabras y signos visuales» (Soler, 2013; p. 13), un ejemplo de ello es la definición dada por Watson, Wright, Long y De L'Aune (1996): proceso que implica la integración visual del texto que se lee con información que se encuentra almacenada en la memoria. Sin embargo, con el tiempo se han ido incorporando más elementos en estas definiciones, tal como se evidencia en la concepción de comprensión de lectura del Grupo de Estudio de Lectura de la Corporación RAND (Research and Development): «proceso de extraer y construir significado de manera simultánea, a través de la interacción y la participación con el lenguaje escrito» (p. xiii). Esta definición incorpora tres elementos que interactúan entre sí enmarcados en un contexto sociocultural: el lector, el texto y la actividad (Snow, 2002).

El *lector* hace referencia a todo lo que este le aporta a la lectura, es decir, sus habilidades, conocimientos y experiencia; el *texto* incluye material escrito presentado en cualquier modalidad visual; y la *actividad* abarca los propósitos al realizar la lectura y las consecuencias derivadas de esta (Snow, 2002). La presencia de estos elementos en la definición de comprensión de lectura, al igual que el contexto sociocultural y la noción de interacción, ha sido ampliamente acogida por otros teóricos, tal es el caso de Pérez (2005) que menciona el rol de los conocimientos previos del lector para interpretar la información presentada por el autor del texto, y de Woolley (2011) que clasifica estos elementos como texto, tarea, características del lector y propósito de la actividad. Para estos autores, estos

elementos contribuyen al éxito o fracaso en la extracción de significado de los textos a los que se expone un individuo, extracción que, de acuerdo con Vallés (2005), es un proceso gradual, progresivo y no-lineal.

Los modelos teóricos que enmarcan la mayoría de estas definiciones y que hasta el momento han predominado son los de índole cognitivo. Para Barnes (2015), en función de su enfoque, estos modelos se pueden clasificar en: a) modelos que buscan determinar cuáles son los componentes que explican la varianza en tareas de comprensión de lectura, es decir, los conocimientos y las habilidades —a nivel de palabra y de texto—; y b) modelos de procesos que buscan «describir cómo las representaciones del texto se construyen a través de una serie de procesos cognitivos iterativos que sirven para mantener la coherencia local y global», y que «tienen en cuenta cómo las características del lector y del texto influyen en la comprensión» (p. 5). Este segundo tipo de modelos, de acuerdo con Pérez (2005) se puede clasificar en modelos ascendentes, descendentes o interactivos.

Los modelos del primer tipo incluyen la Visión Simple de la Lectura (SVR, sigla del ingl. *Simple View of Reading*) y el Modelo de Mediación Directa e Inferencial (DIME, sigla del ingl. *Direct and Inferential Mediation Model*), entre otros; y los modelos del segundo tipo incluyen el Modelo de Construcción-Integración de Kintsch y el Modelo de Paisaje, entre otros.

La SVR propuesta por Gough y Tunmer en 1986 sostiene que la lectura equivale al producto de la capacidad de descodificación, es decir, la capacidad para reconocer palabras, y la comprensión lingüística o auditiva, es decir, el uso de información a nivel de palabra para lograr interpretar el sentido del texto en general (Gough y Tunmer, 1986; Hoover y Gough, 1990). Las fallas en alguno de estos elementos desembocan en algún tipo de «discapacidad de lectura», como lo llaman los autores. Los tres supuestos de este modelo son: la descodificación de palabras es necesaria para la comprensión de lectura, los conocimientos y las habilidades involucradas en la comprensión lingüística equivalen a lo que se requiere para entender el texto, y la multiplicación de la descodificación y la comprensión lingüística explican gran parte de la varianza de la comprensión de lectura (Barnes, 2015).

Estos supuestos han sido probados por diversos investigadores, entre ellos, Hoover y Gough (1990) que obtuvieron resultados que apoyaban el modelo, y Dreyer y Katz (1992) que, si bien no encontraron evidencias concluyentes sobre este, sí encontraron evidencias de que los componentes de descodificación y comprensión lingüística son factores esenciales en la comprensión de lectura ya sea asumiéndola como producto de estos dos componentes o como el resultado de su suma. Otros autores le han adicionado a este modelo componentes como la velocidad (Joshi y Aaron, 2000), la memoria de trabajo y la conciencia fonológica (Ahmed et ál., 2016), para hacerlo más preciso.

Por su parte, el Modelo DIME propuesto por Cromley y Azevedo (2007) supone que hay cinco predictores de la comprensión de lectura —conocimientos previos, inferencias, estrategias, vocabulario y lectura de palabras— que se relacionan entre sí y que tienen efectos directos e indirectos sobre la comprensión. De acuerdo con Ahmed et ál. (2016) este modelo se deriva directamente del modelo SVR puesto que la lectura de palabras representa el componente de descodificación, y los conocimientos previos, las inferencias, las estrategias y el vocabulario representan el componente de comprensión lingüística.

Las hipótesis de este modelo son: a) las relaciones entre las cinco habilidades mencionadas dan lugar a la comprensión de lectura, b) los conocimientos previos tienen un efecto directo en la comprensión cuando se trata de comprensión literal, c) el conocimiento previo, con frecuencia, es necesario para implementar estrategias y extraer inferencias, d) las estrategias y las inferencias pueden afectar de forma directa a la comprensión, e) las estrategias implementadas (p. ej. resumir) pueden favorecer la elaboración de inferencias que a su vez tienen efecto en la comprensión, f) la lectura de palabras y el vocabulario tiene efectos directos e indirectos sobre la comprensión, y g) los conocimientos previos y la lectura de palabras están correlacionadas entre sí pero no se afectan de forma directa (Cromley y Azevedo; 2007).

Estas hipótesis, además de haber sido probadas por los autores del modelo con textos informativos y narrativos en una muestra de estudiantes de secundaria fueron probadas por Cromley, Snyder-Hogan y Luciw-Dubas (2010) con textos de dominio específico en estudiantes universitarios. Los resultados de este estudio reflejaron un adecuado ajuste del modelo incluso al haber variado las condiciones con las que fue validado originalmente, es

decir, en tamaño de muestra, tipo de texto, edad de los evaluados, entre otros; no obstante, para que este modelo mostrara un ajuste aún mejor se agregó un efecto directo del vocabulario en las estrategias implementadas por los lectores. Para Ahmed et ál. (2016), estos estudios experimentales representan un gran avance en el desarrollo de modelos de este tipo puesto que la mayoría suele basarse en aproximaciones teóricas, únicamente.

Uno de los modelos más conocidos que se fundamenta en la descripción de los procesos cognitivos para dar cuenta de cómo se lleva a cabo la comprensión de lectura, y que de acuerdo con la clasificación de Pérez (2005) hace parte de los modelos ascendentes, es el Modelo de Construcción-Integración propuesto por Kintsch (1988, 1998). Este modelo se basa en algunas de las premisas del Modelo Proposicional de Kintsch y Van Dijk (1978) que considera que el lector inicia la construcción del significado del texto a nivel de proposiciones y, luego, por medio de «macrooperadores», estas son transformadas o resumidas en macroproposiciones que representan el significado del texto a nivel global. En el modelo de Kintsch (1988, 1998), e incluso en la versión posterior de este modelo proposicional, la noción de «macrooperadores» como determinante para construir el significado del texto fue removida, y se introdujeron otros cambios para hacer que el modelo tuviera un carácter más dinámico y, con ello, se lograra dar una descripción más fiel sobre cómo se llevaba a cabo la comprensión de lectura (ver Van Dijk y Kintsch, 1983).

De acuerdo con Kintsch (1998), existen dos tipos de modelos mentales, los *modelos basados en el texto* que, en palabras de Woolley (2011), corresponden a «una representación mental del discurso real del texto» (p. 16) puesto que se basan en la intencionalidad del autor, y los *modelos de situación* que corresponden a lo que el lector considera que se trata el texto teniendo como base sus conocimientos previos. Según Stinnett (2009), el lector debe apelar a este segundo modelo cuando encuentra brechas en el primer modelo que le impiden construir el significado de lo que lee. Estos dos tipos de modelos operan en dos niveles: en una estructura local o microestructura, es decir, a nivel de la información de cada palabra y oración; y en una estructura global o macroestructura, la cual está ordenada jerárquicamente por las oraciones que representan el significado del texto en general.

Este modelo ha sido ampliamente acogido por el importante rol que se le atribuye tanto al texto como al lector, y por la simultaneidad de los procesos involucrados. Rapp y Van den Broek (2005) introducen algunos de estos conceptos en su Modelo de Paisaje de la comprensión de lectura, el cual se encuentra enmarcado en los modelos que consideran que este constructo es un proceso continuo y dinámico que involucra fluctuaciones en la activación de conceptos a medida que se avanza en la lectura e interacciones entre el lector y el texto que afectan y son afectadas por estas fluctuaciones. De acuerdo con estos autores, el proceso de comprensión de lectura tiene una naturaleza cíclica donde cada ciclo corresponde a una oración, y donde entre ciclo y ciclo se presentan fluctuaciones en la activación de conceptos. «Estas fluctuaciones dan lugar a un "paisaje" de activaciones, con conceptos que crecen y disminuyen en activación durante el transcurso de la lectura» (Rapp y Van den Broek, 2005; p. 277). Desde que inicia el primer ciclo se produce una representación en la memoria que se actualiza a medida que se avanza a lo largo de los ciclos, lo que al final produce una representación estable de la lectura.

Este modelo surgió como una forma de integrar todas las «miniteorías» que se habían formulado hasta entonces y que solo se centraban en la definición de los componentes de la comprensión de lectura, sin abordar la interacción entre estos (Rapp y Van den Broek, 2005); lo que, según Barnes (2015), resultó ventajoso para el modelo puesto que le permitió tener un poder predictivo superior al de las «miniteorías» sobre las que se estructuró.

Como se puede observar en los planteamientos anteriores, si bien las posturas respecto a cómo se lleva a cabo el proceso de la comprensión de lectura son muy variadas, en la mayoría de los casos, resultan ser complementarias, lo que, en teoría, podría facilitar la formulación de un modelo integral de comprensión de lectura que abarcara todos sus elementos y que fuera aplicable a todas las situaciones de lectura, tal como lo proponen McNamara y Magliano (2009); pero, que, hasta el momento, solo ha permitido llegar a acuerdos en relación con algunos factores que guardan estrecha relación con este constructo, como es el caso de ciertos procesos cognitivos.

De acuerdo con Seigneuric y Ehrlich (2005), estos procesos involucran desde aquellos que se ejecutan a nivel de grafema hasta aquellos que se dan a nivel de palabra, oración y texto completo, bien sea que se considere que la comprensión de lectura sigue un

procesamiento de abajo hacia arriba o de arriba hacia abajo. Vallés (2005) clasifica estos procesos en *perceptivos,* los cuales permiten la identificación y la decodificación de los grafemas que componen el texto; *básicos,* que abarcan el reconocimiento de las palabras y su asociación con conocimientos almacenados en la memoria, entre otras acciones; y *cognitivo-lingüísticos,* que permiten el acceso al significado del texto y la extracción de la información necesaria para poder comprenderlo.

Los procesos *perceptivos* se ponen en marcha en el momento en el que el lector se enfrenta al texto, ya sea mediante la vista o el tacto, y son los encargados de transmitirle al cerebro la información que se presenta allí (Vallés, 2005). Si la información ingresa al cerebro por vía visual, la lectura se realiza por medio de *fijaciones* y *movimientos sacádicos*. Las primeras se refieren a las pausas requeridas para identificar las letras, las sílabas y todo aquello que se encuentre alrededor del punto en el que se focalice la vista, y los segundos, a los deslizamientos de la vista entre los puntos de fijación (Bäckman, 1999). Los lectores expertos tienden a hacer fijaciones más cortas y movimientos sacádicos más largos que los lectores principiantes, omitiendo palabras de función, como preposiciones, conjunciones, artículos y pronombres, y fijando las palabras de contenido, como sustantivos, verbos, adjetivos y adverbios, en particular, si estas no les resultan familiares (Singleton, 2008).

Si la información ingresa al cerebro por vía táctil, es decir, si el texto se encuentra codificado en braille, la lectura se realiza por medio del patrón disjunto de reconocimiento, el cual implica el uso simultáneo del dedo índice de cada mano, siendo el de la mano derecha el que se utiliza para reconocer la grafía y el de la mano izquierda para corroborarla (Vallés, 2005). El hecho de que el lector deba ir reconociendo letra por letra para poder avanzar, hace de este reconocimiento un proceso secuencial (Singleton, 2008), y, por ende, menos veloz que aquel que se realiza haciendo uso de la vista (Veispak, Boets y Ghesquière, 2012).

Dentro de los procesos *básicos*, Vallés (2005) incluye la atención selectiva, que le permite al lector concentrarse en lo que está leyendo; el análisis secuencial y la síntesis, que le facilitan la vinculación de los significados de cada palabra nueva que aborda con los de las palabras anteriores; y la memoria, que, según Gómez, Vila, García, Contreras y Elosúa

(2013), tiene dos funciones esenciales para el lector: almacenar la información que va extrayendo del texto y evocar sus conocimientos previos para darle sentido a lo que está leyendo.

Para estos autores, si bien la memoria largo plazo desempeña un rol importante durante el proceso de comprensión, el tipo de memoria sobre la que recae la mayor carga cognitiva durante la lectura es la memoria de trabajo puesto que esta interviene en la recuperación y la integración de información a nivel de palabra —manteniendo el significado en función del contexto dado por el texto—, oración —almacenando las ideas contenidas en cada una de ellas— y texto —permitiendo la construcción de un modelo mental coherente sobre el significado dado al texto completo—. Lo anterior significa que la capacidad del lector para comprender un texto depende, en gran medida, de su capacidad de memoria de trabajo, tal como lo confirma el estudio de Seigneuric y Ehrlich (2005) en el que el desempeño de niños en tareas de este tipo de memoria resultó ser un buen predictor de su nivel de comprensión de lectura.

Finalmente, siguiendo con la clasificación de Vallés (2005), los procesos *cognitivo-lingüísticos* incluyen el acceso al léxico, que se refiere a la asociación que debe realizar el lector entre las palabras que lee y la información que tiene almacenada en la memoria a largo plazo para identificar el significado de cada una de ellas; el análisis sintáctico, que se refiere a la configuración del significado de las oraciones a partir de las relaciones estructurales entre las palabras que las conforman; y la interpretación semántica, que da lugar a la representación del significado del texto como un todo y que implica la elaboración de inferencias, tanto para establecer relaciones entre los significados de las oraciones y el contexto del texto, como para atribuirle significados al contenido de lo que se lee cuando existen dudas; inferencias que «pueden alcanzar diferentes niveles de elaboración en función de la capacidad cognitiva (imaginación, motivación por el tema de la lectura, memoria, abstracción, generalización, etc.) del lector» (p. 59).

En el caso del proceso de comprensión de lectura de las personas con LV, la demanda de estos procesos cognitivos presenta variaciones importantes que pueden ser más o menos profundas dependiendo de las estrategias a las que deba recurrir el lector para lograr acceder a la información que se presenta en el texto. Un ejemplo de ello se muestra

en el estudio de Gompel, Van Bon y Schreuder (2004) en el que se encontró que las personas con baja visión deben dedicar más tiempo a extraer la información del texto del que dedican las personas sin ningún tipo de deterioro visual, lo cual, de acuerdo con Bosman, Gompel, Vervloed y Van Bon (2006), se debe a que estas personas presentan dificultades para fijar el texto e identificar las letras, pueden ver de forma simultánea una menor cantidad de información a causa de que tienen un campo visual periférico más reducido, están menos expuestos al aprendizaje incidental y pueden experimentar menores niveles de motivación para leer.

Para algunos autores, las personas con baja visión pueden llegar a compensar esta menor velocidad de descodificación dependiendo en mayor medida de la información que suministra la oración en la que se encuentra incrustada la palabra que intentan descodificar (Gompel, Van Bon y Schreuder, 2004), lo que puede llegar a hacer comparables sus niveles de comprensión de lectura con los de las personas sin ningún tipo de LV; sin embargo, la evidencia recopilada para apoyar esta idea ha resultado contradictoria, tal como lo señalan Omar y Mohammed (2005). Según estos autores, al tener una menor velocidad de descodificación, las personas con baja visión requieren una mayor capacidad de procesamiento y de memoria de trabajo, y, por lo tanto, presentan menores niveles de comprensión de lectura que las personas sin LV.

Las variaciones en la demanda de los procesos cognitivos involucrados en la comprensión de lectura de carácter más profundo se presentan en las personas que, debido a la gravedad de su LV, deben acceder a los textos haciendo uso de un canal de entrada diferente al visual. Al respecto, en el estudio realizado por Veispak et ál. (2012) con niños y adultos, se encontró que los individuos que acceden a la información por vía táctil (braille), además de leer con menor rapidez, también lo hacen con menor precisión, lo cual, resulta particularmente cierto cuando se trata de lectores jóvenes e inexpertos. Para estos autores, las razones de esta menor precisión radican en la dificultad para identificar las letras, puesto que, si el lector pasa por alto uno de los puntos que componen cada una de ellas, la identificación de dicha letra puede resultar incorrecta.

Lo anterior, para Ochaíta, Rosa, Fernández y Huertas (1988) se traduce en que las personas que acceden a la información mediante este canal requieren una mayor cantidad de

recursos atencionales y memorísticos para activar el fonema adecuado de cada letra e ir configurando las palabras; requerimiento que, de acuerdo con Lorenzo (2001), también aplica para el caso de las personas que capturan la información mediante el oído puesto que el mensaje oral tiene características como el volumen, las pausas, la fluidez y el ritmo, entre otras, que no se aplican al mensaje escrito e implican un esfuerzo adicional para lograr filtrar, retener y asociar la información relevante y, por tanto, comprender lo planteado en el texto.

Estas variaciones en el proceso de la compresión de lectura, junto con las vistas en los párrafos anteriores, han hecho que el diseño de mecanismos para evaluar este constructo deba enfrentar grandes desafíos, que se hacen visibles desde la definición de los objetivos de la evaluación y la elección del modelo teórico de base, hasta la compensación de las limitaciones impuestas por las particularidades de los instrumentos de medida a la evaluación de este constructo, tal como se expone en el siguiente apartado.

Evaluación de la comprensión de lectura

De acuerdo con Pérez (2005), los mecanismos de evaluación de la comprensión de lectura se pueden clasificar en dos categorías dependiendo de si se quieren abordar los *procesos* que se llevan a cabo durante la comprensión, o si se quieren abordar sus *productos*. Para Rapp y Van den Broek (2005), esta clasificación parece desconocer que los dos tipos de evaluación se encuentran interconectados, ya que, por un lado, «lo que ocurre durante la lectura debe ser de algún modo el fundamento de lo que el lector retiene después» (p. 277) y, por el otro, «los lectores no esperan hasta que la lectura esté completa para empezar a construir sus representaciones» (p. 277) sino que las van construyendo a medida que van leyendo el texto. No obstante, esta clasificación resulta útil para efectos de la ejemplificación de los mecanismos de evaluación desarrollados para evaluar la comprensión de lectura.

Algunos tipos de *medidas de proceso* son el registro del tiempo de lectura y el seguimiento de los movimientos de los ojos a medida que se va leyendo el texto, y de *medidas de producto*, la evocación del contenido del texto y la aplicación de lo leído en situaciones novedosas (Rapp y Van den Broek, 2005). Aunque, según Pérez (2005), «la utilización de las medidas de producto supone una visión restrictiva de la comprensión

lectora que no coincide con los actuales desarrollos teóricos» (p. 125), debido a los intereses de los investigadores y a las facilidades que brindan este tipo de medidas en términos de aplicación y calificación, la mayoría de instrumentos diseñados para evaluar este constructo corresponden a estas medidas.

Los ítems más utilizados para conformar estos instrumentos incluyen los de recuerdo libre, que le exigen al evaluado producir un texto que refleje su comprensión de lo leído; los de respuesta abierta, que también le exigen al evaluado dicha producción, pero en relación con aspectos particulares del texto; los de verdadero/falso, en los que los evaluados deben determinar el valor de verdad de algunas afirmaciones derivadas del texto leído; y los de elección múltiple, en los que los evaluados deben elegir, entre varias opciones, la respuesta que consideren que coincide con la información presentada en el texto (Pérez 2005). Otros ítems que han sido ampliamente utilizados y que, a diferencia de los anteriores, no parten de la lectura previa de un texto, son los ítems de espacios en blanco («cloze»), en los cuales se les presenta a los evaluados algunos pasajes de texto a los que les ha sido borrada una palabra, y, con base en las pistas contextuales del pasaje, los evaluados deben inferir cuál es la palabra faltante (Watson et ál., 1996).

Cada uno de estos tipos de ítems tiene ventajas y desventajas, por ejemplo, si bien los ítems de recuerdo libre son fáciles de aplicar, les exigen a los evaluados el uso de habilidades expresivas que pueden no estar relacionadas de forma directa con su capacidad de comprensión de lectura y, además, al partir de la idea de que para que haya una producción textual lo leído debe haber sido comprendido, no tienen en cuenta que el evaluado podría llegar a configurar su respuesta haciendo uso solo de recursos memorísticos. Algunas de estas dificultades también se observan en los ítems de respuesta abierta, sin embargo, este tipo de ítems presenta ventajas sobre el de recuerdo libre al facilitar que se lleven a cabo tipos de procesamiento diferentes a los memorísticos (Pérez, 2005).

Por su parte, aunque los ítems de verdadero/falso tienen como ventajas que pueden ser calificados de forma muy rápida y que no le exigen al evaluado una producción textual, presentan una alta probabilidad de ser acertados por azar y un bajo valor diagnóstico. Los de elección múltiple reducen la probabilidad de ser respondidos al azar, pero dificultan la tarea de los constructores de ítems al tener que garantizar que solo una de las opciones

sea correcta y, además, le pueden demandar al evaluado estrategias de resolución de problemas en lugar de las requeridas para comprender un texto (Pérez, 2005). Finalmente, los ítems de completar espacios en blanco si bien resultan fáciles de aplicar, le imponen altas demandas a la memoria de trabajo del lector debido a que este debe integrar la información presentada en las frases que se encuentran alrededor de la palabra faltante para poder identificarla (Mostow et ál., 2004).

Para la conformación de pruebas de comprensión de lectura de aplicación masiva, al igual que para el resto de pruebas de este tipo, se ha optado por el uso de ítems de elección múltiple con única respuesta debido a que, como se mencionó, estos ítems resultan más fáciles de aplicar y calificar que otros y, además, permiten evaluar de forma óptima una gran multiplicidad de habilidades. En Colombia, un ejemplo de este tipo de instrumentos de aplicación masiva es la subprueba de *Lenguaje* (transformada a mediados de 2014 en *Lectura crítica*) de la prueba Saber 11.º desarrollada por el Icfes, en la que se concibe la comprensión de lectura como una competencia transversal a todas las áreas de conocimiento y cuya estructura de prueba fue formulada con base en los supuestos del Modelo Proposicional de Kintsch y Van Dijk (Icfes, 2016).

De acuerdo con esta estructura, la comprensión de lectura está dada por el cruce de acciones y componentes. Las *acciones* corresponden a aquello que el individuo debe ejecutar sobre los contenidos de un texto e incluyen la *acción interpretativa*, que le implica al lector constituir los sentidos que circulan en el texto; la *acción argumentativa*, que le exige al lector explicar las ideas que articulan y le dan sentido al texto, y la *acción propositiva*, que le demanda al lector incorporar sus conocimientos previos para plantear alternativas ante las situaciones expuestas en el texto. Los *componentes*, por su parte, corresponden a las unidades del texto en las que el individuo debe centrar sus acciones, e incluyen: el *componente de función semántica de la información local*, que indaga por la función que cumplen los elementos microestructurales del texto (adjetivos, sustantivos, etc.) para darle sentido a este; el *componente de configuración del sentido global del texto*, que indaga por el sentido el texto a nivel macroestructural e implica la conformación de relaciones entre el contenido explícito e implícito del texto; y *el componente del sentido del texto hacia otros textos*, que indaga por la relación que existe entre el contenido de dos o más textos (Icfes, 2012).

Para evaluar estos cruces de acciones y componentes, el Icfes ha hecho uso de una amplia variedad de textos que ha clasificado de acuerdo con su forma y su temática. En la clasificación formal diferencia entre los *textos continuos*, que siguen una estructura lineal y están conformados por párrafos; los *textos discontinuos*, que no siguen una estructura lineal, tales como gráficos, tablas, etc.; y los *textos mixtos*, que incorporan los dos tipos de formato anteriores. En la clasificación temática diferencia entre los *textos literarios*, que incluyen cuentos, novelas, poesías y obras de teatro; y los *textos informativos*, que incluyen los expositivos, los argumentativos y los descriptivos (Icfes, 2016).

A la hora de conformar instrumentos que evalúen comprensión de lectura es frecuente que no se realice este tipo de clasificaciones, ni que se tengan en cuenta otras características de los textos como su longitud, la densidad de su contenido y la cantidad de información nueva que le provee al lector; características que, de acuerdo con Pérez (2005), han sido ampliamente estudiadas y han demostrado tener influencia directa en el desempeño de los evaluados en tareas de comprensión de lectura. Lo anterior, según Gorin (2005), puede llegar a afectar la validez de las interpretaciones que se realicen a partir de las puntuaciones derivadas de estos instrumentos al pasar por alto la evaluación de ciertas habilidades requeridas para la ejecución de este constructo.

Para autores como O'Reilly, Weeks, Sabatini, Halderman y Steinberg (2014) y Hansen, Lee y Forer (2002) este problema de que los instrumentos diseñados para evaluar comprensión de lectura aborden diferentes partes de la comprensión e incluso constructos diferentes y, por lo tanto, no resulte claro lo que representan las puntuaciones que se obtienen a partir de estos instrumentos, puede deberse no solo a las características de los textos que se utilicen para formular las preguntas sino también a las restricciones propias del tipo de instrumento que se decida usar. Para O'Reilly et ál. (2014) estas restricciones incluyen los costos de desarrollo de los instrumentos, los límites de tiempo destinados para su aplicación y la forma en la que deben ser entregados sus resultados; y para Hansen et ál. (2002) incluyen los mecanismos utilizados para su aplicación.

En relación con este último aspecto, Hansen et ál. (2002) resaltan que las dificultades potenciales de los mecanismos de aplicación se hacen particularmente visibles en el caso de la evaluación de las personas con LV que, por su condición, no pueden acceder a la

información por vía visual y para las cuales se han diseñado estrategias de aplicación que incorporan el uso de cuadernillos de prueba escritos en sistema braille o de lectores entrenados para que les lean la prueba y marquen la respuesta que ellas elijan para cada pregunta.

La primera estrategia tiene como problema principal que muy pocas personas con LV saben cómo manejar este sistema de escritura, lo cual resulta particularmente cierto en el caso de Colombia por varias razones, la primera radica en que las instituciones educativas públicas del país no cuentan con los recursos suficientes para la enseñanza del braille (Gómez y González, 2008) y son precisamente estas instituciones a las que tiene acceso la mayoría de personas con LV ya que, de acuerdo con Cañón (2011), alrededor del 80 % pertenece a los estratos 1 y 2; y la segunda razón radica en que el porcentaje de personas con LV que se encuentra escolarizado es muy bajo: en 2011, de acuerdo con el INCI (2013), este porcentaje era del 44 % en el caso de los niños entre los 3 y los 5 años, y, de acuerdo con el MEN (2012), del 58,4 % en el caso de los niños entre los 5 y los 11 años.

Las dificultades que presenta la segunda estrategia, que suele ser la más utilizada, incluye la calidad inconsistente de la lectura por parte de la persona de apoyo, la ansiedad del evaluado al tener que depender de otro, la vergüenza que puede sentir el evaluado de tener que pedirle al lector que le vuelva a leer alguna parte del texto, los errores que pueden llegar a cometer los lectores al marcar las respuestas del evaluado, la fatiga que puede experimentar el evaluado por la lentitud del proceso y por mantener un diálogo con el lector, y el tiempo adicional que se requiere para completar la prueba.

Debido a estos inconvenientes, autores como Douglas, Kellami, Long y Hodgetts (2001) se han interesado en probar formas de aplicación que impliquen la mediación de herramientas computarizadas y que ayuden a disipar los problemas mencionados, encontrando que el uso de estas herramientas representa grandes ventajas para la evaluación de personas con LV, tales como la posibilidad de responder la prueba de manera independiente y de modificar la forma de presentación de los ítems en función de las necesidades de estas personas. Ventajas que, de acuerdo con Olea y Hontangas (1999), entre otros, se pueden hacer aún más grandes haciendo uso de TAI, ya que, como se verá en el siguiente apartado, esta estrategia de evaluación genera medidas más precisas (Muñiz y Hambleton, 1999), reduce el tiempo de aplicación (Weiss y Betz,

1973) y los niveles de fatiga experimentados por los evaluados (Shermis y Lombard, 1998) y brinda otras ventajas que pueden favorecer la realización de evaluaciones más equitativas al garantizar la obtención de medidas más justas.

TAI como estrategia de evaluación alternativa

Los TAI son pruebas computarizadas, desarrolladas a partir de la IRT, que adaptan la presentación de los ítems en función del nivel de atributo manifestado por el individuo a través de sus respuestas (aciertos/desaciertos) y buscan seleccionar para cada examinado el grupo de preguntas que estima con mayor precisión y eficiencia su nivel de atributo (Olea y Ponsoda, 2004). De acuerdo con López y Sánchez (2005), el mecanismo subyacente a este tipo de instrumentos (mostrado en la figura 1) inicia con la presentación de un ítem que ha sido seleccionado según una estrategia de arranque definida previamente, luego, con la respuesta dada por el individuo a este primer ítem, realiza una estimación provisional de su nivel de atributo y, con base en esta estimación, elige el siguiente ítem y repite el proceso de estimación hasta que se haya satisfecho cierto criterio de parada, culminando con una estimación definitiva del nivel de atributo del evaluado.

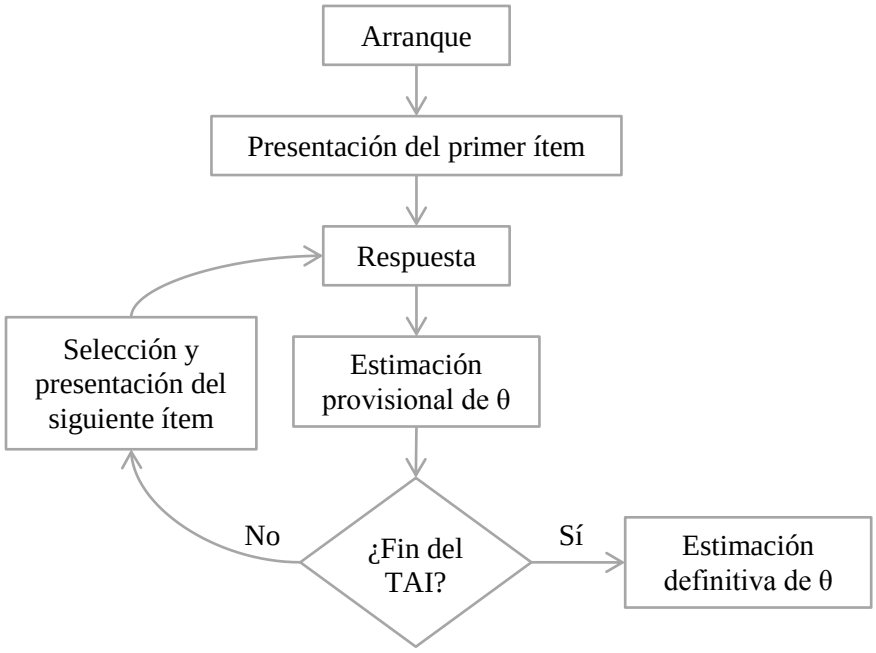

Figura 1. Algoritmo de administración del TAI. Tomada de «GenTAI: Generador de test adaptativos informatizados» por J. López y J. M. Sánchez, 2005, *Revista Iberoamericana de Informática Educativa, 2*, p. 11.

Según Olea y Ponsoda (2004) para desarrollar un TAI se necesitan, como elementos básicos, un banco de ítems con parámetros estimados a partir de un modelo IRT, procedimientos para iniciar y finalizar la prueba y para escoger los ítems que se le presentarán al evaluado, y un método estadístico de estimación del atributo. Para la construcción del primer elemento —banco de ítems— se deben tener en cuenta varias consideraciones, algunas aplicables al diseño de bancos para cualquier tipo de prueba, otras aplicables específicamente al diseño de bancos para TAI y otras que tienen en cuenta las particularidades de los constructos que se pretendan evaluar.

De acuerdo con Herrera (1996) y Tristán y Vidal (2006), si bien las etapas para construir un instrumento de medida pueden variar dependiendo de los intereses de los autores o del tipo de instrumento que se desee construir, existen algunos pasos generales que se deben tener en cuenta a la hora de llevar a cabo esta tarea. Estos pasos incluyen: la conformación de un grupo de expertos en el área que se desea evaluar y en psicometría; la planeación de la prueba, es decir, la definición de aquello que se va a medir, el objetivo de dicha medición, la población a la que irá dirigida y el tipo de ítems que se utilizarán, entre otros; la elaboración del marco teórico que servirá de sustento; la conformación de la estructura de prueba, es decir, la distribución de los ítems en los temas que se evaluarán; la definición de las especificaciones psicométricas de la prueba, tales como la forma en la que se probará el instrumento, los criterios de selección de ítems que se usarán y la forma en la que se presentarán los resultados; y, finalmente, la elaboración de material complementario.

En relación con los aspectos particulares que se deben tener en cuenta para la conformación de bancos para TAI, Olea y Ponsoda (2004) señalan la necesidad de asegurarse de: a) contar con una suficiente cantidad de ítems, la cual estará determinada por las restricciones impuestas por el algoritmo de elección de ítems, el número de aplicaciones del TAI que se piensa realizar y la posible tasa de exposición de los ítems, entre otros aspectos; b) realizar la calibración de los ítems cumpliendo con el tamaño mínimo de muestra requerido de acuerdo con la cantidad de bloques en los que esté dividido el banco y el modelo de interpretación escogido, ya sea de uno o más parámetros; c) comprobar qué tanto se está cumpliendo con el supuesto de unidimensionalidad, lo cual debe hacerse cada vez que se introduzcan nuevos ítems en el banco o se apliquen los ítems existentes a

muestras diferentes a la inicial; d) verificar el grado de invarianza de los parámetros estimados para el banco y el grado en el que el modelo usado predice los datos observados; y e) renovar o ampliar el banco una vez se detecten ítems con propiedades insatisfactorias o que tengan una alta tasa de exposición.

Adicionalmente, para estos autores es muy importante que los ítems que sean seleccionados para conformar el banco de un TAI sean aquellos que brinden más información sobre el nivel de atributo de los evaluados y que, además, abarquen todo el continuo del atributo que se desea estimar; sin embargo, este último requisito puede variar dependiendo de los objetivos de la evaluación puesto que, por ejemplo, si se desea hacer una diferenciación entre individuos con un alto nivel de rasgo, no es necesario incluir en el banco del TAI ítems con una dificultad muy baja.

Finalmente, respecto a las consideraciones derivadas del constructo que se pretende evaluar, algunos autores se han puesto en la tarea de identificar cuáles son las características que pueden hacer que un ítem resulte más o menos difícil que otro dependiendo del constructo que se quiera evaluar, lo cual, resulta ser muy importante para ayudar a garantizar que los ítems que conforman un TAI se encuentran distribuidos a lo largo del continuo. Por ejemplo, en el caso de la evaluación de la comprensión de lectura, Embretson y Wetzel (1987) encontraron que si se utiliza un lenguaje complejo solo en la opción de respuesta correcta, la dificultad del ítem aumenta, mientras que si este vocabulario se usa solo en los distractores, la dificultad del ítem disminuye; Gorin (2005), por su parte, encontró que el amplio uso de una redacción negativa aumenta en gran medida la dificultad de un ítem; y Brizuela y Montero (2013) encontraron que cuantos menos conectores tiene un texto, más difíciles resultan los ítems que se desprenden de él, que los ítems en los que se hace uso de cláusulas subordinadas son más difíciles ya que le demandan al evaluado una mayor cantidad de memoria de trabajo para tener presente el sujeto, y que si en el texto y en las opciones de respuesta de los ítems se hace uso de un lenguaje similar, la dificultad del ítem disminuye, mientras que si existen grandes diferencias entre el lenguaje utilizado en el texto y las opciones de respuesta, la dificultad del ítem aumenta.

Tener en cuenta todas estas consideraciones, por un lado, optimiza la elección de los ítems que conformarán el banco puesto que cuantos más controles se lleven a cabo,

menos probable es que deban desecharse ítems por no cumplir con las características definidas en la planeación de la prueba y, por el otro, ayuda a garantizar que los ítems elegidos para conformar el banco cuenten con propiedades psicométricas deseables que se adapten a los requerimientos de la evaluación, convirtiendo el proceso de construcción en sí mismo en una fuente de evidencia de la validez de las interpretaciones que se realicen a partir de las puntuaciones obtenidas mediante la aplicación de dicho banco.

En relación con el segundo elemento básico para conformar un TAI —procedimientos para iniciar y finalizar la prueba y para escoger los ítems que se le presentarán al evaluado—, Revuelta, Ponsoda y Olea (1998) afirman que si bien la elección de los ítems que se le presentarán al individuo a lo largo del TAI se basa en la cantidad de información que los ítems ofrecen acerca del nivel de atributo del individuo, es decir, mediante el Método de Máxima Información (MMI), existen criterios adicionales para esta elección que varían dependiendo de si se trata del ítem inicial, de los ítem intermedios o del ítem final.

La elección del ítem inicial puede darse en función de tres criterios: el nivel de atributo del individuo —conocido mediante información previa—, un nivel de atributo asignado al individuo de manera arbitraria o intuitiva, o un nivel de dificultad determinado (Revuelta, Ponsoda y Olea, 1998). Los ítems intermedios pueden ser elegidos mediante el método de estratificación de dos niveles, que inicia con la presentación de un conjunto de ítems con diferentes niveles de dificultad, continúa con una estimación preliminar del atributo del evaluado a partir de sus respuestas y culmina con la elección de los ítems que se presentarán posteriormente; o mediante el método de estratificación de múltiples niveles, en el que la respuesta del individuo a cada ítem presentado determinará el ítem subsiguiente. Finalmente, el ítem de terminación puede estar definido por haber aplicado cierto número de ítems, haber logrado cierta precisión de la medida o haber cumplido con un criterio que permita ubicar al individuo en un grupo de acuerdo con una clasificación previamente definida (Muñiz, 1997).

De acuerdo con Revuelta, Ponsoda y Olea (1998), además de tener en cuenta estos criterios para la elección de los ítems que se le presentarán al individuo, es necesario definir mecanismos que controlen las tasas de exposición de los ítems ya que, como se mencionó, su elección depende de la cantidad de información que brinden sobre el nivel de atributo de

los individuos, lo que hace que algunos de ellos tiendan a ser presentados con mayor frecuencia que otros y, por lo tanto, puedan llegar a ser conocidos por los individuos antes de que les sea aplicada la prueba. Las estrategias más comunes para evitar la sobreexposición de los ítems, además de la renovación constante de los bancos, incluyen los métodos indirectos y los métodos directos. Los primeros adicionan un componente aleatorio al algoritmo basado en MMI, seleccionando aleatoriamente entre los ítems más informativos los que se van a presentar; y los segundos, además de contener un componente aleatorio como criterio de elección, agregan parámetros que fijan las tasas de exposición de los ítems de forma explícita.

Finalmente, para Abad, Olea, Real y Ponsoda (2002) los métodos estadísticos más utilizados para estimar el nivel de atributo de los evaluados —tercer elemento básico para conformar un TAI— son la estimación de Máxima Verosimilitud (ML, sigla del ingl. Maximum Likelihood) que se basa en datos empíricos, y los métodos bayesianos que tienen en cuenta información sobre la distribución a priori del nivel de atributo de los evaluados, y que incluyen la estimación máxima a posteriori (MAP) que asume que el estimador del atributo es la moda, y la estimación esperada a posteriori (EAP) que asume que el estimador del atributo es la media. De acuerdo con estos autores, la eficacia de cualquiera de estos métodos de estimación depende «del sesgo y del error típico de medida que generen para diferentes niveles de habilidad y para distintas condiciones de aplicación» (p. 2).

Según Wise y Kingsbury (2000), a pesar de que los principios básicos para desarrollar un TAI son relativamente sencillos, la implementación y el mantenimiento de ese tipo de instrumentos resultan complejos debido a la gran cantidad de variables que intervienen en su desarrollo, lo cual les impone a los evaluadores una serie de desafíos prácticos. Algunos de estos desafíos incluyen la definición de estrategias de control de las tasas de exposición, ya mencionada, el alto costo del mantenimiento del banco de ítems puesto que es necesario actualizarlo con frecuencia y asegurar que cuente con cierto número mínimo de ítems (Olea y Ponsoda, 2004), el alto costo de los equipos y de los software necesarios para realizar las calibraciones y demás estimaciones requeridas (Muñiz, 1997), el manejo de los niveles de ansiedad experimentados por los evaluados ya que pueden sentir que no tienen control

sobre el proceso evaluativo debido a que no pueden modificar sus respuestas (Fritts y Marszalek, 2010), entre otros.

Superados estos desafíos, los TAI pueden representar grandes ventajas sobre las pruebas de lápiz y papel tradicionales ya que, como se ha mencionado, esta estrategia de evaluación favorece la obtención de medidas más precisas (Muñiz y Hambleton, 1999), permite diferenciar con mayor claridad entre las personas que presentan niveles de atributo extremos (Brown y Weiss, 1977), favorece la inclusión de formatos de ítems novedosos, posibilita el control de tiempos de exposición de los ítems (Arribas, 2004), permite la calificación automática de la prueba (Huff y Sireci, 2001), y reduce la cantidad de ítems que se le presentan al evaluado (Embretson, 1992), lo cual implica un menor tiempo de aplicación (Weiss y Betz, 1973) y menores niveles de fatiga (Shermis y Lombard, 1998), entre otros.

Adicionalmente, entidades como el Educational Testing Service (ETS) consideran a los TAI como una estrategia de evaluación muy prometedora en el campo de la educación, ya que puede ayudar a favorecer la inclusión de poblaciones con algún tipo de discapacidad al contar con algunas de las ventajas mencionadas como la facilidad en la elección del sitio de aplicación, los novedosos formatos de ítems y la posibilidad de interactuar con la prueba, retrocediendo y avanzando a lo largo de cada ítem, entre otras (Stone y Davey, 2011); no obstante, es claro que falta la realización de estudios que den solidez a estos planteamientos y que brinden una guía para manejar adecuadamente los retos que estos instrumentos entablan al concepto de validez, puesto que la mayoría de estudios acerca de los TAI tienen como foco de atención los bancos de ítems (Cella, Gershon, Lai y Choi, 2007; Molina, Pareja y Sanmartín, 2008; Abad, Olea, Aguado, Ponsoda y Barrada, 2010; Lozzia et ál., 2015; Şahina, A. y Weiss, D. J., 2015), los métodos de elección de ítems (Wang y Chang, 2011; Yao, 2012) y algunos desarrollos prácticos (Olea, Abad, Ponsoda y Ximénez, 2004; Rubio y Santacreu, 2004).

Validez

El concepto de validez en el campo de la evaluación y la medición psicológica ha sido objeto de múltiples debates y modificaciones a lo largo de la historia debido, principalmente,

a los constantes cambios en la forma en la que son concebidas las pruebas objetivas, la inclusión de nuevas tecnologías y el aumento del interés por garantizar equidad en las evaluaciones, en particular, cuando se cuenta con poblaciones con características disímiles. La AERA, la APA y el NCME, en un intento por dar cuenta de estos debates, en cada una de las versiones de sus estándares han ofrecido definiciones de validez que corresponden con aquellas que en el momento de su publicación han sido las más aceptadas.

En 1966, estas entidades definieron validez como el grado en el que la prueba o instrumento mide lo que pretende medir, y la dividieron en tres formas —validez de constructo, de contenido y de criterio— cada una de las cuales contaba con procedimientos y estrategias específicas de estimación, desarrollados a partir de los postulados de la TCT. De acuerdo con Gómez e Hidalgo (2005) esta noción de validez representaba varios riesgos para los constructores de pruebas al asumir que estas formas de validez eran independientes y que se debían recopilar evidencias de las tres formas en cualquier estudio que se realizara de validez, lo cual favoreció la aparición de cuestionamientos en torno a esta concepción tripartita y tomó fuerza la idea de desarrollar un concepto que se centrara en la teoría subyacente al atributo por estimar y que integrara, en una sola, las formas de validez mencionadas. Esta nueva visión unitaria se condensó en los estándares de la AERA, la APA y el NCME de 1974 y 1985.

Para Pérez, Chacón y Moreno (2000) la raíz de la concepción de validez de constructo como concepto unificador se encuentra en el trabajo de Cronbach y Meehl (1955) en el que se definió validez 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). Esta definición trasladó el enfoque del «desarrollo de un test para una interpretación dada, a la relación entre el test y una interpretación propuesta» (Kane, 2013; p. 5) y dio lugar a multiplicidad de planteamientos posteriores como los de Loevinger (1957), Guion (1977) y Messick (1980; 1988; 1989; 1995).

En coherencia con el espíritu de esta nueva concepción de validez, Cronbach (1971) asumió la validación como un proceso interminable que se enfoca en la «precisión de una predicción específica o inferencia realizada a partir de las puntuaciones de una prueba» (p. 443), y, en 1988, propuso cinco perspectivas a partir de las cuales el validador podría

formular las preguntas relacionadas con el uso de las pruebas y las inferencias hechas de sus puntuaciones: la *funcionalista*, que valora las consecuencias reales o posibles del uso de la prueba para individuos e instituciones; la *política*, que se enfoca en la utilización de los resultados de las pruebas con el fin de desarrollar leyes y reglamentaciones; la *operacionalista*, que busca determinar si los elementos que componen la prueba contemplan aspectos relevantes del atributo que se está midiendo; la *económica*, que se centra en los costos que representan la construcción de las pruebas y en las posibles decisiones tomadas a partir de sus resultados; y, por último, la *explicativa*, que pretende dar cuenta de la información que se requiere para la construcción de teorías en torno a los procesos psicológicos.

Posteriormente, Messick —el defensor más representativo de la visión unificada de validez— definió validez como «un juicio evaluativo global de la adecuación y la conveniencia de las inferencias derivadas de las puntuaciones de las pruebas» (Messick, 1980; p. 1023) y propuso que el propósito de la validación del uso de las pruebas, así como de las interpretaciones de sus puntuaciones, fuera visto en función de una matriz interconectada que contenía *bases evidenciales*, es decir, aproximaciones prácticas de las relaciones con otros constructos que contemplan, por una parte, la recopilación de evidencias que soporten el significado o interpretación plausible de las puntaciones de la prueba, y, por otra, la relevancia del constructo y la utilidad de sus puntuaciones en aplicaciones específicas; y *bases consecuenciales*, es decir, «evaluaciones de juicios de valor contextual de las relaciones implicadas en diversos atributos y conductas» (Wainer y Braun, 1988, p. 2) que abarcan la estimación del valor de las implicaciones de las interpretaciones de las puntuaciones de la prueba, y las consecuencias sociales reales y potenciales del uso propuesto (Messick, 1989).

Luego de la publicación de los estándares de la AERA et ál. de 1985, este autor, argumentando que la definición de validez dada en esta versión de los estándares daba cabida a diversas interpretaciones frente al proceso de validación al no brindar parámetros prácticos que orientaran la labor de los investigadores (Messick, 1988), propuso una nueva conceptualización de validez que hacía énfasis en los usos e interpretaciones de las puntuaciones, concebía la validación como un proceso de recolección de evidencia e

intentaba disipar del todo la clásica consideración de que la validez es una propiedad del instrumento de medida *per se* (Messick, 1989).

Estos planteamientos, junto con los de Cronbach (1988), sirvieron de insumo para que Kane (1992) también intentara disipar la ambigüedad presente en los estándares de 1985 asociada con el proceso de validación, proponiendo un mecanismo «basado en argumento» para validar las inferencias y los usos en el que el validador debía evaluar la evidencia disponible por medio de la formulación de hipótesis y el desarrollo de argumentos que defendieran el uso o la inferencia propuesta; lo cual, para Sireci (2009) representó la posibilidad de establecer de forma más explícita qué tipos de evidencia eran más adecuados para soportar determinados usos o inferencias, tal como se reflejó en los estándares de la AERA et ál. de 1999 y 2014 en los que se definió validez como «el grado en el cual la evidencia y la teoría soportan las interpretaciones de las puntuaciones de una prueba asociadas con los usos propuestos» (2014; p. 11), y se propusieron cinco fuentes de evidencia asociadas con diferentes aspectos de las pruebas: contenido, procesos de respuesta, estructura interna, relación con otras variables y consecuencias.

La fuente de evidencia referida al *contenido* de la prueba abarca los análisis tradicionales de la validez de contenido, es decir, la valoración de si los ítems que componen la prueba son una muestra representativa del atributo que se pretende medir e incluye estrategias de recopilación como el juicio de expertos. La segunda fuente, enfocada en los *procesos de respuesta*, se refiere al grado de concordancia existente entre el constructo y la naturaleza de las respuestas dadas por los individuos e incluye, como estrategias de recopilación, entrevistas cognitivas, observaciones conductuales de las personas mientras responden las pruebas, entre otras. La tercera fuente de evidencia, que se centra en la *estructura interna* de la prueba, se refiere a la definición de las dimensiones evaluadas por la prueba y abarca análisis estadísticos de los ítems, escalamiento multidimensional y análisis factoriales. La cuarta fuente se enfoca en las *relaciones con otras variables* y se refiere a la valoración de la validez de criterio tradicional, involucrando estudios multirasgo-multimétodo además de estudios correlacionales. Por último, la quinta fuente se interesa en las *consecuencias* de las pruebas, se refiere a la evaluación de sus

consecuencias esperadas e inesperadas, e incluye estudios de impacto adverso, entre otros (AERA et ál., 1999).

Según Kane (2001), de esta definición deben rescatarse cuatro aspectos principales: (a) la validez implica una evaluación integral de las interpretaciones, no solo una simple aplicación de técnicas; (b) el juicio resultante de la validación refleja la adecuación y conveniencia de las interpretaciones, así como el grado en el que la evidencia las soporta; (c) la validez involucra la valoración de la plausibilidad general de un uso o inferencia propuesta a partir de las puntuaciones de una prueba; y (d) la validación puede abarcar la evaluación de las consecuencias sociales del uso de las pruebas. Lo anterior, para este autor evidencia que el problema de la validez no reside en su definición teórica, sino en la forma en la que se enlaza dicha teoría con la práctica (Kane, 2009).

De manera complementaria, Sireci (2009) afirma que si bien las fuentes de evidencia postuladas en esta definición brindan «una estructura útil para evaluar el uso de una prueba para un propósito particular y para documentar la evidencia de validez de forma coherente» (p. 31), junto con el planteamiento de los distintos mecanismos para valorarlas, aún hace falta una ejemplificación práctica; y que, además, al no existir un ente o autoridad que garantice que una prueba es válida para cierto uso o inferencia, ni que determine cuándo se puede afirmar que se cuenta con suficiente evidencia para soportarlos, la tarea del validador aún continúa siendo confusa.

Otro de los aspectos de esta definición de 1999 que generó grandes debates fue la inclusión de las consecuencias sociales como fuente de evidencia de validez. Para algunos autores, que venían hablando del tema antes de la publicación de esta versión de los estándares, esta inclusión resulta apropiada dado que al hablar de adecuación, significado y utilidad están implicados juicios de valor (Linn, 1997); mientras que, para otros, si bien es importante tener en consideración las consecuencias sociales del uso de las pruebas, no se puede asumir esta indagación como otra faceta de validez puesto que esto complica de manera innecesaria este concepto (Popham, 1997) y propicia que se presenten confusiones entre la precisión de la inferencia acerca del nivel de atributo en el individuo y la utilización de los datos en un proceso de toma de decisiones (Mehrens, 1997).

Algunas de estas críticas a la definición de validez de la AERA et ál. de 1999 lograron ser «parcialmente superadas en los 25 criterios de validez descritos en la última versión de los estándares» (Herrera, Barajas y Jiménez, 2015; p. 301), los cuales se encuentran agrupados alrededor de los usos y las interpretaciones de las pruebas, las muestras y el contexto de validación, y las formas específicas de evidencia de validez. Este nuevo acercamiento a la conceptualización de validez y validación favoreció que varios expertos se interesaran en proveer algunos ejemplos de cómo recolectar evidencias de cada una de las fuentes descritas en estos estándares; ejemplos que, de acuerdo con Sireci y Padilla (2014), representan para los profesionales en evaluación psicológica una mayor comprensión sobre cada una de estas fuentes de evidencia.

Las estrategias de recolección de evidencias de validez propuestas por estos expertos incluyen, para las evidencias basadas en la estructura interna de la prueba, la evaluación de aspectos relacionados con la dimensionalidad, la invarianza de la medida y la confiabilidad, mediante la realización de análisis factoriales y el uso de modelos bifactor (Ríos y Wells, 2014); para las evidencias basadas en los procesos de respuesta, el uso de entrevistas cognitivas y de métodos cuantitativos y cualitativos para su análisis (Padilla y Benítez, 2014); para las evidencias basadas en las relaciones con otras variables, la realización de los clásicos estudios de validez predictiva (Oren, Kennet-Cohen, Turvall y Allalouf, 2014); para las evidencias de validez basadas en las consecuencias, la elaboración de un argumento de validez enfocado en las consecuencias deseadas y no deseadas de las pruebas; y para las evidencias basadas en el contenido de la prueba, la conformación de paneles de expertos que evalúen la forma en la que está definido operacionalmente el constructo que se va a medir, el grado en el que la prueba diseñada representa el constructo, la relevancia de cada ítem para evaluar el constructo y la idoneidad de los procesos y controles llevados a cabo en la construcción de la prueba (Sireci y Faulkner-Bond, 2014).

Si bien para algunos autores, como Shepard (2016) y Sireci (2016), esta aproximación al concepto de validez de la AERA et al. (2014) representa una visión abarcadora de la validez y brinda un marco de validación útil, aún no se encuentra exenta de críticas. Lo anterior, sumado a los planteamientos presentados a lo largo de este apartado, pone de manifiesto, por un lado, que la validez es un concepto de gran complejidad que requiere

continuar siendo perfeccionado y, por el otro, que se deben continuar realizando estudios de validación que sean aplicables a distintos tipos de instrumentos de medida, ya que dependiendo de las condiciones que cada uno de estos le impongan a la evaluación, las estrategias que se deben utilizar para recolectar evidencias de las distintas fuentes de validez pueden variar, lo cual se muestra en el siguiente apartado con la incorporación de TAI en la evaluación psicológica.

Validez en TAI

De acuerdo con Huff y Sireci (2001), a pesar de las múltiples ventajas que los TAI representan para la evaluación psicológica frente a las pruebas de lápiz y papel convencionales, son pocos los estudios que han intentado determinar las implicaciones de este tipo de instrumentos en términos de validez. No obstante, antes de hablar de ello, es necesario abordar el tema de la confiabilidad, entendida como precisión de la medida, puesto que, como señalan autores como Aiken (2013), aunque insuficiente, esta es una condición necesaria para que las interpretaciones que se realicen a partir de las puntuaciones de los instrumentos resulten válidas.

Para Muñiz (2010), una de las ventajas de la IRT (modelo que fundamenta el desarrollo de los TAI) en la evaluación psicológica es la introducción de la Función de Información como mecanismo que permite determinar el nivel de precisión con el que está siendo evaluado cada nivel de atributo, de modo que, según Baker (2001), si los ítems de un instrumento proveen una gran cantidad de información sobre un determinado nivel de atributo, los evaluados que tengan dicho nivel serán evaluados con una alta precisión. De acuerdo con Olea y Ponsoda (2004), esta información será mayor cuanto: a) mayor sea la capacidad de los ítems que componen el instrumento para diferenciar entre los que poseen y no poseen el atributo que se está evaluando, b) menor sea la probabilidad de que las personas que no poseen el atributo acierten los ítems del instrumento, c) mayor sea la cantidad de ítems que componen el instrumento, y d) mayor sea la convergencia entre el nivel de atributo y la dificultad de los ítems.

En el caso de los TAI, si bien el contar con una pequeña cantidad de ítems podría afectar el nivel de precisión de las medidas que brinda este tipo de instrumentos, el hecho

de que les presenten a los evaluados los ítems que resultan más apropiados para su nivel de atributo hace que estos instrumentos logren alcanzar altos niveles de precisión, sobrepasando las posibles desventajas que podría tener el contar con pocos ítems (Olea y Ponsoda, 2004). Lo anterior resalta la importancia de desarrollar bancos de ítems que cumplan con las recomendaciones presentadas en apartados anteriores como las postuladas por Herrera (1996) y Olea y Ponsoda (2004).

En relación con la validez en los TAI, Huff y Sireci (2001) señalan que los aspectos que los hacen potencialmente ventajosos en términos de validez frente a las pruebas de lápiz y papel convencionales son los mismos que los hacen potencialmente riesgosos, lo cual representa grandes retos para los diseñadores de este tipo de instrumentos. A continuación, se expondrán estos aspectos clasificándolos de acuerdo con las fuentes de evidencia de validez propuestas por la AERA et ál. (2014) —contenido, estructura interna, procesos de respuesta, relación con otras variables y consecuencias—.

En cuanto a las evidencias de validez basadas en el contenido de la prueba, se ha encontrado que los TAI, al posibilitar el uso de nuevos formatos de ítems dado su carácter informático, favorece una mayor eficiencia en la evaluación de habilidades cognoscitivas de alto nivel y la obtención de medidas más amplias de dominios de constructo (Huff y Sireci, 2001); sin embargo, al valorar un atributo que presenta varios dominios de contenido, se puede poner en riesgo la validez del instrumento si no se cuenta con mecanismos de selección de ítems que garanticen que todos los contenidos se están evaluando con la misma precisión, lo cual, a su vez puede representar un problema puesto que al imponer estos mecanismos se corre el riesgo de aumentar la longitud de la prueba y, por ende, de reducir su eficiencia (Weiss, 2004).

Adicionalmente, Huff y Sireci (2001) afirman que debido a que cuando se calibran los ítems se suelen anteponer los criterios estadísticos frente a los cualitativos, los ítems que son relevantes para el constructo que se desea estimar y no se adaptan al modelo de interpretación utilizado suelen ser rechazados de inmediato; y si los ítems de un dominio de contenido particular se ajustan menos que los de otros dominios, es probable que el atributo por estimar se vea significativamente alterado, lo que puede favorecer la subrepresentación del constructo. Lo anterior podría ser solventado mediante la

aplicación de la estrategia propuesta por Sireci y Faulkner-Bond (2014) para recolectar evidencias de esta fuente de validez que consiste en la conformación de paneles de expertos que, para este caso, tendrían que evaluar si el banco diseñado para el TAI representa el constructo que se pretende medir.

Respecto a las evidencias basadas en los procesos de respuesta, se ha encontrado que los TAI favorecen la reducción de la fatiga al aplicar menos ítems, lo cual descarta esta fuente de varianza irrelevante para el constructo que se va a estimar. Además, posibilita el registro de variables como el tiempo de respuesta de cada evaluado a cada ítem, permitiendo evaluar la posible relación entre estas y el constructo de interés (Hornke, 2000) y favoreciendo la incorporación de modelos cognitivos que permitan dar cuenta de la respuesta de los evaluados, lo cual resulta especialmente importante en la evaluación de ciertos grupos poblacionales como el de personas con LV sobre el que se podría suponer que sus procesos de respuesta difieren de los procesos del grupo de personas sin LV.

Por otra parte, para Lee, Moreno y Sympson (1984), los TAI también pueden favorecer que los evaluados experimenten niveles de ansiedad más altos que los que pueden llegar a experimentar cuando se enfrentan a pruebas de lápiz y papel convencionales, lo cual, de acuerdo con Fritts y Marszalek (2010), puede deberse al descubrimiento del evaluado de que el algoritmo subyacente a los TAI se basa en los niveles de dificultad y al hecho de que no se le permita revisar sus respuestas a cada uno de los ítems. En relación con este último aspecto, Papanastasiou y Reckase (2007) señalan que para algunos teóricos el permitirle a los evaluados revisar sus respuestas puede favorecer la validez de las interpretaciones que se realicen a partir de las puntuaciones de los evaluados ya que les permite repensar sus respuestas y hacer correcciones; mientras que para otros esto puede representar problemas al aumentar la probabilidad de que los evaluados hagan trampa al resolver la prueba.

Las implicaciones que estas variables tienen sobre los procesos que llevan a cabo los evaluados para responder los ítems pueden ser abordados mediante el uso de entrevistas cognitivas, tal como lo proponen Padilla y Benítez (2014), o mediante el registro de variables de interés y la comparación del efecto que estas variables sobre la evaluación para diferentes grupos, lo cual a su vez podría favorecer el desarrollo de mecanismos que permitan controlar los efectos de estas variables.

En lo referente a las evidencias de validez basadas en la estructura interna de la prueba, Green (1988) afirma que los TAI pueden introducir varianza irrelevante para el constructo que se desea abordar, debido al medio por el que son aplicados; factores como la familiaridad con las herramientas informáticas, la ansiedad, los límites de tiempo establecidos para la presentación de cada uno de los ítems, la imposibilidad para el individuo de revisar sus respuestas, la plataforma utilizada, el tamaño de la pantalla y la ubicación de los elementos que componen el ordenador, pueden ser variables que interfieren con una adecuada medida del atributo de interés. Por ejemplo, Taylor, Kirsch, Eignor y Jamieson (1999) tras realizar una revisión acerca de si la familiaridad que un individuo tiene con el manejo de herramientas informáticas puede llegar a favorecer o entorpecer su ejecución respecto del contenido de los TAI, encontraron resultados contradictorios, siendo que para algunas áreas de evaluación estos efectos son más evidentes cuando los formatos de los ítems son mucho más complejos que los usualmente utilizados en pruebas de lápiz y papel convencionales.

En cuanto a las evidencias de validez basadas en las relaciones con otras variables, Olea y Ponsoda (2004) mencionan algunos estudios en los que se han realizado correlaciones entre las puntuaciones de los TAI y algunas medidas de interés como el desempeño laboral, sin encontrar diferencias contundentes entre las pruebas de lápiz y papel y los TAI. De igual manera, otros autores han señalado la importancia de que esta fuente de validez también sea abordada mediante la recolección de información que podría llegar a explicar el desempeño de los evaluados en función de ciertas variables de interés.

Finalmente, en relación con las evidencias de validez basadas en las consecuencias de las pruebas, Huff y Sireci (2001) destacan que los TAI permiten la calificación automática tras haber respondido la prueba, lo que favorece la validez al eliminar la dependencia de los evaluadores; sin embargo, Clauser, Harik, y Clyman (2000) señalan que a pesar de que los algoritmos desarrollados para generar puntuaciones automatizadas tienen como fin garantizar uniformidad y precisión en la estimación del atributo abordado, estos pueden presentar un funcionamiento inadecuado y variar en función del grupo de expertos que los diseñen, lo cual favorece la posibilidad de cometer errores en las puntuaciones y, por lo tanto, en las inferencias realizadas a partir de ellas.

Adicionalmente, las puntuaciones obtenidas también pueden llegar a presentar sesgo por no tener en cuenta que con el paso del tiempo los parámetros de los ítems tienden a cambiar dados factores de exposición y enseñanza, lo que evidencia la necesidad de concebir la validación como un proceso constante y no dar por sentado que los criterios que en algún momento sirvieron como base para afirmar que una prueba era válida para tal uso o inferencia servirán para siempre (Huff y Sireci, 2001), para lo cual pueden tenerse en cuenta las recomendaciones de Olea y Ponsoda (2004) relacionadas con el mantenimiento de los bancos de ítems de los TAI.

En cuanto a las consecuencias sociales imprevistas más representativas de los TAI se encuentran: el alto costo del desarrollo de este tipo de herramientas, lo que supone una desventaja para la población que presenta una condición socioeconómica desfavorable dado que no tienen acceso en la misma proporción a la tecnología que las demás personas; el cambio en la forma en la que las personas se preparan para enfrentarse a situaciones en las que serán valoradas por medio de TAI, la cual puede entorpecer su rendimiento al percibir que deben prepararse menos puesto que se les presentará un menor número de ítems; y el efecto de las estrategias de preparación para enfrentarse a un TAI promulgadas por las compañías, las cuales en lugar de intentar mejorar el rendimiento de los individuos respecto al constructo por medir, procuran manipular el algoritmo subyacente al TAI para obtener mejores puntuaciones (Huff y Sireci, 2001).

Como se puede observar, debido al carácter informático de los TAI, para la recolección de evidencias de validez de cada una de estas fuentes, los TAI suponen ventajas sobre las pruebas tradicionales de lápiz y papel ya que permiten el registro de una gran cantidad de variables, lo cual, puede ser aprovechado por los investigadores interesados en usar este tipo de mecanismos para la evaluación de diferentes atributos psicológicos o de grupos poblacionales con diferentes características.

Método

Esta investigación se llevó a cabo en dos fases. En la primera se diseñaron los instrumentos que permitieron la recolección de evidencias de validez del TAI que sirvió de insumo para esta investigación, el cual contaba con dos plataformas, una para el acceso a la información por vía visual y otra para el acceso a la información por vía auditiva. Esta fase involucró el acompañamiento en el desarrollo de este aplicativo, la construcción, validación y calibración de su banco de ítems, y la construcción de los instrumentos mencionados.

En la segunda fase, se aplicaron el TAI y los demás instrumentos, y se analizó la información recolectada. Esta fase abarcó el acompañamiento en la selección y el contacto con los individuos que respondieron el TAI, el registro de variables de interés asociadas con validez, y la depuración de la información resultante con su respectivo análisis.

Fase 1

Participantes.

Para la aplicación piloto del banco que conformó el TAI se contactaron estudiantes de grados 10.º y 11.º vinculados a instituciones educativas públicas y privadas de los municipios de Bogotá, Medellín, Puerto Tejada, San José del Palmar y Yopal. Estos estudiantes fueron convocados a través de los rectores de los colegios y se les solicitó su autorización y la de sus padres para participar en el estudio; en el primer caso, se hizo uso de un asentimiento informado y, en el segundo, de un consentimiento informado. Como resultado, se contó con la participación de 1325 estudiantes. La tabla 1 presenta su distribución por municipio.

Tabla 1

Cantidad y porcentaje de estudiantes que participaron en la aplicación piloto del banco de ítems del TAI, diferenciando por municipio

Municipio	Estudiantes	
	Cantidad	Porcentaje
Medellín	698	53 %
Bogotá	545	41 %
Yopal	36	3 %
Puerto Tejada	31	2 %
San José del Palmar	15	1 %
Total	1325	100 %

Instrumentos.

• *Protocolos de construcción y validación de ítems:* Para llevar a cabo los talleres de construcción y validación de los ítems se hizo uso de protocolos en los que se detallaron los criterios metodológicos necesarios para su realización, el personal requerido y los aspectos logísticos que debían ser tenidos en cuenta durante su desarrollo.

• *Formato de construcción de ítems:* En este formato se registraron los ítems construidos por los expertos y se incluyeron las observaciones de los validadores a cada uno de ellos.

• *Acta de validación de ítems:* En esta acta se registraron las modificaciones sugeridas por los validadores a cada uno de los ítems del banco diseñado.

• *Formatos de asentimiento y consentimiento informado.* Estos formatos se utilizaron para informar a los estudiantes y a sus padres acerca de las características del estudio y para que autorizaran su participación en la aplicación del TAI.

• *Manual de aplicación de pruebas:* Este manual se diseñó para guiar las actividades que debían desarrollar las personas encargadas de la aplicación piloto de los ítems del TAI.

• *Paquetes de análisis de datos:* Para analizar la información recolectada tras la aplicación piloto de los ítems del TAI, se hizo uso de los programas SPSS v.22, y Winsteps v.3.73.

Procedimiento.

Dado que el objetivo de esta fase era el diseño del TAI y de los demás instrumentos que sirvieran para la recolección de evidencias de validez, se planearon cuatro grandes actividades: participar en el diseño del TAI, participar en la conformación del banco de ítems, participar en la calibración de banco y diseñar los instrumentos de recolección de las variables de interés.

La primera actividad, correspondiente al *diseño del TAI*, fue realizada por los miembros del macroproyecto del que se deriva este estudio en compañía de un ingeniero de sistemas e implicó la definición de los criterios de selección de ítems y de estimación del atributo del TAI, con base en el objetivo de la evaluación y las características de la población evaluada; el acuerdo sobre las condiciones de accesibilidad para la población invidente, con base en la normatividad vigente; y la definición de los mecanismos para la recolección de variables

asociadas con validez, que incluyeron variables generales como sexo, edad, estrato socioeconómico, lugar de procedencia, sector de la institución educativa en la que los participantes cursaron la secundaria y tipo de LV; y de variables propias del TAI como tiempo que los individuos tardaban en abordar cada ítem y número de veces que reproducían cada fragmento de texto, incluyendo las opciones de respuesta.

La segunda actividad, *conformación del banco ítems*, también fue realizada en compañía de los miembros del macroproyecto que engloba este estudio. Estos ítems eran de selección múltiple con única respuesta y las fuentes de las que provinieron fueron: banco del Laboratorio de Psicometría, banco diseñado como parte de las tesis de maestría de Soler (2013) y Espinosa (2013), y banco construido dentro del macroproyecto. La consolidación de este banco implicó realizar las siguientes actividades:

1. *Selección de los ítems del banco del Laboratorio de Psicometría.* Los ítems de lenguaje que hacían parte del banco del Laboratorio fueron revisados, se seleccionaron aquellos que cumplían con los requisitos establecidos en el *Manual de construcción de preguntas para evaluar comprensión de lectura en personas con y sin LV*, desarrollado por Espinosa (2013) y, posteriormente, fueron clasificados de acuerdo con la estructura propuesta por el Icfes para la subprueba de Lenguaje de la prueba Saber 11.º; es decir, que aunque se acogió la estructura de prueba propuesta por el Icfes, los ítems fueron adaptados para cumplir con los requisitos establecidos por Espinosa (2013).

Esta adaptación incluyó, entre otros aspectos, la reducción de los textos que eran muy largos y cuya longitud era innecesaria para abordar las competencias que se pretendían evaluar, y el reemplazo en los enunciados de expresiones como «*En el texto, la palabra subrayada se refiere a*» por «*En la frase "[...]", la palabra "[...]" se refiere a*», ya que la expresión inicial le implicaba a los evaluados volver al texto a verificar de qué palabra se trataba y esto representaba barreras para la adecuada evaluación de la comprensión de lectura, particularmente, en las personas con LV.

2. *Organización del banco de ítems desarrollado por Soler (2013) y Espinosa (2013).* Teniendo en cuenta que estos ítems fueron construidos siguiendo el manual desarrollado por Espinosa (2013) y con base en la estructura definida por el Icfes para la subprueba de Lenguaje, esta actividad de organización consistió en codificar los ítems, trasladarlos al

formato de construcción diseñado y, finalmente, validarlos junto con los ítems desarrollados dentro del macroproyecto.

3. Construcción del banco de ítems del macroproyecto. Esta actividad implicó la selección y capacitación del personal para construir los ítems y la elaboración de protocolos de revisión de ítems. Los diez constructores seleccionados, que contaban con una amplia experiencia en el área de psicometría y, algunos de ellos, con amplios conocimientos en procesos cognitivos y comprensión de lectura en personas con LV, asistieron a una capacitación en la que se les dio a conocer la estructura de prueba, se les informó acerca de los aspectos que debían tener en consideración al construir los ítems de acuerdo con los protocolos diseñados, se les mostraron ejemplos de ítems de cada cruce de la estructura propuesta y se realizaron ejercicios de construcción.

Después de esta capacitación, se armaron dos grupos de constructores y se llevaron a cabo los talleres de construcción de manera independiente. Estos talleres tuvieron una duración aproximada de tres horas cada uno y, en total, se realizaron alrededor de diez sesiones por grupo.

4. Validación del banco de ítems. Esta validación solo se llevó a cabo con los ítems desarrollados por Soler (2013) y Espinosa (2013), y con los ítems construidos como parte del macroproyecto, ya que los ítems del Laboratorio de Psicometría fueron validados en años anteriores. El proceso de validación de los ítems mencionados se realizó siguiendo los protocolos diseñados para tal fin, los cuales incluían la evaluación de los aspectos formales de los ítems, la posible dificultad que estos podrían representar para los evaluados (clasificada en baja, media o alta), la respuesta correcta de cada uno de ellos y su justificación, y el cruce de componente y competencia al que pertenecían, entre otros.

Este proceso de validación se realizó haciendo uso de dos estrategias: una *validación cruzada*, realizada entre los dos grupos de constructores, y una *validación externa*, en la que participaron cinco expertos, dos del área de psicometría y tres del área de comprensión de lectura. Esta validación se llevó a cabo en tres sesiones de cuatro horas cada una.

El banco de ítems resultante de las actividades anteriores incluyó 21 ítems del banco del Laboratorio de Psicometría, 61 del banco desarrollado por Soler (2013) y Espinosa (2013), y 196 construidos como parte del macroproyecto, para un total de 278 ítems.

Para llevar a cabo la tercera actividad, *calibración del banco de ítems,* los 278 ítems fueron ensamblados y diagramados en cinco formas de prueba que fueron revisadas por varios miembros del equipo para verificar que no tuvieran errores y fueron codificadas con sus hojas de respuesta para facilitar su identificación; cada forma de prueba quedó compuesta por 55 ítems (aprox.) y con dos versiones diferenciadas por el orden de presentación de los ítems (versiones A y B). Estas cinco formas fueron aplicadas a muestras de 211 a 308 evaluados, tal como se expone en la tabla 2.

Tabla 2

Número de ítems y de evaluados por forma de prueba para la calibración del banco de ítems del TAI

Formas de prueba	Número de ítems	Número de evaluados		
		Versión A	Versión B	Total
1	55	170	120	290
2	54	170	138	308
3	54	171	134	305
4	58	98	113	211
5	57	101	110	211
Total	278	710	615	1325

Para convocar a las personas que participaron en la aplicación piloto de estos ítems, se contactaron rectores de colegios públicos y privados de diferentes ciudades del país, se les solicitó que permitieran que su institución participara en el estudio y, una vez aceptaron, se les pidió a sus estudiantes de grados 10.º y 11.º y a sus padres, su autorización para que les fueran aplicadas estas pruebas. El proceso de aplicación se realizó siguiendo el *Manual de aplicación de pruebas* diseñado dentro del macroproyecto y estuvo a cargo de personas con experiencia en estos procesos.

Después de realizar la aplicación, se consolidaron las bases de datos y se realizaron análisis de dificultad, discriminación y flujo de opciones desde la TCT. Estos resultados fueron revisados en compañía de los constructores y, derivado de este proceso, el banco del TAI quedó conformado por 257 ítems, 245 vinculados a textos continuos y 12 a textos discontinuos.

Finalmente, para dar cumplimiento al primer objetivo de este trabajo se analizó la información recolectada a partir del modelo de Rasch, se identificó la Función de Información de los ítems y se estableció la distribución de todos ellos a lo largo del continuo de habilidad.

La cuarta y última actividad de esta primera fase del proyecto, *diseño de instrumentos de recolección de evidencias de validez*, fue realizada de manera individual e incluyó el diseño del *Instrumento de familiaridad informática* y el *Instrumento de percepción de validez* (ver apéndices «A» y «B»), los cuales fueron diseñados con base en la revisión teórica realizada y en las características del TAI desarrollado.

Fase 2

Participantes.

La muestra de participantes de esta segunda fase estuvo conformada por estudiantes que presentaron la prueba Saber 11.º en las aplicaciones 2013-II y 2014-I. El total de estudiantes con LV fue contactado, mientras que para seleccionar al grupo de estudiantes sin LV se tuvo en cuenta que sus puntajes equiparados en la subprueba de Lenguaje presentaran la misma media y desviación estándar que los puntajes de la muestra de personas con LV, buscando que los dos grupos —personas con y sin LV— fueran comparables en su nivel de habilidad.

Estas personas fueron invitadas por los miembros del macroproyecto a participar en la aplicación del TAI. En total, 128 personas aceptaron, 40 de ellas presentaban algún tipo de LV (ceguera total, ceguera parcial o baja visión) y las 88 restantes presentaban una visión normal. Las personas sin LV provenían, en su mayoría, del departamento de Cundinamarca, específicamente, de la ciudad de Bogotá, salvo cuatro personas, una de las cuales provenía de Leticia, otra de San Andrés, otra de Buga y otra de Paipa. Las personas con LV residían en los departamentos de Cundinamarca, Atlántico, Antioquia, Santander, Valle de Cauca, Bolívar, Quindío, Sucre, Norte de Santander, Risaralda, Boyacá, Tolima, Caldas y Huila.

La muestra general se distribuyó en tres grupos en función del canal mediante el cual tuvieron acceso a la prueba, tal como se expone en la tabla 3.

Tabla 3

Distribución de la muestra que presentó el TAI

Plataforma del TAI	Evaluados		Total
	Sin LV	Con LV	
Auditiva	35	40	75
Visual	53	--	53
Total	88	40	128

Instrumentos.

• *Formatos de consentimiento y asentimiento informado.* Estos formatos se utilizaron para informar a los participantes mayores y menores de edad acerca de las características del estudio y para que autorizaran su participación en la aplicación del TAI.

• *Instrumento de familiaridad informática (FI).* Este instrumento se diseñó para evaluar la familiaridad de los individuos con las herramientas informáticas y presentaba dos versiones, la FI-S que incluía 17 preguntas sobre si los participantes tenían acceso a computadores e internet, la frecuencia de uso y los usos que le daban, entre otros aspectos; y la FI-O que incluía cinco actividades concretas que daban cuenta de la destreza de los individuos en el manejo del computador y se calificaban en una escala de 0 a 2 (0 si el individuo no lograba realizar la actividad, 1 si la realizaba con ayuda del evaluador, y 2 si la realizaba de manera autónoma).

• *Protocolo de entrenamiento en manejo del TAI.* Este protocolo se utilizó para igualar a la muestra en el manejo del TAI. Este entrenamiento ejemplificó la forma en la que los participantes podían y debían interactuar con la prueba, fue desarrollado con base en la manera en la que se diseñaron las plataformas usadas en el TAI, y su diseño estuvo a cargo de uno de los miembros del macroproyecto.

• *TAI de comprensión de lectura.* Este instrumento fue diseñado por los miembros del macroproyecto del que se desprende este estudio, está conformado por un banco de 245 ítems vinculados a textos continuos, calibrado y construido con base en los criterios establecidos por Espinosa (2013) para evaluar comprensión de lectura en personas con LV y en la estructura definida por el Icfes para la construcción de la subprueba de Lenguaje,

descrita en apartados anteriores. Adicionalmente, cuenta con dos plataformas, una para el acceso a la información por vía visual y otra para el acceso a la información por vía auditiva.

Los criterios de elección de ítems y de estimación de habilidad del TAI fueron definidos con base en los resultados obtenidos por las personas con LV en la subprueba de Lenguaje de la prueba Saber 11.º aplicada en 2013-II. Para la elección del ítem de arranque se tuvo en cuenta la habilidad media de esta población y se fijó un rango de habilidad en -1 desviación estándar, con el fin de seleccionar, de forma aleatoria, un ítem que se encontrara entre -1,524 y -0,887, es decir, en un valor cercano al nivel medio de habilidad de esta población; para la estimación de la habilidad se utilizó el algoritmo de máxima verosimilitud con ajuste de media ponderada de Warm; para la elección de los ítems intermedios se empleó el criterio de máxima información de Fisher; y, finalmente, como criterio de parada del TAI se optó por un valor fijo de error típico de medida de 0,4.

- *Instrumento de percepción de validez (PV).* Este cuestionario de 11 ítems se utilizó para evaluar las percepciones de los individuos sobre la calidad y la precisión de la evaluación. Las preguntas iniciales se enfocaban en cómo se sentía el individuo al ser evaluado mediante el TAI, las preguntas siguientes le solicitaban al individuo comparar su experiencia al haber presentado el TAI y la prueba Saber 11.º en el formato tradicional, y las preguntas finales abordaban particularidades de la versión de prueba que el individuo presentó (visual o auditiva), los aspectos positivos del TAI y, en caso de haber, otras observaciones.

- *Cuestionario de variables sociodemográficas.* Este instrumento fue diseñado por los miembros del macroproyecto del que se deriva este estudio e incluyó preguntas que abordaban variables de interés para los distintos subproyectos. Las variables de este cuestionario analizadas en este proyecto fueron sexo, edad, estrato socioeconómico, lugar de procedencia, sector de la institución educativa en la que estudiaron los participantes y tipo de LV, en caso de presentarla.

- *Paquetes de análisis de datos.* De acuerdo con las exigencias de los métodos que se utilizaron, se hizo uso de los programas SPSS v.22, y Winsteps v.3.73.

Procedimiento.

Teniendo en cuenta que el objetivo de esta fase era la aplicación del TAI y de los instrumentos adicionales, y el análisis de la información de las variables asociadas con validez, se realizaron tres grandes actividades: conformación de la muestra, aplicación de los instrumentos y el entrenamiento diseñados en la primera fase, y análisis de los datos recolectados. Las dos primeras actividades reportadas en esta fase fueron ejecutadas junto con los demás miembros del macroproyecto que engloba este estudio.

Para llevar a cabo la primera actividad —*conformación de la muestra*— el Icfes contactó a las personas seleccionadas por el equipo de investigación (104 personas con LV y 120 personas sin LV), les informó acerca de las características del macroproyecto y les pidió consentimiento de transmitir sus datos para que fueran contactados para la aplicación del TAI. Esta estrategia de contacto tuvo varios inconvenientes puesto que una gran cantidad de personas ya no se encontraba viviendo en el mismo lugar que había reportado cuando presentó la prueba Saber 11.º, había cambiado de número de teléfono o había dado datos erróneos, entre otros; y los que lograban ser contactados decidían no participar por no obtener remuneración. No obstante, se logró que 40 personas con LV aceptaran participar en el proyecto.

En vista de que no se obtenía respuesta de las personas sin LV contactadas, se decidió ampliar la muestra, pero la respuesta de este grupo continuó siendo reducida. Por lo anterior, el grupo de investigación tuvo que buscar otras estrategias de contacto que permitieran completar la muestra requerida, una de las cuales consistió en solicitarles a estudiantes de diferentes instituciones universitarias que presentaron la prueba Saber 11.º en los periodos 2013-II y 2014-I que participaran en el estudio. Esto permitió contar con una muestra de 88 personas sin LV.

La segunda actividad, correspondiente a la *aplicación de los instrumentos y entrenamientos*, en el caso de las personas sin LV, se llevó a cabo en las instalaciones del Laboratorio de Cognición y del Laboratorio de Psicometría de la Universidad Nacional de Colombia, sede Bogotá, dado el fácil acceso y las adecuadas condiciones de iluminación; y en el caso de las personas con LV, se llevó cabo en sus domicilios. Las sesiones de

aplicación se realizaron entre abril de 2015 y abril de 2016 en los horarios definidos por los participantes, y se desarrollaron de la siguiente manera:

Inicialmente, se procuró garantizar que las condiciones de aplicación fueran óptimas, después se aplicó el *Instrumento FI* (duración aproximada: 10 minutos) y el *entrenamiento en el manejo del TAI* (duración aproximada: 10 minutos); después se dio paso a la aplicación del *TAI de comprensión de lectura,* cuyo tiempo dependió de cada individuo, aunque en promedio tuvo una duración de hora y media; y, finalmente, a todos los individuos se les aplicó el *Instrumento PV* (duración aproximada: 10 minutos).

Para realizar la tercera actividad, *análisis de datos*, la información recolectada mediante los instrumentos diseñados fue clasificada de acuerdo con la categorización de evidencias de validez realizada por la AERA et ál. (2014) y a partir de allí se definieron los siguientes análisis:

• *Estructura interna de la prueba:* Con el fin de determinar la estructura del banco de ítems del TAI, inicialmente, se corroboró la viabilidad de realizar análisis factoriales mediante la medida de adecuación de Kaiser-Meyer-Olkin (KMO) y la Prueba de esfericidad de Bartlett, luego, se realizaron análisis exploratorios mediante el método de ejes principales con rotación varimax para cada forma de prueba y, teniendo en cuenta sus resultados, se volvieron a correr los análisis forzando las soluciones a tres factores, contemplando las posibles agrupaciones de acuerdo con la estructura de prueba, ya fuera por acciones (interpretativa, argumentativa y propositiva) o por componentes (función semántica de la información local, configuración del sentido global del texto y del sentido del texto hacia otros textos). Los ítems que no cargaron en ningún factor, es decir, que tuvieron cargas inferiores a |0,25|, fueron excluidos de los análisis.

• *Relación con otras variables:* Los estadísticos que se estimaron para establecer la relación entre la habilidad de los evaluados (estandarizada en z) y las variables sociodeomográficas abordadas fueron: U de Mann-Whitney para las variables *sexo*, *sector educativo* y *lugar de procedencia* (clasificada en ciudades principales y municipios secundarios en función de la cantidad de habitantes); coeficiente de Pearson para la variable *edad*; y H de Kruskal Wallis para las variables *estrato socioeconómico* y *tipo de LV* (Sarriá, Guàrdia y Freixa, 1999).

Los resultados de la aplicación de la primera parte del *Instrumento FI* (FI-S) fueron analizados de forma descriptiva, mientras que los de la segunda parte (FI-O), al medir la destreza de los evaluados mediante tareas concretas y asignársele una puntuación a cada tarea, fueron correlacionados con la habilidad estimada mediante el coeficiente Rho de Spearman. Finalmente, para analizar los resultados del *Instrumento PV*, además de realizar análisis descriptivos, se estimaron los estadísticos H de Kruskal Wallis y U de Mann Whitney para ver la relación entre las categorías de respuesta escogidas por los evaluados y su nivel de habilidad.

- *Procesos de respuesta.* La relación entre las variables tiempo por ítem y número de reproducciones, y el grupo al que pertenecían los evaluados (con o sin LV) fue evaluada mediante el estadístico U de Mann-Whitney.

Resultados

Los resultados de esta investigación se presentan organizados de acuerdo con los objetivos específicos definidos y las fases en las que se llevó a cabo el proyecto. Inicialmente, se aborda la confiabilidad de la prueba, y luego, se da paso a los objetivos relacionados con las evidencias de validez del TAI: Estructura interna, procesos de respuesta y relación con otras variables. Los resultados de confiabilidad y estructura interna de la prueba se derivan de la primera fase del proyecto, y los resultados de procesos de respuesta y relación con otras variables, de la segunda.

Evidencia de la estructura interna del banco para el TAI

En la tabla 4 se presentan los resultados generales de la calibración del banco de ítems para el TAI. Allí se puede observar que la dificultad de los ítems de este banco oscila entre -2,11 y 2,55 (desviación estándar = 0,74) y que el error de estimación promedio es de 0,14 (desviación estándar = 0,02). En relación con los indicadores de ajuste, se observa que algunos ítems están desajustados (ver mínimos y máximos de Infit y Outfit); no obstante, este resultado es esperable ya que se trata de los ítems que se encuentran ubicados en los extremos del continuo de habilidad y donde no hay personas con esos niveles de rasgo.

Tabla 4

Estadísticos descriptivos de la calibración del banco para el TAI

	Dificultad	Error estándar	Infit		Outfit	
			MNSQ	ZSTD	MNSQ	ZSTD
Media	0,00	0,14	1,00	0,00	1,00	0,10
D. E.	0,74	0,02	0,06	1,40	0,09	1,50
Mínimo	- 2,11	0,12	0,85	- 3,80	0,78	- 3,70
Máximo	2,25	0,19	1,21	5,40	1,55	5,30

**D. E. = Desviación estándar*

Confiabilidad.

En las figuras 2 y 3, que respectivamente presentan la Función de Información del TAI y el error estándar de medida, se observa que la prueba mide con una alta precisión los niveles medios de habilidad (-1 a 1) presentando errores de medida cercanos a 0,1, y que esta precisión decrece en los niveles extremos del continuo, llegando a presentar niveles de error cercanos a 0,3; no obstante, estos valores de error en general son bajos.

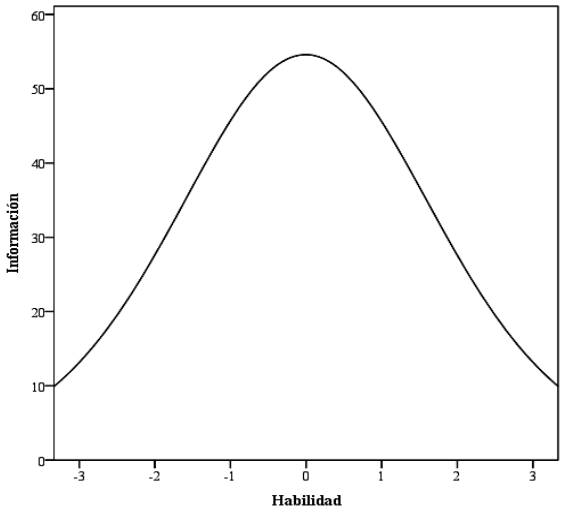

Figura 2. Función de Información del TAI

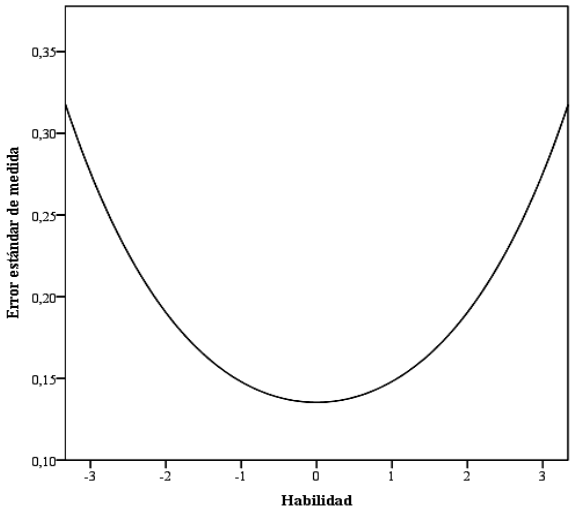

Figura 3. Error estándar de medida del TAI

La información anterior se reitera al observar el mapa de ítems-personas presentado en la figura 4, en el cual se observa que hay una mayor concentración de los ítems en los niveles medios de habilidad (-1 a 1) que en los niveles extremos, y que si bien los ítems solo abarcan una parte del continuo (-2 a 2), la distribución de las personas es similar a la de los ítems, y no se presentan casos de personas que presenten niveles de habilidad que se encuentren fuera de los niveles cubiertos por los ítems. Lo anterior, también es un indicador de la precisión de la medida que brinda el TAI.

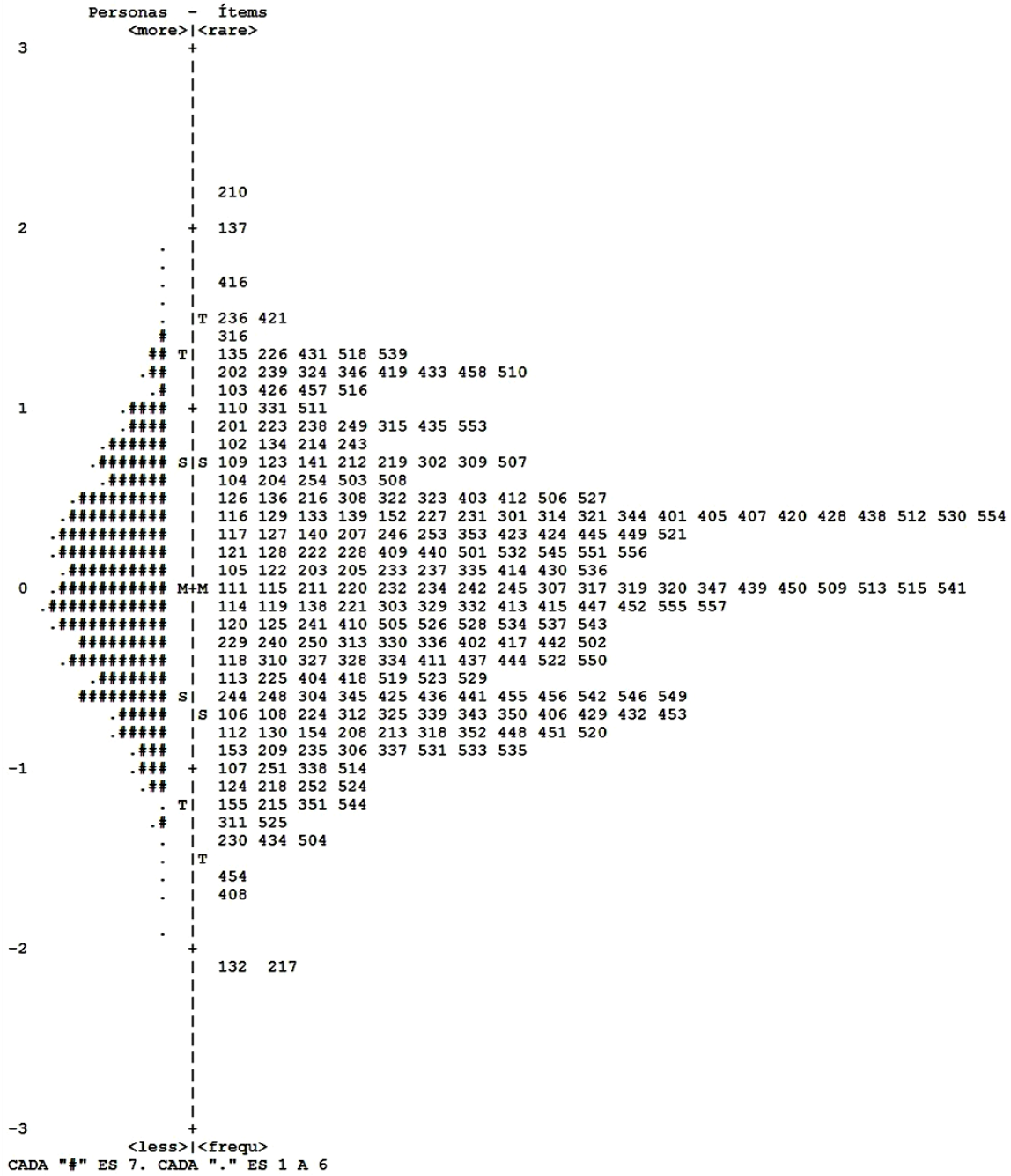

Figura 4. Mapa de ítems-personas del banco del TAI

Estructura factorial.

En la tabla 5, que muestra los resultados de la medida de adecuación de Kaiser-Meyer-Olkin (KMO) y la Prueba de esfericidad de Bartlett para las cinco formas de prueba, se puede observar que resulta adecuado realizar análisis factoriales con la información disponible.

Tabla 5

Estadístico KMO y prueba de esfericidad de Bartlett por forma de prueba del banco del TAI

Forma de prueba	KMO	Significancia Esfericidad de Bartlett
1	0,60	0,00
2	0,64	0,00
3	0,66	0,00
4	0,51	0,00
5	0,55	0,00

Los primeros análisis factoriales exploratorios realizados mediante el método de ejes principales arrojaron soluciones rotadas compuestas por tres factores con posible sentido teórico de acuerdo con la estructura de prueba sobre la que se elaboró el banco, los cuales explicaban alrededor del 17 % de la varianza. Los ítems que en estas primeras soluciones no cargaron en ningún factor se eliminaron, y los análisis se volvieron a correr forzándolos a tres factores, lo cual trajo como resultado porcentajes de varianza explicada del 20 % para la forma 1, 18 % para la forma 2, 21 % para la forma 3, 18 % para la forma 4 y 19 % para la forma 5. Las matrices factoriales resultantes se presentan en las tablas 6 a 10.

Como se puede observar en estas tablas, a pesar de haber excluido de los análisis los ítems que no cargaron en ningún factor, el porcentaje de varianza explicada por el primer factor de cada forma de prueba continuó siendo reducido; en las formas 1, 2 y 5 este porcentaje fue del 10 %, en la forma 3 fue del 11 % y en la forma 4 del 9 %, es decir, que los ítems del TAI están evaluando un constructo que no es unidimensional. No obstante, con el fin de dar soporte teórico a lo encontrado, el contenido de los ítems que cargaron en cada factor fue analizado.

Tabla 6
Matriz factorial forzada a tres factores
Forma 1 banco del TAI, excluyendo ítems

Ítem	Factor		
	1	2	3
I02			0,25
I03			
I06	0,27		
I07	0,31		
I09	0,41		
I11	0,32		
I12			
I13	0,40	0,28	
I14	0,27		
I15	0,28		
I16			0,49
I17			
I18		0,34	0,25
I19	0,58		
I20			
I21			0,43
I23			0,30
I24		0,39	
I25			0,40
I26			
I27			
I28			
I29			
I36			
I37			
I38	0,26	0,35	
I39		0,37	
I40			
I41			0,29
I52			
I53		0,53	
I54		0,32	
I55		0,39	
σ^2 explicada	9,8 %	5,7 %	4,8 %

Tabla 7
Matriz factorial forzada a tres factores
Forma 2 banco del TAI, excluyendo ítems

Ítem	Factor		
	1	2	3
I01	0,26	0,30	
I03		0,33	
I04			
I05			
I07		0,51	
I08		0,39	
I09		0,27	
I10			-0,25
I11			
I12			
I13	0,38		
I14	0,31		
I15			
I16	0,45		
I17			
I18			
I19	0,42		
I20		0,42	
I21		0,36	
I22			
I23	0,29		
I25			
I27	0,34		
I28			
I29	0,25		
I30			0,36
I31	0,26	0,28	
I32			
I33	0,28	0,27	
I34	0,34		
I35			
I36		0,30	
I37	0,26	0,30	
I38			
I39			
I40			
I41			
I42			
I43			
I44			
I45			0,31
I48	0,41		
I51			0,48
I52	0,33		0,44
σ^2 explicada	9,7 %	4,3 %	4,1 %

Tabla 8
Matriz factorial forzada a tres factores
Forma 3 banco del TAI, excluyendo ítems

Ítem	Factor		
	1	2	3
I02			
I03		0,37	
I04			
I07	0,26	0,35	
I08			
I09	0,26		
I10	0,39		
I11			
I12	0,31		
I13	0,35	0,27	
I14			
I15			
I17		0,29	
I18	0,32		
I19			
I20	0,37		
I21	0,29		
I22			
I23			0,33
I25	0,34		
I27	0,50		
I28	0,42		
I31			
I35	0,31		
I36			
I37	0,38		
I38	0,46	-0,28	
I39	0,38		
I43	0,27		
I44			0,33
I45	0,43		
I46			
I47	0,46		
I50	0,37	-0,33	
I51	0,41		
I52	0,35		
I53			0,26
σ^2 explicada	11,2 %	5,2 %	4,3 %

Tabla 9
Matriz factorial forzada a tres factores
Forma 4 banco del TAI, excluyendo ítems

Ítem	Factor		
	1	2	3
I01			
I02		0,30	
I03			0,42
I04		0,36	
I05		0,31	-0,30
I06		0,51	-0,26
I07			
I09	0,29		
I10			
I11		0,27	
I12			
I13			0,47
I16			
I17	0,35		
I18		0,38	
I19			
I20	0,27		
I21			
I23		0,31	
I25			
I29	0,32		
I30			
I31			
I32	0,38		
I35			
I36		0,38	
I37			
I38		0,27	
I39			0,31
I41	0,29		
I42	0,29		
I44			0,38
I47		0,37	
I48			
I49	0,30		
I50			
I51	0,41	0,30	
I52			
I54			
I55	0,53		
I56	0,39		
I57			
σ^2 explicada	8,8 %	4,7 %	4,5 %

Tabla 10

Matriz factorial forzada a tres factores Forma 5 banco del TAI, excluyendo ítems

Ítem	Factor		
	1	2	3
I01			
I02			0,26
I03			
I04	0,39		0,28
I05			0,34
I06	0,32		
I07	0,31		
I08			
I09	0,49		
I10			
I11			
I12	0,32		
I14	0,55		
I15	0,26		
I16	0,28		
I18			
I19	0,36		
I21			
I22		0,32	
I23	0,26	0,26	
I24		0,53	0,28
I25		0,47	
I26	0,26		
I28	0,29		
I29	0,31	0,32	
I31			
I32			
I33		0,41	
I34	0,27		
I35			
I37			
I39	0,30		
I41			
I42	0,35		
I43			
I44			-0,41
I46		0,32	
I50		0,41	
I51	0,41		
I55		0,40	-0,27
I56			
I57		0,46	-0,28
σ^2 explicada	9,5 %	4,9 %	4.5 %

En la revisión de contenido de los ítems, inicialmente, se intentó ver si los factores hallados coincidían con las dimensiones definidas en la estructura de la subprueba de Lenguaje de la prueba de Saber 11.º diseñada por el Icfes (acciones, componentes o cruce entre ambos); sin embargo, no se encontró coincidencia alguna. Dados estos resultados, se exploraron otras posibles agrupaciones, algunas derivadas de la forma en la que se construyó el banco (autor, grupo de construcción, procedencia del banco y tipos de textos utilizados) y otras derivadas de otros modelos de evaluación de la comprensión de lectura más sencillos como el propuesto por la Agencia de la Calidad de la Educación Escolar (ACEE) de Chile para la conformación de la prueba de *Lenguaje y Comunicación: Lectura* del Sistema de Medición de la Calidad de la Educación (Simce), que contempla la evaluación de tres habilidades necesarias para que se lleve a cabo la comprensión de lectura: localizar, interpretar y relacionar, y reflexionar (ACEE, 2014). No obstante, tampoco se encontraron correspondencias entre los factores hallados y estas posibles agrupaciones.

Evidencia de validez basada en la relación con otras variables

Algunas variables individuales que podrían estar relacionadas con el desempeño de los grupos fueron registradas. Estas variables son: sexo, edad, estrato socioeconómico, lugar de procedencia, sector de la institución educativa en la que estudiaron los participantes, tipo de LV —en caso de presentarla—, familiaridad informática y percepción de validez. Para efectos de la interpretación de la información que se presenta a continuación, las estimaciones de habilidad fueron estandarizadas en z.

Variables sociodemográficas.

El grupo de personas con LV estuvo compuesto por 55 % mujeres y 45 % hombres, y el grupo de personas sin LV, por 64 % mujeres y 36 % hombres. En relación con la habilidad estimada en los dos grupos, en la figura 5 se puede observar que la distribución de las estimaciones de habilidad de los hombres con y sin LV fue similar, sin embargo, hubo una mayor cantidad de hombres sin LV con niveles de habilidad inferiores a 0. Por otra parte, se observa que las mujeres con LV presentaron niveles de habilidad más homogéneos que las

mujeres sin LV (ver mínimos y máximos), y que, en general, los hombres alcanzaron niveles de comprensión de lectura superiores a los de las mujeres.

Finalmente, el estadístico U de Mann-Whitney estimado para hombres y mujeres de cada grupo —personas con y sin LV— arrojó como resultado que los niveles de habilidad de los hombres no presentaban diferencias (U = 389, p = 0,51), mientras que los de las mujeres sí los presentaban (U = 667, p = 0,04); siendo las mujeres con LV las que tenían los niveles de habilidad más altos. Adicionalmente, al calcular este estadístico para comparar las estimaciones de habilidad entre hombres y mujeres de cada grupo, se obtuvo como resultado que estas no presentaban diferencias. En el caso de personas sin LV este estadístico fue de 815 (p = 0,48) y en el caso de las personas con LV fue de 208,5 (p = 0,78).

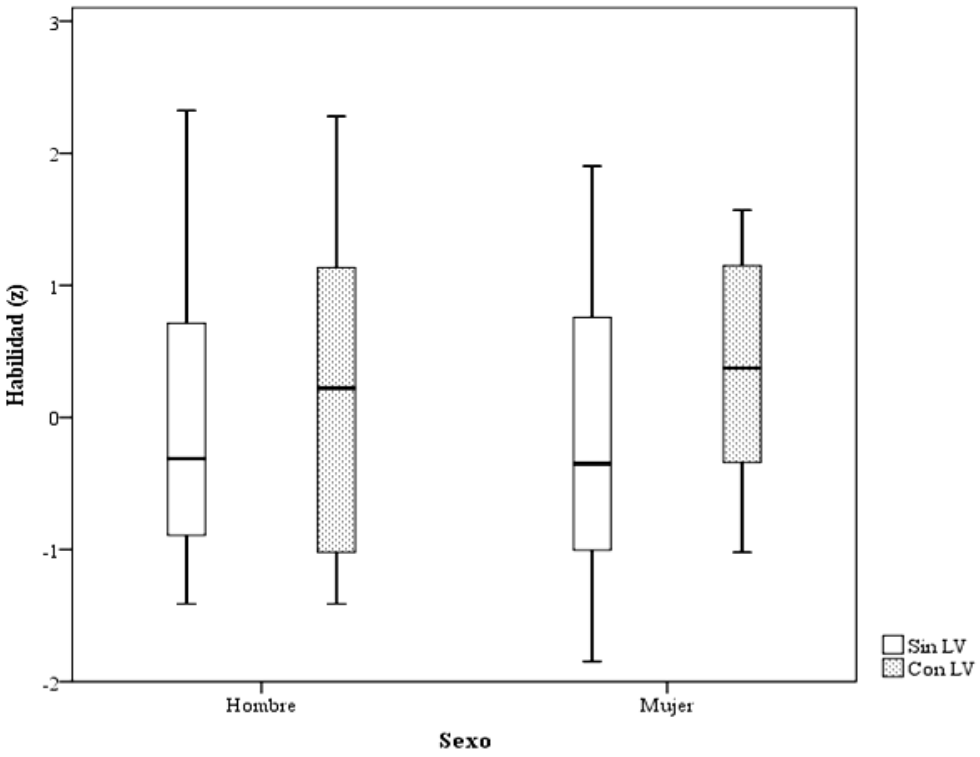

Figura 5. Comparación de las estimaciones de habilidad de las personas con y sin LV diferenciando por sexo

En relación con la variable ***edad***, se encontró que las personas con LV tenían entre 15 y 37 años, siendo la edad más común 18 años; mientras que las personas sin LV tenían entre 15 y 47 años, siendo la edad más común 17 años (en este grupo hubo 15 datos perdidos). Al estimar el coeficiente de Pearson para ver si las variables edad y habilidad se encontraban

asociadas linealmente, en el caso de las personas sin LV se encontró que estas variables no estaban correlacionadas (ρ = -0,06; *p* = 0,60); mientras que en el caso de las personas con LV esta correlación resultó baja positiva, siendo significativa al 0,05 (ρ = 0,31; *p* = 0,04), es decir, que cuanta mayor edad, mayor habilidad (ver figura 6).

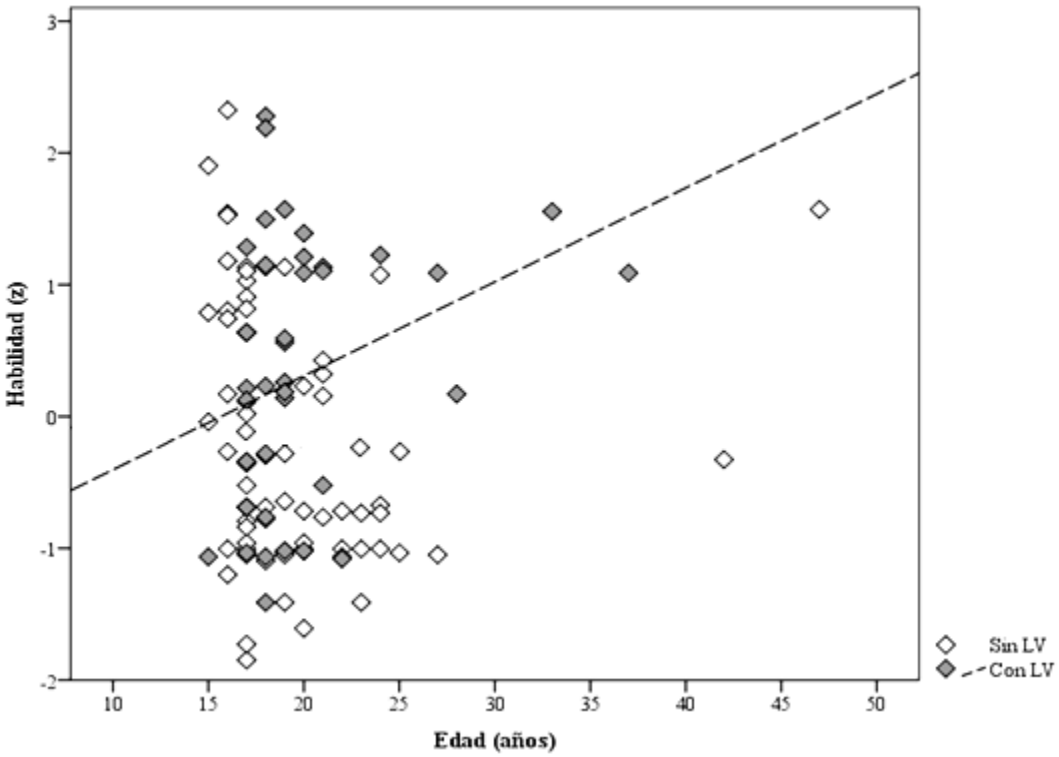

Figura 6. Comparación de las estimaciones de habilidad de las personas con y sin LV según edad

En relación con la variable de ***estrato socioeconómico***, se encontró que en el grupo de personas con LV hubo un 5 % de datos perdidos, correspondiente a dos personas, y en el grupo de personas sin LV, un 48 %, correspondiente a 42 personas. Las personas con LV pertenecían a estratos 0 a 6 y las personas sin LV a estratos 1 a 5 y, en los dos grupos, la mayor cantidad de participantes pertenecía al estrato 2 seguido del 3. Respecto a la estimación de habilidad por estrato, en la figura 7 se puede observar que los niveles de habilidad de las personas con LV se distribuyeron de manera similar en los estratos 1, 2 y 3, que eran los estratos en los que había más de un participante; y que los niveles de habilidad de las personas sin LV presentaron una menor dispersión cuanto mayor era el estrato

socioeconómico al que pertenecían (ver estratos 2, 3 y 4) aunque se debe tener en cuenta que solo cuatro personas pertenecían al estrato 4.

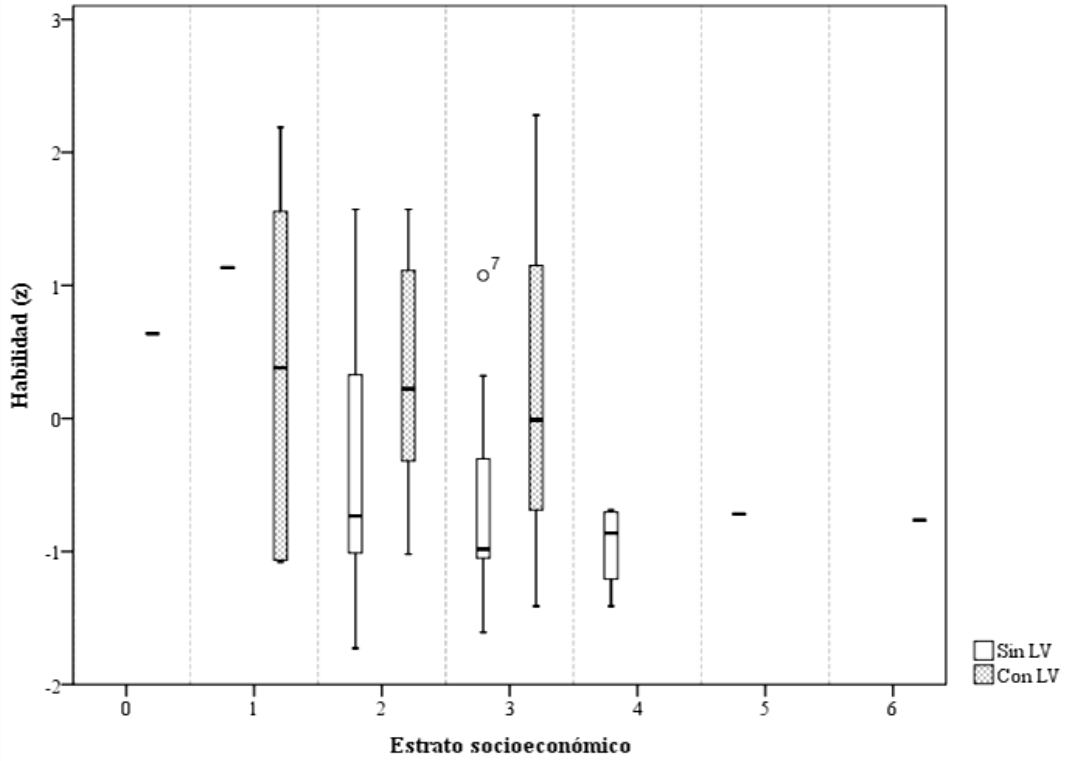

Figura 7. Comparación de las estimaciones de habilidad de las personas con y sin LV diferenciando por estrato socioeconómico

Por otro lado, al estimar el estadístico U de Mann Whitney para ver si existían diferencias entre las estimaciones de habilidad de las personas con y sin LV de los estratos 2 y 3 (estratos en los que había una mayor cantidad de personas) se encontró que, en los dos casos, las personas con LV presentaron niveles de habilidad significativamente más altos que las personas sin LV; este estadístico fue de 273 (p = 0,03) en el estrato 2, y de 161,5 (p = 0,04) en el estrato 3. Finalmente, al explorar la relación del estrato socioeconómico y el nivel de habilidad de las personas evaluadas, por medio del estadístico H de Kruskal Wallis, se encontró que estas variables no presentaban relación en ninguno de los grupos; en el caso de las personas sin LV este estadístico fue de 4,41 (p = 0,35) y en el caso de las personas con LV fue de 1,42 (p = 0,84).

Para analizar la variable de ***lugar de procedencia***, los municipios de los que provenían los participantes se clasificaron en *ciudades principales* y *municipios secundarios*, de

acuerdo con la cantidad de habitantes por municipio. Las ciudades principales, por tener más de 200 000 habitantes, incluyeron Bogotá, Medellín, Cali, Barranquilla, Cartagena, Cúcuta, Soledad, Ibagué y Bucaramanga, y los municipios secundarios incluyeron todos los municipios restantes del país. Con base en esto, se encontró que el 52,5 % de las personas con LV provenía de alguna de las ciudades principales nombradas y el 47,5 % restante de municipios secundarios; también, se encontró que el 95,5 % de las personas sin LV provenía de ciudades principales, específicamente de Bogotá, y el 4,5 % restante, de municipios secundarios.

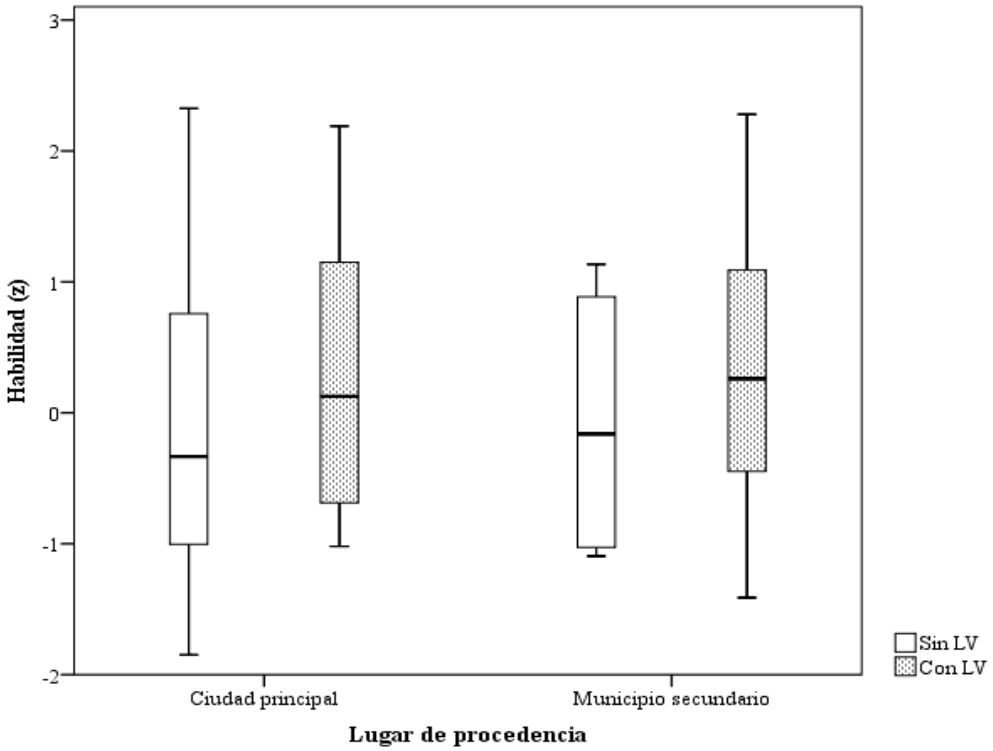

Figura 8. Comparación de las estimaciones de habilidad de las personas con y sin LV diferenciando por lugar de procedencia

Con respecto a la habilidad estimada en los grupos, en la figura 8 se puede observar que los niveles de habilidad de las personas sin LV provenientes de ciudades principales fueron más heterogéneos que los de aquellas provenientes de municipios secundarios (aunque cabe recordar que este segundo grupo estaba conformado solo por cuatro personas), y que las personas con LV provenientes de municipios secundarios alcanzaron niveles de habilidad más bajos que los provenientes de ciudades principales. Sin embargo, al estimar el estadístico

U de Mann-Whitney se encontró que no había diferencias en las estimaciones de habilidad de las personas con y sin LV que provenían de ciudades principales (U = 1113; p = 0,06) ni en las distribuciones de las que provenían de municipios secundarios (U = 43,5; p = 0,07).

Finalmente, al indagar si se presentaban diferencias entre las habilidades estimadas de las personas con y sin LV dependiendo de si procedían de ciudades principales o de municipios secundarios, mediante el estadístico U de Mann-Whitney, se encontró que en ninguno de los grupos había diferencias. En el caso de personas con LV este estadístico fue de 196 (p = 0,09), y en el caso de personas sin LV, fue de 164 (p = 0,09).

En cuanto al ***sector educativo*** de las instituciones en las que los participantes cursaron la secundaria, se encontró que en el grupo de personas con LV hubo un 8 % de datos perdidos en esta variable, y en el grupo de personas sin LV, un 20 %, por lo cual, los análisis se realizaron con 32 personas con LV y 46 sin LV. En la tabla 11 se presenta la cantidad y el porcentaje de participantes con y sin LV que se usó para el análisis, diferenciando entre aquellos que estudiaron en instituciones educativas públicas y aquellos que estudiaron en instituciones privadas; allí se puede observar que el porcentaje de estudiantes con LV que estuvo vinculado a instituciones educativas privadas es bastante reducido, a diferencia del grupo de personas sin LV, en el que este porcentaje fue similar al de personas que estudiaron en instituciones públicas.

Tabla 11

Distribución de personas con y sin LV que presentaron el TAI según sector de institución educativa

Sector institución educativa	Con LV		Sin LV	
	Cantidad	Porcentaje	Cantidad	Porcentaje
Público	25	78 %	21	44 %
Privado	7	22 %	25	46 %
Total	32	100 %	46	100 %

En relación con la habilidad estimada en los dos grupos, en la figura 9 se puede observar que en el caso de las personas sin LV, aquellos que estudiaron en colegios públicos presentaron niveles de habilidad más bajos que aquellos que estudiaron en colegios privados. Lo opuesto se presentó en el caso de las personas con LV, en el que las

personas que estudiaron en colegios públicos presentaron los niveles de habilidad más altos. No obstante, al estimar el estadístico U de Mann-Whitney para comparar las estimaciones de habilidad de los que estudiaron en colegios públicos y los que estudiaron en colegios privados de cada grupo —personas con y sin LV—, se obtuvo como resultado que estas no presentaban diferencias. En el caso de personas sin LV este estadístico fue de 279,5 (p = 0,70) y en el caso de las personas con LV fue de 54 (p = 0,13).

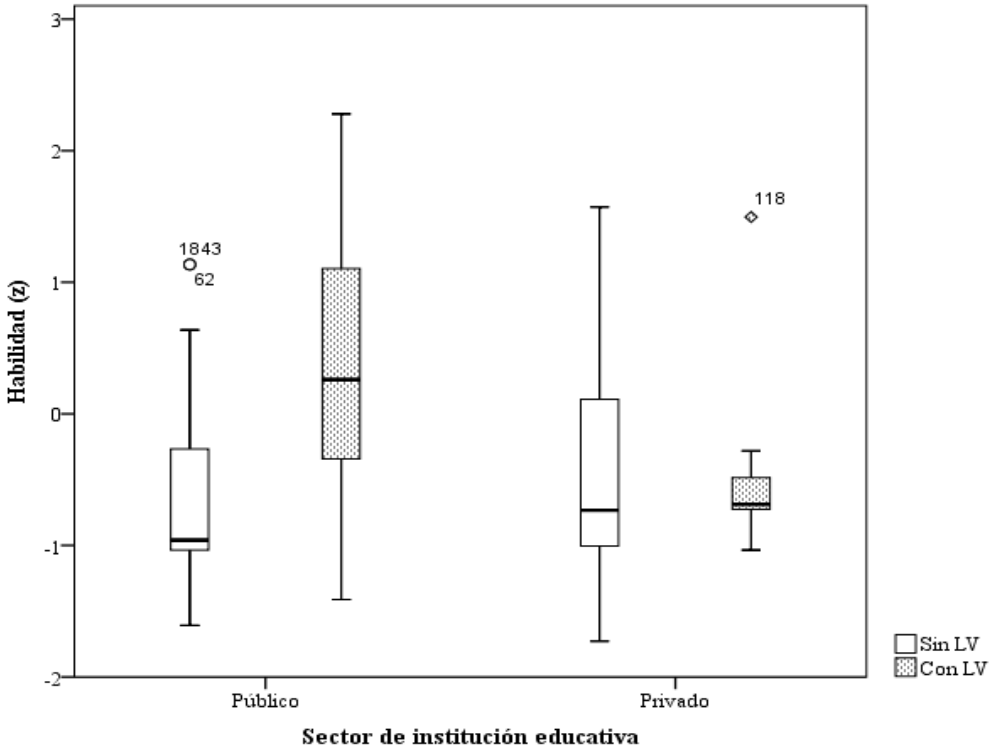

Figura 9. Comparación de las estimaciones de habilidad de las personas con y sin LV diferenciando por sector de institución educativa

Adicionalmente, al comparar el desempeño de las personas con y sin LV que estudiaron en colegios públicos se puede observar que las personas con LV alcanzaron valores máximos de habilidad más altos que las personas sin LV y que, incluso, los valores atípicos de este segundo grupo fueron inferiores a los valores máximos de habilidad alcanzados por las personas con LV. Esta idea se ve reforzada al ver los resultados del estadístico U de Mann-Whitney, los cuales indican que las habilidades estimadas de los que estudiaron en colegios públicos presentaban diferencias significativas (U = 377,5; p = 0,01); siendo el grupo con LV el que presentó los niveles de habilidad más altos.

Finalmente, en relación con las personas con y sin LV que estudiaron en colegios privados, en la figura 9 también se puede observar que las personas con LV presentaron niveles de habilidad mucho más homogéneos que las personas sin LV (recuérdese que este grupo estaba conformado por siete personas). No obstante, el estadístico U de Mann-Whitney arrojó como resultado que los niveles de habilidad entre estas personas no presentaban diferencias (U = 104,5; p = 0,44).

En lo que respecta a la variable de tipo de LV, en caso de presentarla, se encontró que el grupo de personas con LV estuvo conformado por 40 personas, 16 de ellas (40 %) presentaban baja visión, 8, ceguera parcial (20 %) y las 16 restantes ceguera total (40 %). En relación con el nivel de habilidad estimado para cada grupo con LV, en la figura 10 se puede observar que, en general, las personas con ceguera parcial presentaron los niveles más bajos de habilidad y las personas con ceguera total alcanzaron los niveles más altos; no obstante, al estimar el estadístico H de Kruskal-Wallis se encontró que no existen diferencias significativas entre los niveles de habilidad de las personas de los tres grupos (H = 2,78; p = 0,25).

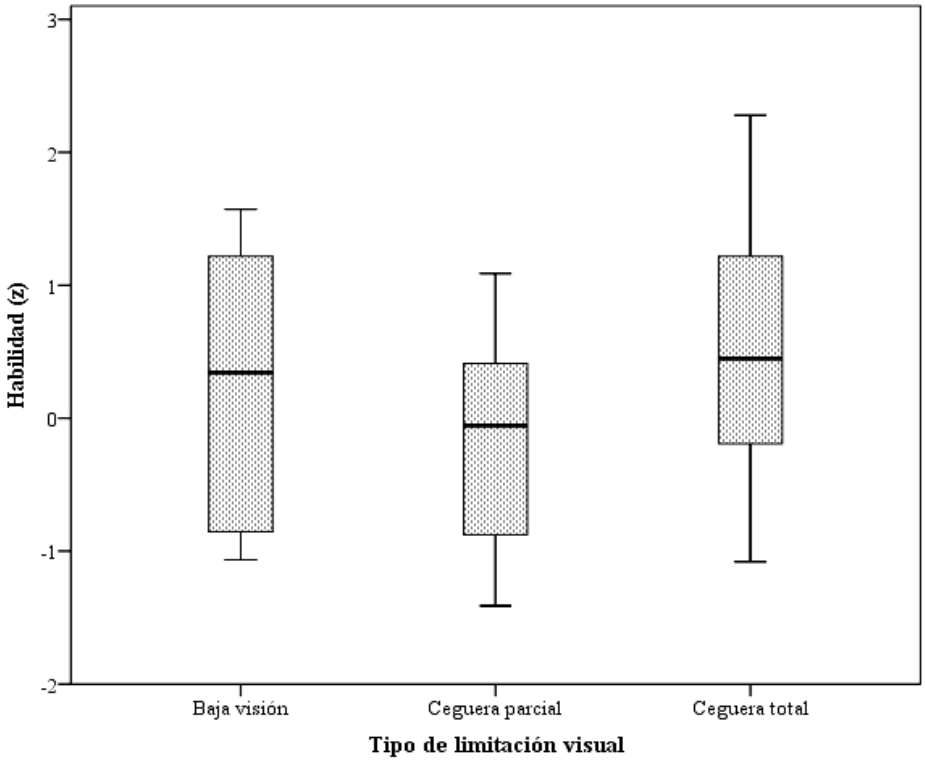

Figura 10. Comparación de las estimaciones de habilidad de las personas con LV diferenciando por tipo de LV

Variable de familiaridad informática.

El Instrumento FI, utilizado para recoger información sobre la variable de familiaridad informática, fue respondido por 51 participantes, 16 sin LV (31 %) y 35 con LV (69 %). A continuación, se presentará un resumen de la información recolectada mediante las 17 preguntas que conformaron la primera parte de este instrumento (FI-S), y luego, se dará paso a los resultados de la segunda parte (FI-O).

Las dos primeras preguntas del Instrumento FI fueron: *¿Tiene acceso a un computador?*, y si lo tiene *¿en qué lugares?* En el grupo de personas sin LV todos los participantes afirmaron tener acceso a esta herramienta y en el grupo de personas con LV dos personas (6 %) manifestaron no tenerlo. La mayoría de participantes de los dos grupos tenían acceso al computador en sus casas o en las instituciones educativas a las que se encontraban vinculados; tres personas (con LV) solo tenían acceso a computador en un lugar diferente a sus casas, y 12 personas sin LV (75 %) y 11 con LV (32 %) tenían acceso a esta herramienta en más de un lugar.

Las siguientes preguntas para las personas que afirmaron tener acceso a un computador fueron: *¿lo usa?*, si no lo usa *¿cuál es la razón?*, y si lo usa *¿con qué frecuencia semanal?*, *¿qué actividades realiza?* y *¿necesita ayuda para usarlo?* El 100 % de las personas sin LV y el 94 % con LV manifestaron usarlo; las razones dadas por las dos personas con LV que afirmaron no utilizar el computador a pesar de tener acceso a él fueron desconocimiento sobre cómo manejarlo y falta de práctica para escribir con el teclado. En cuanto al tiempo de uso de esta herramienta, se encontró que las personas sin LV dedicaban, en promedio, 26 horas semanales (desviación estándar = 14 h), y las personas con LV, 14 horas (desviación estándar = 10 h), y el tipo de actividades que realizaban eran de carácter académico y de ocio. Adicionalmente, el 100 % de las personas sin LV afirmó no necesitar ayuda para utilizar el computador, mientras que el 31 % de las personas con LV afirmó necesitarla para hacer uso de programas con los que no estaban familiarizados, ejecutar comandos que no podían usar con el lector de pantalla, ubicar en el teclado letras que no recordaban, solucionar fallas del computador, y acceder a información almacenada allí cuando no se tenía acceso a un lector de pantalla.

Las preguntas 8 y 9 fueron: *¿Tiene acceso a internet?* y si lo tiene *¿en qué lugares?* En el grupo de personas sin LV el 100 % de los participantes afirmó que sí lo tenía, mientras que en el grupo de personas con LV cuatro personas (11 %) manifestaron no tenerlo. El lugar en el que todas las personas sin LV y el 86 % de las personas con LV tenían acceso a internet era en casa, y el 44 % de las personas del primer grupo y el 20 % del segundo tenían acceso a internet en más de un lugar.

Las siguientes preguntas para las personas que afirmaron tener acceso a internet fueron: *¿la usa?*, si no la usa *¿cuál es la razón?*, y si la usa *¿con qué frecuencia semanal?*, *¿qué actividades realiza?* y *¿necesita ayuda para usarla?* El 100 % de las personas sin LV y el 97 % con LV manifestaron usar internet; la razón dada por la persona con LV que afirmó no utilizarla a pesar de tener acceso a ella fue desconocimiento sobre cómo usarla. En cuanto al tiempo semanal de uso de esta herramienta, se encontró que las personas sin LV dedicaban, en promedio, 26 horas semanales (desviación estándar = 14 h), y las personas con LV, 11 horas (desviación estándar = 9 h), y el tipo de actividades que realizaban las personas de los dos grupos también eran de carácter académico y de ocio. Finalmente, el 100 % de las personas sin LV afirmó no necesitar ayuda para usar internet, mientras que el 31 % de las personas con LV manifestó que requería ayuda para navegar por redes sociales, y para los mismos aspectos mencionados en el apartado anterior.

Las últimas preguntas fueron: *¿Está familiarizado con algún software de lectura?*, si lo está *¿cuál?* y *¿para qué lo utiliza?* En el grupo de personas sin LV, todos los participantes manifestaron no estar familiarizados con este tipo de programas, mientras que en el grupo de personas con LV el 54 % manifestó sí estarlo. Estas personas afirmaron conocer y usar los programas Jaws, Convertic, Text Aloud y NVDA, para todo lo que tenía que ver con el uso del computador, desde prenderlo hasta descargar programas, usar internet, y consultar información de interés personal como mencionó uno de los evaluados: «lo utilizo para investigar cómo arreglar aparatos electrónicos y buscar contenidos de humor».

La segunda parte del Instrumento FI (FI-O) contenía cinco actividades que evaluaban la destreza de los participantes en el uso del computador y se calificaban en una escala de 0 a 2. Para determinar si la puntuación obtenida en esta parte del instrumento se encontraba relacionada con el nivel de habilidad de los evaluados se estimó el estadístico Rho de

Spearman; sin embargo, solo se calculó para el grupo de personas con LV ya que todas las personas sin LV obtuvieron una puntuación perfecta en las actividades. La correlación obtenida en el grupo de personas con LV fue de -0,07 (p = 0,68), es decir, que no se encontró relación lineal entre la familiaridad informática y la estimación de habilidad de los evaluados con LV. Este hallazgo se ve reforzado al observar los resultados de la correlación estimada entre la habilidad de los evaluados y el tiempo de uso del computador y de internet para los dos grupos, que en el caso de personas sin LV fue de 0,09 (p = 0,74) y -0,12 (p = 0,54), respectivamente, y en el caso de personas con LV fue de 0,09 (p = 0,74) para la variable de tiempo de uso del computador y -0,33 (p = 0,08) para la variable de tiempo de uso internet.

Variable de percepción de validez.

El instrumento diseñado para recoger información de esta variable (Instrumento PV) fue respondido por 52 participantes, 31 con LV (60 %) que por su condición presentaron la versión auditiva de la prueba, y 21 sin LV (40 %), de los cuales 10 presentaron la versión auditiva y 11 la versión visual. A continuación, se presenta la información recolectada mediante las 11 preguntas que conformaron este instrumento.

En relación con la pregunta inicial: *¿Siente que le fue mejor, peor o igual en esta prueba que en la prueba de Lenguaje que presentó en Saber 11.º?*, en la figura 11 se puede observar que las opciones escogidas por la mayoría de los participantes sin LV fueron «peor» (48 %) e «igual» (43 %). Los argumentos dados por aquellos que escogieron la opción «peor» y que presentaron la versión auditiva de la prueba fueron falta de concentración y dificultad para responder debido al canal de entrada de la información; y de aquellos que presentaron la versión visual, fueron dificultad para entender los textos y desconocimiento de algunas palabras. Por su parte, las personas que respondieron «igual» argumentaron que las dos pruebas eran «del mismo estilo», tenían el mismo objetivo y el mismo tipo de preguntas, y hacían uso de textos fáciles y difíciles, además, que habían obtenido resultados similares en las dos pruebas (en esta categoría no se hizo distinción entre los que presentaron la versión auditiva y la visual ya que en los dos grupos se argumentó lo mismo). Finalmente, dos personas seleccionaron la opción «mejor», una de

ellas afirmó que había entendido más las lecturas del TAI y, la otra, que el TAI le permitía responder de forma autónoma.

En la figura 11 también se observa que la opción de respuesta más escogida por las personas con LV fue «mejor» (68 %), seguida de «peor» (16 %), «igual» (10 %), y por último, «no sabe» (6 %). Los argumentos dados por las personas que manifestaron que les fue «mejor» en el TAI que en la prueba de Lenguaje de Saber 11.º se enfocaron en la autonomía para responder la prueba, la posibilidad de repetir la información, la buena calidad de los audios y de la lectura realizada por el locutor, la claridad de las instrucciones, la posibilidad de acceder a la información por vía auditiva y la ausencia de imágenes. Las personas que consideraron que les fue «peor» argumentaron que esto se debía a que experimentaron cansancio, falta de presión puesto que los resultados que obtuvieran no tendrían incidencia en su vida, y falta de comprensión de los textos. Finalmente, las personas que escogieron la opción «igual» afirmaron que habían obtenido resultados similares en las dos pruebas y que en las dos había preguntas fáciles y difíciles; una de las dos personas que respondieron «no sabe», dijo que se había tensionado en la mitad de la prueba, y la otra, no argumentó su respuesta.

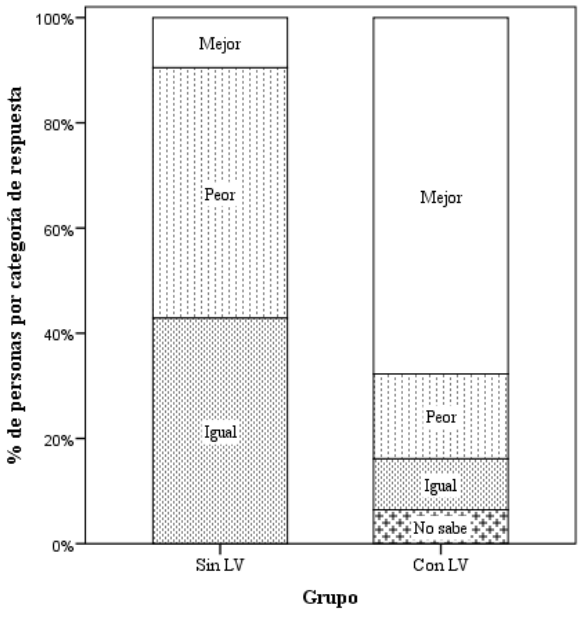

Figura 11. Comparación de la distribución de los participantes con y sin LV por categoría de respuesta del ítem 1 del Instrumento PV

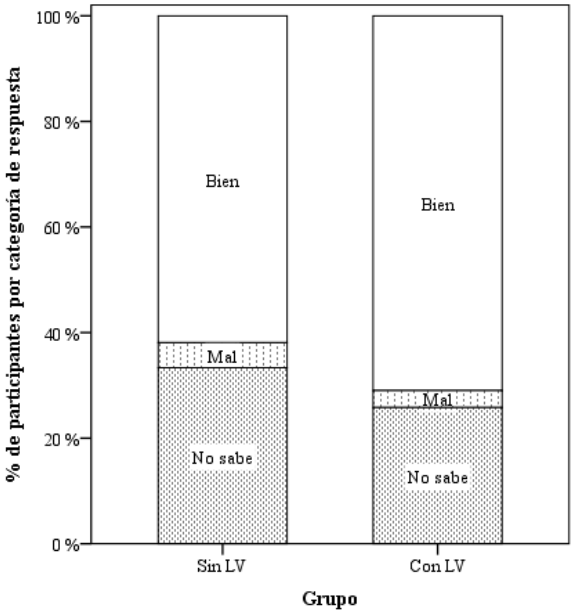

Figura 12. Comparación de la distribución de los participantes con y sin LV por categoría de respuesta del ítem 2 del Instrumento PV

La segunda pregunta fue: *¿Siente que fue bien o mal evaluado con esta prueba o no sabe?* Al respecto, en la figura 12 se puede observar que la opción escogida por la mayoría de participantes de los dos grupos fue «bien»; en el caso de las personas sin LV, esta opción fue elegida por el 62 %, y en el caso de las personas con LV, por el 71 %. Las personas sin LV que seleccionaron esta opción afirmaron que sintieron que fueron bien evaluados porque la prueba evaluaba comprensión de lectura y estaba bien elaborada; las preguntas se derivaban de los textos; las instrucciones, los textos y las preguntas eran claras; la prueba era computarizada y hecha por expertos; y el sistema brindaba las herramientas necesarias para responder la prueba. Además de dar estos argumentos, las personas con LV afirmaron que sintieron que fueron bien evaluados porque entendían más que cuando un tercero les leía, tenían control sobre sus respuestas y podían repetir los audios y pensar mejor.

En cada grupo hubo una persona que afirmó haber sentido que fue «mal» evaluado con el TAI; al respecto, la persona con LV afirmó que había palabras desconocidas o muy técnicas que no le permitieron responder la prueba adecuadamente, y la persona sin LV afirmó que no había comprendido bien los textos. Finalmente, la opción «no sabe» fue seleccionada por 33 % de las personas sin LV y el 26 % de las personas con LV; el argumento predominante de haber seleccionado esta opción fue que los participantes desconocían sus resultados.

En cuanto a la tercera pregunta: *¿Se sintió cómodo o incómodo presentando esta prueba?*, en la figura 13 se puede observar que, en los dos grupos, la mayoría de personas manifestó haberse sentido «cómoda»; en el grupo de personas sin LV esta opción fue escogida por el 81 %, y en el grupo de personas con LV, por el 94 %. Las personas sin LV manifestaron haber escogido esta opción porque el aplicativo era algo novedoso y fácil de manejar, no había tiempo límite, se encontraban en un ambiente tranquilo y sin ruido, y los textos, las preguntas y la forma en la que narraba el locutor eran claros. Además de lo anterior, las personas con LV afirmaron que su comodidad se debía a que tenían autonomía para responder la prueba, no dependían de un tercero y podían concentrarse más debido al uso de audios.

En la figura 13 también se puede observar que la cantidad de personas que manifestó sentirse «incómoda» presentando la prueba fue mayor en el grupo de personas sin LV que

en el grupo de personas con LV. El argumento de las cuatro personas sin LV que escogieron esta opción residió en su imposibilidad de ver los textos, ya que presentaron la versión auditiva; y el de las dos personas con LV fueron, respectivamente, haber sentido estrés y encontrarse en época de parciales.

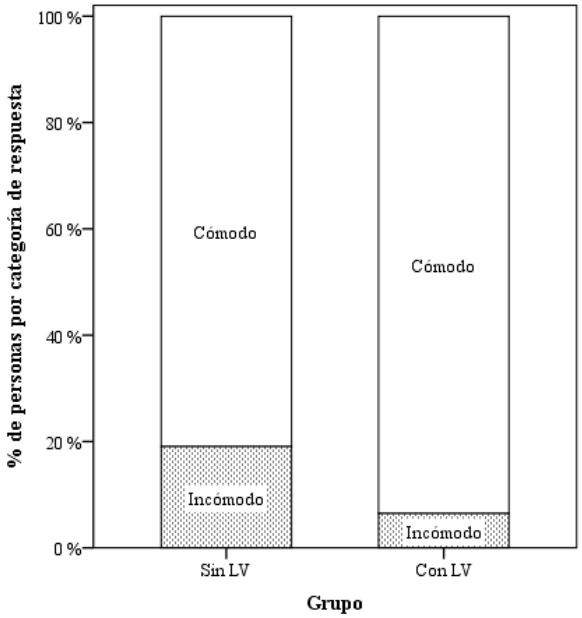

Figura 13. Comparación de la distribución de los participantes con y sin LV por categoría de respuesta del ítem 3 del Instrumento PV

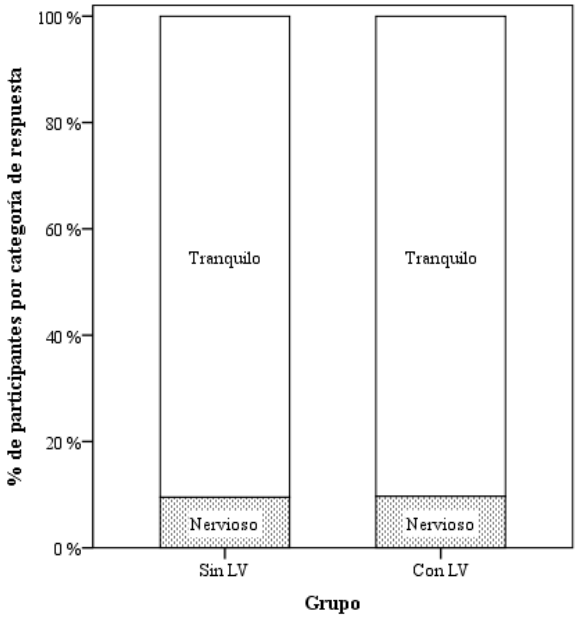

Figura 14. Comparación de la distribución de los participantes con y sin LV por categoría de respuesta del ítem 4 del Instrumento PV

La cuarta pregunta fue: *¿Se sintió nervioso o tranquilo presentando esta prueba?* En la figura 14 se puede observar que las opciones de respuesta se distribuyeron de manera equivalente en los dos grupos —personas con y sin LV—, siendo la opción más escogida «tranquilo». Los participantes sin LV que seleccionaron esta opción afirmaron que se sentían así porque la prueba no tenía consecuencias para ellos, no sentían presión por obtener un buen resultado, las instrucciones dadas por los aplicadores fueron claras, no tenían límite de tiempo, no había ruido en el lugar de aplicación, y tenían las herramientas necesarias para dar respuesta a la prueba. Además de dar estos argumentos, las personas con LV agregaron que estaban familiarizados con el tipo de pregunta utilizado, tenían autonomía para responder la prueba, el lugar en el que se realizó la aplicación era familiar,

y el acceder a la información mediante audios pregrabados les hacía sentir confianza y les quitaba la presión de tener que solicitarle a un tercero que les repitiera.

Finalmente, cinco personas manifestaron haberse sentido «nerviosas» respondiendo la prueba, dos sin LV y tres con LV. Una de las personas sin LV afirmó que se había sentido así debido a que estaba desconcentrada y no sabía si estaba respondiendo correctamente, y la otra, pensaba que la prueba tendría consecuencias para ella. Por su parte, dos de las tres personas con LV afirmaron que habían escogido esta opción porque no sabían si estaban respondiendo correctamente y, la tercera persona, que le faltaba experiencia en ese tipo de pruebas.

Respecto a la quinta pregunta: *En comparación con la prueba que presentó en formato de lápiz y papel, ¿considera que esta prueba computarizada tiene más, menos o igual probabilidad de presentar errores en la calificación?*, en la figura 15 se puede observar que las opciones escogidas por la mayoría de personas sin LV fueron «menos» (43 %) e «igual» (43 %). Las personas de este grupo que consideraron que el TAI, por ser una prueba computarizada, presentaba menos probabilidad de error en la calificación afirmaron que esto se debía a que el sistema utilizado era «confiable», había sido programado, generaba los resultados automáticamente y disipaba la posibilidad de que se cometieran errores en la lectura de las hojas de respuesta; las personas que consideraron que esta probabilidad era «igual» argumentaron que tanto en la prueba de lápiz y papel como en el TAI, era una máquina la que realizaba la calificación, de manera que cualquiera de estos mecanismos podía presentar errores. Finalmente, tres personas afirmaron que el TAI podía presentar «más» errores en la calificación debido a que, en general, los programas presentaban fallas y a que el TAI no permitía revisar o corregir las respuestas de las preguntas respondidas con anterioridad.

La opción más escogida por las personas con LV fue «menos» (84 %) seguida de «igual» (10 %), y de «más» (3 %) y «no sabe» (3 %). Las razones reportadas por las personas de este grupo para escoger la opción «menos» fueron la posibilidad brindada por el sistema para confirmar las respuestas antes de realizar su envío y para repetir la información cuantas veces fuera necesario; la practicidad del sistema ya que permitía responder de forma autónoma, eliminando la posibilidad de que los lectores se equivocaran o «alteraran» las respuestas; y el diseño del sistema para entregar los resultados de forma

inmediata. Las personas de este grupo que seleccionaron la opción «igual» afirmaron que las dos pruebas evaluaban lo mismo y, por tanto, no variaba la probabilidad de que se cometieran errores al calificar. Por su parte, el participante que escogió la opción «más» afirmó que esto se debía a que había un mayor número de variables involucradas en la evaluación; y por último, quien escogió la opción «no sabe» dijo que la cantidad de errores dependía de qué tan bien estuviera hecho el software utilizado.

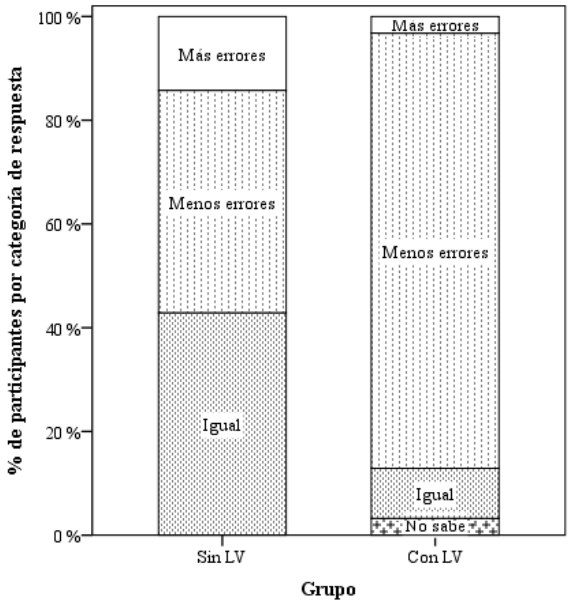

Figura 15. Comparación de la distribución de los participantes con y sin LV por categoría de respuesta del ítem 5 del Instrumento PV

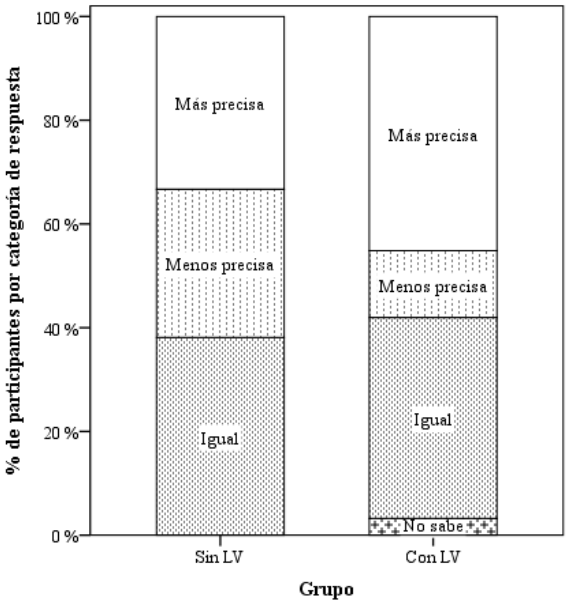

Figura 16. Comparación de la distribución de los participantes con y sin LV por categoría de respuesta del ítem 6 del Instrumento PV

La sexta pregunta fue: *En comparación con la prueba que presentó en formato de lápiz y papel, ¿considera que, al tener menos ítems, esta prueba computarizada es más, menos o igual de precisa en la calificación?* En la figura 16 se puede observar que, en el caso de las personas sin LV, la mayoría consideró que la prueba computarizada y la de lápiz y papel generaban calificaciones «igual» de precisas (38 %), y en el caso de las personas con LV, que esta precisión era mayor en el TAI (45 %). Aquellos que consideraron que la calificación del TAI era «menos» precisa representaron el menor porcentaje de personas en los dos grupos, aunque cabe anotar que hubo una persona con LV que manifestó no saber la respuesta a esta pregunta.

En ambos grupos, las personas que consideraron que, por tener menos ítems, el TAI generaba calificaciones «más» precisas afirmaron que esto se debía a que experimentaron menos cansancio al resolver la prueba, tuvieron mayor claridad sobre la situación de examen y sobre los resultados que obtendrían, las preguntas utilizadas en el TAI eran más claras y estaban dirigidas a evaluar lo mismo, y los resultados se podían obtener en poco tiempo. Por su parte, las personas que consideraron que el nivel de precisión de la calificación era «igual» afirmaron que este no dependía de la cantidad de ítems sino de la calidad que tuvieran, que las dos pruebas evaluaban lo mismo y que si bien el TAI tenía menos ítems, el tiempo de aplicación había sido extenso, así que esto nivelaba las dos pruebas (esto último lo afirmó solo un par de personas). Las personas que consideraron que el TAI generaba calificaciones «menos» precisas argumentaron que si la prueba tenía menos preguntas los evaluados tenían menor probabilidad de acertar, la evaluación era más superficial ya que no se evaluaba todo lo que se podría evaluar, y había «mayor margen de error» y «menor eficiencia estadística». Finalmente, una persona afirmó que «no sabía» si el TAI generaba calificaciones más, menos o igual de precisas a la prueba de lápiz y papel debido a que no recordaba cómo eran los ítems de la otra prueba y por lo tanto, no tenía punto de comparación.

Respecto a la séptima pregunta: *En comparación con la prueba que presentó en formato de lápiz y papel ¿considera que esta prueba computarizada es más fácil, más difícil o igual de fácil o difícil?*, en la figura 17 se puede observar que la mayoría de personas sin LV (57 %) consideró que el TAI y la prueba de lápiz y papel presentaban «igual» dificultad debido a que las dos pruebas tenían preguntas sencillas y complejas, la tarea que se debía realizar era la misma, lo único diferente entre las dos pruebas era la forma de presentación de la información, y la capacidad de respuesta solo dependía de la concentración y de la comprensión de lectura que cada quién tuviera. El 24 % de las personas de este grupo consideró que el TAI era «más fácil» debido a que no experimentaron cansancio, el sistema les permitía pensar más las respuestas y responder con tranquilidad, no había tiempo límite y se sintieron más seguros de sí mismos. El 19 % restante consideró que la prueba computarizada era más «difícil» que la de lápiz y papel, ya que presentaron la versión auditiva de la prueba y no lograron concentrarse.

Por su parte, la mayoría de personas con LV (74 %) consideró que el TAI era «más fácil» debido a que se podía manejar de forma autónoma y sencilla, tenía menos preguntas, requería menos tiempo para su resolución y, por tanto, no generaba cansancio, los audios se podían controlar y repetir, no había preguntas asociadas a imágenes, se podía interactuar con el texto y con las preguntas y, por ello, era más fácil interpretar la información presentada, y resultaba ser una forma de evaluación más incluyente puesto que daba la posibilidad de ser respondida por personas con y sin LV. El 19 % consideró que la prueba computarizada y la de lápiz y papel tenían una dificultad equivalente debido a que la tarea que se debía realizar era la misma y que las preguntas eran similares. Por último, dos participantes, que representaban el 7 %, consideraron que la prueba computarizada era más «difícil» que la de lápiz y papel; una de estas personas afirmó que esto se debía a que no logró concentrarse y, la otra, que el contenido de las preguntas del TAI era más complejo.

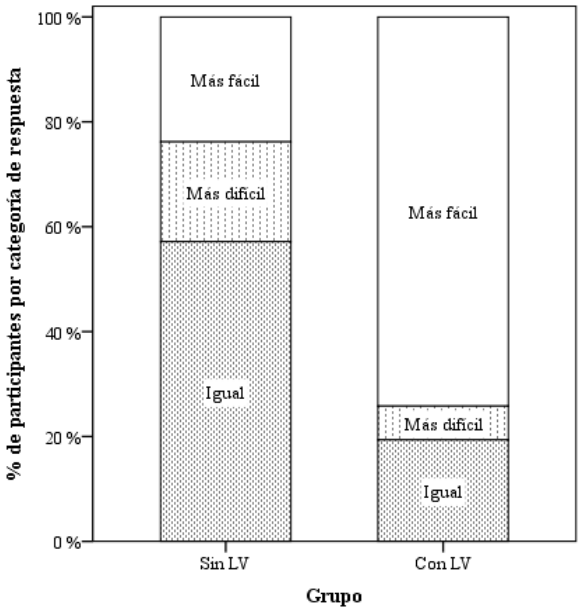

Figura 17. Comparación de la distribución de los participantes con y sin LV por categoría de respuesta del ítem 7 del Instrumento PV

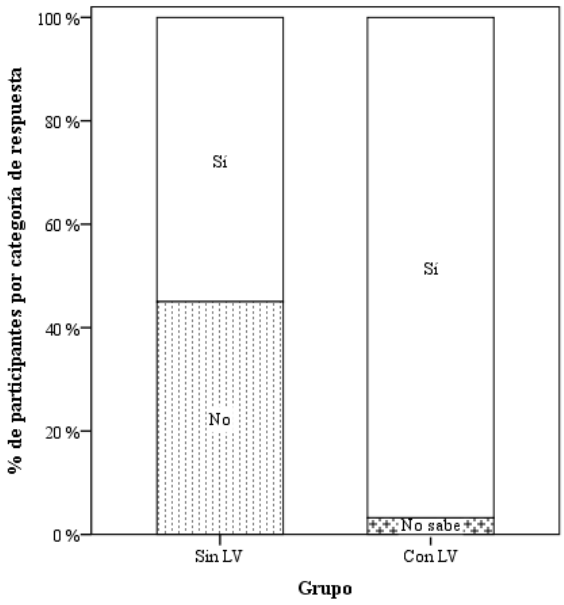

Figura 18. Comparación de la distribución de los participantes con y sin LV por categoría de respuesta del ítem 8 del Instrumento PV

La octava pregunta fue: *¿Prefiere ser evaluado mediante este mecanismo que mediante la prueba tradicional de lápiz y papel?* En la figura 18 se puede observar que el 97 % de las personas con LV y el 52 % de las personas sin LV prefirieron ser evaluados mediante el TAI. Las personas con LV que dieron esta respuesta afirmaron que la prueba

computarizada era más sencilla, les daba la posibilidad de repetir la información cuantas veces lo consideraran necesario, igualaba sus condiciones de evaluación con las de las personas sin LV, les permitía acceder a la información por vía auditiva, les brindaba autonomía, generándoles confianza y seguridad al responder, los hacía sentir más cómodos y tranquilos, les permitía pensar con mayor claridad, y eliminaba la posibilidad de que alteraran sus respuestas. Por su parte, las personas sin LV que manifestaron preferir ser evaluados con este mecanismo afirmaron que ello se debía a que el TAI ayudaba al medio ambiente al no requerir papel, no dañaba el espacio físico, era más cómodo y flexible, no generaba cansancio, les daba seguridad a los evaluados para responder, los textos utilizados eran más claros, y resultaba más fácil leer por ser presentada en el computador.

Las personas sin LV que manifestaron preferir ser evaluados mediante las pruebas de lápiz y papel tradicionales afirmaron que estaban acostumbrados a ser evaluados con este mecanismo, les resultaba más cómodo leer en papel, el computador les cansaba la vista, y al poder ver la información tenían la posibilidad de cambiar las respuestas y de conocer la cantidad de texto que debían leer. En el grupo de personas con LV, ningún participante manifestó preferir ser evaluado mediante una prueba de lápiz y papel, y solo uno (con diagnóstico de baja visión) afirmó no saber si prefería ser evaluado mediante TAI o mediante el formato tradicional ya que cada forma de evaluación tenía desventajas, por ejemplo, con la prueba de lápiz y papel se agotaba su vista, y con la computarizada se demoraba más.

La novena pregunta indagaba sobre la calidad de la forma de presentación de la prueba; a aquellos que presentaron la versión auditiva se les preguntó: *¿La voz del aplicativo fue clara, presentaba una entonación apropiada? ¿Tuvo inconvenientes relacionados con la calidad del sonido?,* y a aquellos que presentaron la versión visual: *¿La letra del aplicativo fue clara, presentaba tamaño y color apropiados? ¿Tuvo inconvenientes relacionados con la calidad de la imagen?* En el grupo de personas sin LV, nueve de los diez evaluados que presentaron la versión auditiva de la prueba respondieron esta pregunta y afirmaron que no tuvieron inconvenientes con la calidad del sonido y que la voz fue clara y tenía una entonación adecuada; por su parte, uno de los once evaluados que presentaron la versión visual de la prueba afirmó que la letra del aplicativo era muy pequeña, y los demás

afirmaron que no tuvieron inconvenientes con la calidad de la imagen y que la letra había sido clara y con tamaño y color apropiados. Finalmente, el 84 % de las personas con LV manifestó que no tuvo inconvenientes con la calidad del sonido, y el 16 % restante (5 personas), que algunos audios tenían más volumen que otros, la entonación de las poesías no era la mejor y el sonido a veces se paraba.

En relación con la décima pregunta: *¿Considera que esta prueba tiene aspectos positivos?,* solo dos evaluados del grupo de personas sin LV respondieron negativamente, y todos los evaluados del grupo de personas con LV dieron una respuesta afirmativa. Al indagar acerca de cuáles eran esos aspectos positivos de la prueba, las personas sin LV afirmaron que el TAI aportaba elementos para el mejoramiento de la calidad de la evaluación, era cómodo, ecológico y de fácil manejo, contaba con ítems claros y bien elaborados, y con textos interesantes, brindaba tranquilidad para responder, tenía pocas preguntas, y representaba muchas ventajas para las personas con LV permitiéndoles tener las mismas condiciones de evaluación que ellos tenían. Por su parte, las personas con LV afirmaron que el TAI les brindaba autonomía, confianza y tranquilidad, tenía instrucciones de manejo claras y sencillas, los audios eran de buena calidad, las preguntas y las lecturas eran claras, concretas y bien elaboradas, era un mecanismo de evaluación interesante, novedoso, inclusivo, cómodo, accesible para personas con LV, bien estructurado y menos costoso que las pruebas tradicionales, mejoraba las posibilidades de las personas con LV ya que estaban siendo mejor evaluadas y facilitaba la concentración al contar con una voz humana, no robotizada.

Por último, respecto a la undécima pregunta: *¿Tiene alguna observación adicional sobre la prueba?,* las personas sin LV sugirieron incluir imágenes, definir el formato de prueba dependiendo de lo que se le facilitara más a cada quién (ver o escuchar), revisar el aplicativo ya que presentó algunas fallas técnicas, y, en el caso de la versión auditiva, describir cada uno de los textos, es decir, dar a conocer la cantidad de párrafos y de palabras que los componen. Las personas con LV sugirieron insertar en el aplicativo la opción de graduar la velocidad de la lectura, utilizar este mecanismo para evaluar a la población con LV en todos los campos, modificar u omitir la lectura de los enlaces de los que se sacaban algunos textos porque a veces generaba confusión, utilizar una voz

femenina para las lecturas, diseñar un mecanismo similar para evaluar áreas como matemáticas e idiomas, y para confirmar el envío de cada respuesta, repetir no solo la opción elegida, sino también la pregunta.

Respecto a la relación entre el nivel de habilidad de los evaluados y esta variable de percepción de validez, en la tabla 12 se presentan los estadísticos obtenidos para comparar las estimaciones de habilidad de las personas con y sin LV por categoría de respuesta de los ítems del Instrumento PV (H de Kruskal Wallis y U de Mann Whitney), con excepción de aquellos ítems en los que alguna categoría solo fue escogida por un participante.

Tabla 12

Resultados de la comparación de las estimaciones de habilidad de las personas con y sin LV de los ítems del Instrumento PV con categorías de respuesta

Ítem	Sin LV		Con LV	
	Estadístico	Valor p	Estadístico	Valor p
1	H = 3,6	0,16	H = 0,4	0,94
3	U = 23	0,36	U = 40	0,43
4	U = 24	0,61	U = 49	0,68
5	H = 2,9	0,23	--	--
6	H = 0,7	0,70	H = 5,8	0,12
7	H = 6,2	0,04*	--	--
8	U = 8,3	0,02*	--	--

* Diferencias significativas

En esta tabla se puede observar que en el grupo de personas con LV no se encontraron diferencias significativas entre las categorías de respuesta, y en el grupo de personas sin LV solo se encontraron en los ítems 7 y 8. En el caso del ítem 7, que hacía referencia a si se consideraba el TAI más fácil, más difícil o igual de fácil o difícil a la prueba de lápiz y papel tradicional, las personas que obtuvieron estimaciones de habilidad más altas fueron aquellas que consideraron que el TAI resultaba más fácil y las personas que obtuvieron las estimaciones más bajas fueron aquellas que consideraron que el TAI era más difícil. En el caso del ítem 8, que indagaba por la preferencia de los participantes de ser evaluados con el TAI o con la prueba tradicional de lápiz y papel, las personas que prefirieron ser evaluadas mediante el TAI presentaron los niveles de habilidad más altos, y las personas que prefirieron ser evaluadas mediante pruebas de lápiz y papel, los niveles de habilidad más bajos.

Evidencia de validez asociada con procesos de respuesta

Las variables abordadas para explorar los procesos de respuesta de las personas con y sin LV fueron tiempo por ítem y número de reproducciones. En cuanto a la variable ***tiempo***, se encontró que las personas con LV se tardaron entre 12 y 976 segundos en dar respuesta a cada ítem (media = 176 s; desviación estándar = 116 s), mientras que las personas sin LV se tardaron entre 5 y 854 segundos (media = 140 s; desviación estándar = 79 s). Estas diferencias resultaron significativas de acuerdo con el estadístico U de Mann Whitney (U = 1 418 647 p = 0,00), siendo las personas con LV las que tardaron más tiempo. Adicionalmente, al estimar este estadístico para determinar si existían diferencias significativas entre las personas de cada grupo —con y sin LV— que respondieron correcta e incorrectamente cada ítem, se encontró que en el grupo de personas sin LV no hubo diferencias significativas (U = 626 786; p = 0,34), mientras que en el grupo de personas con LV sí las hubo (U = 144 119; p = 0,02), siendo las personas que respondieron correctamente las que se tardaron más en abordar los ítems. Finalmente, al comparar el tiempo que tardaron en responder cada ítem las personas sin LV que presentaron la versión visual del TAI y la versión auditiva, se encontró que las personas que presentaron la versión auditiva se tardaron más que las que presentaron la visual (U = 519 453; p = 0,00).

En relación con la variable ***número de reproducciones***, registrada solo para las personas que presentaron la versión auditiva de la prueba, se encontró que el grupo de personas sin LV reprodujo los audios de los ítems entre 7 y 35 veces (media = 11,90; desviación estándar = 4,74), mientras que el grupo de personas con LV los reprodujo entre 7 y 46 veces (media = 13,54; desviación estándar = 6,21). Estas diferencias resultaron significativas de acuerdo con el estadístico U de Mann Whitney (U = 56 560; p = 0,00), siendo las personas con LV las que reprodujeron los audios más veces. Finalmente, al estimar este estadístico para determinar si existían diferencias significativas entre las personas de cada grupo —con y sin LV— que respondieron correcta e incorrectamente cada ítem, se encontró que en ninguno de los grupos hubo diferencias significativas (para el grupo de personas sin LV se obtuvo U = 10 596; p = 0,12, y para el grupo de personas con LV, U = 11 074; p = 0,35).

Discusión y conclusiones

El objetivo principal de este estudio fue brindar evidencia de validez de un TAI que evalúa comprensión de lectura en personas con y sin LV, con el fin de aportar elementos que sirvieran para proponer un modelo general de validación de este tipo de instrumentos de evaluación. Con este objetivo se buscaba, por un lado, recopilar evidencias de validez del TAI que evaluaba un constructo transversal a la mayoría de mecanismos de evaluación psicológica y con el que se pretendía obtener medidas igual de precisas para los dos grupos evaluados, y, por el otro, darles algunos insumos a los investigadores para desarrollar algún modelo de validación de TAI dada la ausencia de lineamientos para llevar a cabo este proceso. De allí que los objetivos específicos buscaran brindar evidencia de la estructura interna del TAI, de su relación con otras variables y de los procesos de respuesta de las personas con y sin LV evaluadas mediante este instrumento, y, a partir de la información recopilada de las dos últimas fuentes de validez nombradas, proponer una relación entre variables que describiera el desempeño de cada grupo.

Las evidencias de validez recolectadas a través de las fuentes señaladas demuestran que el uso de los TAI tiene importantes implicaciones en la evaluación psicológica al brindar medidas más precisas, lo cual, en el caso de personas con LV resulta particularmente significativo ya que los mecanismos que suelen utilizarse para evaluar a estas personas introducen numerosas fuentes de varianza irrelevante en su evaluación que no les permiten mostrar su nivel real de atributo (Hansen et ál., 2002). Adicionalmente, las estrategias diseñadas para recopilar estas evidencias de validez, como la aplicación de los instrumentos FI y PV y el registro de ciertas variables, representan un primer paso en el desarrollo de mecanismos de recolección de este tipo de evidencias que puedan ser aplicables a cualquier proceso de validación de TAI, tal como se muestra en los resultados de este trabajo.

En relación con las evidencias de validez basadas en la estructura interna del TAI, los resultados de la estimación de la Función de Información dan cuenta de que, como se mencionó, esta herramienta provee medidas precisas, sobre todo, en niveles medios de atributo. Este hallazgo es muy importante ya que demuestra que el banco de ítems del TAI, al contar con una alta confiabilidad, además de satisfacer una de las condiciones

necesarias para que las inferencias que se realicen a partir de las puntuaciones obtenidas mediante esta herramienta resulten válidas, cumple con los requisitos postulados por autores como Olea y Ponsoda (2004) de contar con ítems que se encuentren distribuidos a lo largo de todo el continuo de habilidad y que brinden una gran cantidad de información sobre el nivel de atributo de los evaluados.

En cuanto a la evaluación de la dimensionalidad de la prueba, los análisis factoriales realizados demuestran que, a pesar de haber construido el banco del TAI siguiendo las especificaciones definidas por el Icfes para la subprueba de Lenguaje de Saber 11.º, la estructura resultante no se deriva de este modelo teórico, ni de otros modelos explorados. Por otra parte, el pequeño porcentaje de varianza explicada por el primer factor en todas las formas de prueba sugiere que el banco de ítems del TAI está evaluando más de un constructo. Estos resultados son consistentes con lo encontrado en el estudio desarrollado por Herrera, Barajas, Casas, Valbuena y Jiménez (2015) en el que se realizaron análisis factoriales para las dos formas de la subprueba de Lenguaje de Saber 11.º aplicadas en 2013-II y se obtuvieron porcentajes de varianza explicada similares a los hallados en esta investigación.

En lo referente a la evidencia de validez basada en la relación con otras variables, se encontró que las variables de sexo, lugar de procedencia, estrato socioeconómico, familiaridad informática y tipo de LV (en caso de presentarla) no estaban relacionadas con el desempeño de los grupos; mientras que tener una mayor edad resultó estar asociada con un mejor desempeño en el grupo de personas con LV. Este primer hallazgo encuentra sustento en que al llegar a la edad adulta las personas con LV logran automatizar ciertas habilidades necesarias para la comprensión de lectura y, esto, a su vez, les permite tener un mejor rendimiento en este tipo de tareas (González y Pérez, 2006).

Por su parte, la ausencia de relación entre las variables de familiaridad informática y estimación de habilidad da cuenta de que los criterios de usabilidad y el entrenamiento diseñado para capacitar a los evaluados en la forma de interactuar con el TAI lograron neutralizar los posibles efectos de la primera variable sobre la segunda (ver Green, 1988 y Taylor et ál., 1999). Esta idea encuentra sustento en las opiniones de los evaluados en el Instrumento PV acerca de las facilidades que les brindaba el mecanismo de aplicación del TAI, puesto que, según ellos, las instrucciones para interactuar con esta herramienta eran

fáciles de entender y de llevar a cabo ya que eran pocos los comandos que se debían utilizar. Cabe anotar que de acuerdo con Taylor et ál. (1999) otra posible explicación de esta ausencia de relación entre estas dos variables puede ser el haber usado un formato de ítems sencillo para conformar el TAI.

Finalmente, en relación con la variable de percepción de validez se encontró que, si bien las personas de los dos grupos manifestaron sentir que fueron bien evaluados y que sintieron que efectivamente mediante el TAI se les estaba evaluando comprensión de lectura, el grupo que manifestó que el TAI resultaba particularmente ventajoso para ellos fue el de personas con LV; las razones de ello se resumen en que el TAI les brindaba independencia para responder la prueba, les hacía sentir que tenían el control sobre la situación de evaluación y les permitía resolver la prueba en un menor tiempo que con la forma de evaluación tradicional. Estas consideraciones van de la mano con las ventajas que para Hansen et ál. (2002), Douglas et ál. (2001) y Edmonds y Pring (2006) representan el uso de herramientas informáticas en la evaluación de personas con LV, y que para Weiss y Betz (1973), Brown y Weiss (1977) y Shermis y Lombard (1998) representan el uso de TAI como mecanismo de evaluación.

En cuanto a la evidencia de validez asociada con procesos de respuesta se encontró que las personas sin LV que presentaron la versión auditiva del TAI requerían más tiempo para responder los ítems que aquellos que presentaron la versión visual, lo cual, de acuerdo con Lorenzo (2001), es esperable debido a las características particulares del lenguaje hablado y a la poca familiaridad de las personas sin LV para acceder a la información mediante este canal de entrada. Adicionalmente, se encontró que, para dar respuesta a los ítems, las personas con LV reprodujeron los audios más veces y se tardaron más que las personas sin LV, y que las personas con LV que respondieron acertadamente los ítems tardaron aún más que aquellas que respondieron incorrectamente, lo cual podría deberse a la motivación experimentada por este grupo de personas para resolver de manera satisfactoria la prueba y obtener buenos resultados dadas las expectativas de ser evaluadas mediante un mecanismo novedoso que igualaba sus condiciones de evaluación con las de las personas sin LV.

Estos hallazgos representan un gran avance en la búsqueda de equidad en la evaluación mediante la incorporación de mecanismos evaluativos que hacen uso de la tecnología y que

desde hace décadas se postulan como una forma de superar muchos de los problemas derivados de los mecanismos de evaluación tradicionales; debido a que hasta el momento no se contaba con estudios de esta índole en los que se pusiera a prueba esta herramienta evaluativa para la evaluación de personas con y sin LV. De igual manera, los resultados de este trabajo constituyen un paso importante en el desarrollo de protocolos que dirijan la forma de evaluar a la población con LV, sabiendo qué variables irrelevantes para lo que se desea estimar pueden ser controladas mediante este mecanismo y favoreciendo la comodidad de estas personas con la evaluación.

En relación con este último aspecto, se destaca que la mayoría de resultados de este trabajo brindan elementos para pensar en la posibilidad de implementar el uso de este tipo de instrumentos en la evaluación de población con LV, ya que presenta muchos beneficios en términos de la calidad de la evaluación y la estimación de medidas más precisas, lo cual, sumado a que el tamaño de esta población es reducido, de acuerdo con Herrera, Barajas, Casas et ál. (2015), podría compensar los altos costos que implica desarrollar estrategias de evaluación como esta.

Adicionalmente, el método, los protocolos y los instrumentos diseñados para recoger evidencias de validez del TAI pueden servir de insumo para los validadores del TAI sobre cómo realizar este proceso de validación, lo cual también representa un gran avance en el campo de la medición y evaluación psicológica, y va de la mano con el interés actual por desarrollar estudios que permitan dar claridad a los validadores sobre cada una de las fuentes de evidencia de validez como los propuestos por Ríos y Wells (2014), Padilla y Benítez (2014), Sireci y Faulkner-Bond (2014), Lane (2014) y Oren et ál. (2014).

Finalmente, para estudios posteriores se recomienda contar con bancos de ítems más amplios y con una muestra más grande para calibrar sus ítems con el fin de garantizar que todos los niveles de habilidad sean estimados con una alta precisión. De igual manera, se recomienda contar con una estructura de prueba que corresponda a un modelo teórico que se encuentre debidamente operacionalizado a través del banco de ítems puesto que al no contar con esto es posible que los dominios de contenido abordados por la prueba no sean evaluados con la misma precisión o no se encuentren representados por los ítems seleccionados a través del aplicativo (Weiss, 2004). Lo anterior, con el fin de tener certeza

de qué se está evaluado lo mismo en todos los evaluados a pesar de que no les sean aplicados los mismos ítems.

En relación con este último aspecto también resulta de gran importancia el desarrollo de estudios que permitan determinar de manera clara cuáles son las dimensiones que componen el constructo de la comprensión de lectura y si cuando se habla de comprensión de lectura en personas con y sin LV se está haciendo referencia al mismo constructo, teniendo en cuenta las diferencias que estos dos grupos de personas presentan en el procesamiento de la información debido al canal mediante el cual acceden a ella.

Por otra parte, teniendo en cuenta que una de las razones que dieron los evaluados para haberse sentido tranquilos durante la aplicación del TAI era que tenían la certeza de que los resultados que obtuvieran en esta evaluación no tendrían incidencia en su vida, se recomienda realizar estudios en condiciones normales de evaluación que permitan determinar con mayor eficiencia el efecto que podría llegar a tener este aspecto sobre la evaluación mediante TAI. Así mismo, se recomienda que se realicen estudios para TAI que evalúen otros constructos.

Referencias

Abad, F. J., Olea, J., Real, E. y Ponsoda, V. (2002). Estimación de habilidad y precisión en test adaptativos informatizados y test óptimos: Un caso práctico. *Revista Electrónica de Metodología Aplicada, 7*(1), 1-20.

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

Agencia de Calidad de la Educación (2014). *Informe técnico Simce 2014*. Recuperado en enero 28, 2016 de http://archivos.agenciaeducacion.cl/InformeTecnicoSimce_2014.pdf

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M. y Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology, 44*(45), 68–82. doi: 10.1016/j.cedpsych.2016.02.002

Aiken, L. R. (2003). *Tests Psicológicos y evaluación*. México: Pearson Educación.

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

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

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

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

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

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.

Bäckman, Ö. (1999). A theoretical reading perspective on training methods for low vision patients. *Visual Impairment Research, 1*(2), 85-94.

Baker, F. B. (2001). *The Basics of Item Response Theory*. Estados Unidos: ERIC Clearinghouse on Assessment and Evaluation

Barnes, M. A. (2015). What do models of reading comprehension and its development have to contribute to a science of comprehension instruction and assessment for adolescents? En K. L. Santi y D. K. Reed (Eds.) *Improving reading comprehension of middle and high school students* (pp. 1-18). Nueva York, Estados Unidos: Springer.

Bosman, A. M. T., Gompel, M., Vervloed, M. P. J. y Van Bon, W. H. J. (2006). Low vision affects the reading process quantitatively but not qualitatively. *The Journal of Special Education, 39*(4), 208-219.

Brizuela, A. y Montero, E. (2013). Prediction of the difficulty level in a standardized reading comprehension test: Contributions from cognitive psychology and psychometrics. *Relieve, 19*(2), 1-21. doi: 10.7203/relieve.19.1.3149

Brown, J. M. y Weiss, D. J. (1977). *An Adaptive Testing Strategy for Achievement Test Batteries*. *Research Report 77-6*. Minneapolis: University of Minnesota.

Cañón, Y. Z. (2011). La baja visión en Colombia y en el mundo. *Ciencia & Tecnología para la Salud Visual y Ocular, 9*(1), 117-123.

Cella, D., Gershon, R., Lai, J. y Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research, 16*(1), 133–141. doi 10.1007/s11136-007-9204-6

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

Clauser, B. E., Harik, P. y Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement, 37*(3), 245-261.

Cromley, J. G. y Azevedo, R. (2007). Testing and refining the Direct and Inferential Mediation Model of reading comprehension. *Journal of Educational Psychology, 99*(2), 311-325.

Cromley, J. G., Snyder-Hogan, L. E. y Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: A domain-specific test of the Direct and Inferential Mediation Model of reading comprehension. *Journal of Educational Psychology, 102*(3), 687–700.

Cronbach, L. J. (1971). Test validation. En R. L. Thorndike (Ed.), *Educational Measurement* (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. En H. Wainer y H. Braun, (Eds.), *Test Validity* (pp. 3-17). Nueva Jersey: Lawrence Erlbaum Associates, Publishers.

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

Douglas, G., Kellami, E., Long, R. y 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. doi: 10.1177/026461960101900105

Dreyer, L. G. y Katz, L. (1992). An examination of «The simple view of reading». En C. K. Kinzer y D. J. Leu (Eds.), *Literacy research, theory, and practice: Views from many perspectives, 41st Yearbook of the National Reading Conference,* Chicago, Estados Unidos: National Reading Conference.

Edmonds, C. J. y Pring, L. (2006). Generating inferences from written and spoken language: A comparison of children with visual impairment and children with sight. *British Journal of Developmental Psychology, 24*(2), 337-351.

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.

Embretson, S. E., y Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension. *Applied* Psychological *Measurement, 11*(2), 175-193.

Espinosa, A. M. (2013). *Evaluación Objetiva de los Procesos Cognitivos involucrados en la Comprensión de Lectura* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

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

Gómez, I., Vila, J. O., García, J. A., Contreras, A. y Elosúa, M. R. (2013). Comprensión lectora y procesos ejecutivos de la memoria operativa. *Psicología Educativa, 19*(2), 103-111. doi: 10.5093/ed2013a17

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

Gómez, J. e Hidalgo, M. D. (2005). La validez de los test, escalas y cuestionarios. *La Sociología en sus Escenarios,* (12), 1-14.

Gómez, J. C. y González, C. I. (2008). *Discapacidad en Colombia: Reto para la inclusión en capital humano.* Bogotá: Fundación Saldarriaga Concha. Recuperado en abril 1, 2017 de http://www.saldarriagaconcha.org/images/fsc/pdf/biblioteca_virtual/discapacidad/estadisticas_e_investigaciones/05_tomo_1_exclusion_DEPTAL_CH.pdf

Gompel, M., Van Bon, W. H. J. y Schreuder, R. (2004). Reading by children with low vision. *Journal of Visual Impairment y Blindless, 98*(2), 77-89.

González, L. (2004). Text comprehension by blind people using speech synthesis systems. En K. Miesenberger, J. Klaus, W. Zagler y D. Burger (eds.), *Computers Helping People with Special Needs* (pp 538-544). París: Springer Science y Business Media.

González, L. y Pérez, M. (2006). Comprensión de textos y modalidades de acceso a la información: Comparación de rendimientos entre personas ciegas y videntes. *Integración: Revista sobre Ceguera y Deficiencia Visual*, (48), 7-24.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*(4), 351-373.

Gough, P. B., y Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1) 6-10.

Green, B. F. (1988). Construct validity of computer-based test. En H. Wainer y H. Braun, (Eds.), *Test Validity* (pp. 77-86). Nueva Jersey: Lawrence Erlbaum Associates, Publishers.

Guion, R. M. (1977). Content validity: Three years of talk-what's the action? *Public Personnel Management, 6*(6), 407-414.

Hansen, E. G., Lee, M. J., y Forer, D. C. (2002). A «self-voicing» test for individuals with visual impairments. *Journal of Visual Impairment y Blindness, 96*(4), 273-275.

Herrera, A. N. (1996). *Algunas consideraciones técnicas sobre la construcción de ítems de pruebas objetivas según la clasificación de objetivos educativos de Bloom*. Bogotá: inédito.

Herrera, A. N., Barajas, R. y Jiménez, G. J. (2015). Validez en Test Adaptativos Informatizados: Alternativa para evaluar población con limitaciones visuales. *Avaliação Psicológica, 14*(3), 299-307. doi: 10.15689/ap.2015.1403.01

Herrera, A. N., Barajas, R., Casas, M., Valbuena, D. y Jiménez, G. J. (2015). *Diseño de una estrategia integral piloto de evaluación alternativa para personas con y sin limitación visual* (Informe técnico de investigación inédito). Bogotá: Icfes.

Hoover, W. A., y Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127–160.

Hornke, L. F. (2000). Item Response Times in Computerized Adaptive Testing. *Psicológica, 21*(1), 175-189.

Huff, K. L. y Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16-25. doi: 10.1111/j.1745-3992.2001.tb00066.x

Instituto Colombiano para la Evaluación de la Educación (2012). *Examen de Estado de la educación media – Icfes Saber 11.º. Qué se evalúa*. Recuperado en enero 12, 2014 de http://www.icfes.gov.co/index.php?searchword=saber+11yoption=com_searchyItemid=307

Instituto Colombiano para la Evaluación de la Educación (2016). *Marco de referencia Módulo de Lectura crítica Saber 11.º y Saber Pro*. Obtenido en noviembre 1, 2016 de http://www2.icfes.gov.co/docman/estudiantes-y-padres-de-familia/saber-pro-estudiantes-y-padres/marcos-de-referencia/2442-marco-de-referencia-lectura-critica/file

Instituto Nacional para Ciegos (2013). *La Inclusión Social de la Niñez con Discapacidad Visual*. Bogotá: Autor.

Jiménez, V. (2004). *Metacognición y Comprensión de la Lectura: Evaluación de los componentes estratégicos (procesos y variables) mediante la elaboración de una Escala de Conciencia Lectora (ESCOLA)* (Tesis de doctorado). Universidad Complutense de Madrid, España.

Joshi, R. M. y Aaron, P. G. (2000). The component model of reading: Simple view of reading made a little more complex. *Reading Psychology, 21*(2), 85-97. doi: 10.1080/02702710050084428

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. En R. W. Lissitz, (Ed.), *The Concept of Validity: Revisions, new directions, and applications* (pp. 39-64). Estados Unidos: Information Age Publishing, INC.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163–182. doi:10.1037/0033-295X.95.2.163.

Kintsch, W. (1998). Comprehension. A paradigm for cognition. Nueva York: Cambridge.

Kintsch, W. y Van Dijk, T. A. (1978) Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127-135. doi: 10.7334/psicothema2013.258

Lee, J. A., Moreno, K. E. y Sympson, J. B. (1984). *The Effects of Mode of Test Administration on Test Performance*. Paper presented at the Annual Meeting of the Eastern Psychological Association, Maryland, Baltimore.

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

Loevinger, J. (1957). Objective test as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

López, J. y Sánchez, J. M. (2005). GenTAI: Generador de test adaptativos informatizados. *Revista Iberoamericana de Informática Educativa,* 2, 9-24.

Lorenzo, J. R. (2001). Procesos cognitivos básicos relacionados con la lectura. Primera parte: La conciencia fonológica. *Interdisciplinaria, 18*(1), 1-33.

Lozzia, G. S., Abal, F. J. P., Blum, G. D., Aguerri, M. E., Galibert, M. S. y Attorresi, H. F. (2015). Construcción de un banco de ítems de analogías verbales como base para un test adaptativo informatizado. *Revista Mexicana de Psicología, 32*(2), 134-148.

McNamara, D. S. y Magliano, J. P. (2009). Toward a comprehensive model of comprehension. En B. Ross (Ed.), *The Psychology of Learning and Motivation*. Nueva York, Estados Unidos: Elsevier Science. doi: 10.1016/S0079-7421(09)51009-2.

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

Messick, S. (1980). Test validity and ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. En H. Wainer y H. Braun, (Eds.), *Test Validity* (pp. 33-45). Nueva Jersey: Lawrence Erlbaum Associates, Publishers.

Messick, S. (1989). Validity. En R. Linn (Ed.), *Educational Measurement* (pp. 13-103). Washington, D.C.: American Council on Education.

Messick, S. (1995). Standards of validity and the validity of the standards in performance assessment. *Educational Measurement: Issues and practice, 14*(4), 5-8.

Ministerio de Educación Nacional (2006). *Estándares básicos de Competencias en Lenguaje, Matemáticas, Ciencias y Ciudadanas.* Bogotá: Autor. Recuperado en enero 12, 2014 de http://www.mineducacion.gov.co/1621/articles-116042_archivo_pdf.pdf

Ministerio de Educación Nacional (2012). *Niños y niñas con discapacidad visual en Colombia*. Bogotá: Autor. Recuperado en abril 1, 2017 de http://www.mineducacion.gov.co/cvn/1665/w3-article-303293.html

Molina, J. G., Pareja, I. y Sanmartín, J. (2008). Modeling item banking: Analysis and design of a computerized system. *Revista Electrónica de Metodología Aplicada, 13*(2), 1-14.

Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., Tobin, B. y Valeri, J. (2004). Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition and Learning, 2*, 103-140.

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

Muñiz, J. (2010). Las teorías de los test: Teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo, 31*(1), 57-66.

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

Ochaíta, E., Rosa, A., Fernández, E. y Huertas, J. A. (1988). *Lectura braille y procesamiento de la información táctil*. Madrid: Inserso.

Olea, J. y Hontangas, P. M. (1999). Test informatizados de primera generación. En J. Olea, V. Ponsoda, y G. Prieto (Eds.), *Test Informatizados: Fundamentos y aplicaciones* (pp. 111-125). Madrid: Pirámide.

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

Olea, J., Abad, F. J., Ponsoda, V., y Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento del inglés escrito: Diseño y comprobaciones psicométricas. *Psicothema, 16*(3), 519-525.

Omar, R. y Mohammed, Z. (2005). Relationship between vision and reading performance among low vision students. *International Congress Series, 1282*, 679-683. doi:10.1016/j.ics.2005.05.063

O'Reilly, T., Weeks, J., Sabatini, J., Halderman, L. y Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review, 26*(3), 403–424.

Oren, C., Kennet-Cohen, T., Turvall, E., y Allalouf, A. (2014). Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel. *Psicothema, 26*(1), 117-126. doi: 10.7334/psicothema2013.257

Padilla, J. L., y Benítez, L. (2014). Validity evidence based on response processes. Psicothema, 26(1), 136-144. doi: 10.7334/psicothema2013.259

Papanastasiou, E. C. y Reckase, M. D. (2007). A «rearrangement procedure» for scoring adaptive test with review options. *International Journal of Testing, 7*(4), 387-407.

Pérez, J. A., Chacón, S. y Moreno, R. (2000). Validez de constructo: El uso de análisis factorial exploratorio-confirmatorio para obtener evidencias de validez. *Psicothema 12*(2), 442-446.

Pérez, M. J. (2005). Evaluación de la comprensión lectora: Dificultades y limitaciones. *Revista de Educación,* (núm. extraordinario), 121-138.

Perfetti, C. y Adolf, S. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. En J. Sabatini, E. Albro, y T. O'Reilly (eds.), *Measuring up: Advances in how to assess reading ability.* Lanham, MD: Rowman y Littlefield.

Popham, W. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Rapp, D. N. y Van den Broek, P. (2005). Dynamic text comprehension: An integrative view of reading. *Current Directions in Psychological Science, 14*(5), 276-279.

Revuelta, J., Ponsoda, V. y Olea, J. (1998). Métodos para el control de las tasas de exposición en test adaptativos informatizados. *Relieve, 4*(2), 1-8-

Rios, J., y Wells, C (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116. doi: 10.7334/psicothema2013.260

Rubio, V., y Santacreu, J. (2004). *TRAS-I: Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción*. Madrid: TEA

Şahina, A. y Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in Computerized Adaptive Testing. *Educational Sciences: Theory y Practice, 15*(6), 1585-1595. doi: 10.12738/estp.2015.6.0102

Sarriá, A., Guàrdia, J. y Freixa, M. (1999). *Introducción a la Estadística en Psicología*. Barcelona: Universitat de Barcelona.

Seigneuric, A. y Ehrlich, M. F. (2005). Contribution of working memory capacity to children's reading comprehension: A longitudinal investigation. *Reading and Writing, 18*, 617-656. doi: 10.1007/s11145-005-2038-0

Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice, 23*(2), 268-280. doi: 10.1080/0969594X.2016.1141168

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

Singleton, C. H. (2008). Visual factors in reading. *Educational and child psychology, 25*(3), 8–20.

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

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23*(2), 226-235. doi: 10.1080/0969594X.2015.1072084

Sireci, S. G., y Faulkner-Bond. M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100-107. doi: 10.7334/psicothema2013.256

Sireci, S. G. y Padilla, J. L. (2014). Validating assessments: Introduction to the Special Section. *Psicothema, 26*(1), 97-99. doi: 10.7334/psicothema2013.255

Snow, C. E. (2002). Defining comprehension. En C. E. Snow (1 ed.), *Reading for Understanding: Toward an RyD program in reading comprehension* (pp. 11-18). California, Estados Unidos: RAND.

Soler, M. P. (2013). *Evaluación de la Comprensión de Lectura en personas con limitación visual* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Stinnett, M. (2009). Research in reading. *Illinois Reading Council Journal, 37*(2), 59-64.

Stone, E., y Davey, T. (2011). *Computer- Adaptive Testing for students with disabilities: A review of the literature. Research Report ETS RR-11-32*. Recuperado en febrero 20, 2014 de http://www.ets.org/Media/Research/pdf/RR-11-32.pdf

Taylor, C., Kirsch, I., Eignor, D. y Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219-274.

Tristan, A. y Vidal, R. (2006). *Estándares de Calidad para Pruebas Objetivas.* Bogotá: Magisterio.

Vallés, A. (2005). Comprensión lectora y procesos psicológicos. *Liberabit: Revista de Psicología,* 11, 49-61.

Van Dijk, T. A. y Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Nueva York: Academic Press.

Veispak, A., Boets, B. y Ghesquière, P. (2012). Parallel versus sequential processing in print and braille reading. *Research in Developmental Disabilities, 33*(6), 2153–2163.

Wainer, H. y Braun, H. I. (1988). *Test Validity*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Wang, C. y Chang, H. (2011). Item selection in multidimensional computerized adaptive testing—gaining information from different angles. *Psychometrika, 76*(3), 363–384. doi: 10.1007/s11336-011-9215-7

Watson, G. R., Wright, V., Long, S., y De L'Aune, W. (1996). A low vision reading comprehension test. *Journal of Visual Impairment y Blindness, 90*(6), 484-494.

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

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

Woolley, G. (2011). *Reading Comprehension: Assisting children with learning difficulties*. Nueva York: Springer. doi: 10.1007/978-94-007-1174-7

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: theory and applications. *Psychometrika, 77*(3), 495–523. doi: 10.1007/s11336-012-9265-5

Apéndice A. Instrumento de familiaridad informática

UNIVERSIDAD NACIONAL DE COLOMBIA SEDE BOGOTÁ — icfes mejor saber — GRUPO DE INVESTIGACIÓN MÉTODOS E INSTRUMENTOS PARA LA INVESTIGACIÓN EN CIENCIAS DEL COMPORTAMIENTO, DEPARTAMENTO DE PSICOLOGÍA

Instrumento FI-S

Número de identificación

Realice las siguientes preguntas y marque con una equis (X) según corresponda.

¿Tiene acceso a un computador? SI ..... NO

¿En qué lugar(es)?

Ej.: Casa, universidad, trabajo.

¿Lo usa? SI ..... NO

¿Con qué frecuencia semanal? Horas. ¿Por qué?

¿Qué actividades realiza cuando lo usa?

¿Cuándo necesita usarlo, requiere ayuda de un tercero? SI ..... NO

¿Qué tipo de ayuda?

¿Tiene acceso a INTERNET? SI ..... NO

¿En qué lugar(es)?

Ej.: Casa, universidad, trabajo.

¿Lo usa? SI ..... NO

¿Con qué frecuencia semanal? Horas. ¿Por qué?

¿Qué actividades realiza cuando lo usa?

¿Cuándo necesita usarlo, requiere ayuda de un tercero? SI ..... NO

¿Qué tipo de ayuda?

¿Está familiarizado con algún software particular? SI ..... NO

¿Cuál?

¿Para qué lo utiliza?

Instrumento FI-O

Califique de acuerdo con los siguientes criterios:

0	No realiza la acción.
1	Realiza la acción con ayuda del evaluador.
2	Realiza la acción de manera autónoma.

	INSTRUCCIÓN	PUNT.			OBSERVACIONES
1	Prenda el computador. *(Si ya está prendido, con que la persona lo simule, basta.)*	0	1	2	
2	*(Abra Word y pídale que realice la siguiente acción)* Escriba su nombre completo.	0	1	2	
3	Ingrese a una página de internet.	0	1	2	
4	Busque una noticia reciente.	0	1	2	
5	*(En el Word abierto pídale que realice la siguiente acción)* Copie y pegue la noticia.	0	1	2	

Carrera 30 No.45-03, **EDIFICIO AULAS DE CIENCIAS HUMANAS**, 2° piso Oficina 227
Teléfono: (57-1) 316 5000 Ext. 16378 - 16347 - 16311
Correo electrónico: labpsico_fchbog@unal.edu.co
Bogotá, Colombia, Sur América

Apéndice B. Instrumento de percepción de validez

UNIVERSIDAD NACIONAL DE COLOMBIA SEDE BOGOTÁ

icfes mejor saber

GRUPO DE INVESTIGACIÓN MÉTODOS E INSTRUMENTOS PARA LA INVESTIGACIÓN EN CIENCIAS DEL COMPORTAMIENTO
DEPARTAMENTO DE PSICOLOGÍA

Instrumento PV

Marcar con una equis (x) según corresponda.

	PREGUNTA			¿POR QUÉ?
1	Siente que le fue	MEJOR PEOR IGUAL	en esta prueba que en la prueba de Lenguaje que presentó en Saber 11.º	
2	Siente que fue	BIEN MAL —NO SABE—	evaluado con esta prueba.	
3	Se sintió	CÓMODO INCÓMODO	presentando esta prueba.	
4	Se sintió	NERVIOSO TRANQUILO	presentando esta prueba.	

En comparación con la prueba que presentó en formato de lápiz y papel, considera que esta prueba computarizada:

1	Tiene	MÁS MENOS IGUAL	probabilidad de presentar errores en la calificación.	
2	Al tener menos ítems es	MÁS MENOS IGUAL de	precisa en la calificación.	
3	Es	MÁS FÁCIL. MÁS DIFÍCIL. IGUAL	de fácil o difícil.	

	PREGUNTA		¿POR QUÉ?
1	¿Prefiere ser evaluado mediante este mecanismo que mediante la prueba tradicional de lápiz y papel?	SI NO	

Preguntar lo siguiente de acuerdo con la versión de prueba que haya presentado:

VERSIÓN AUDITIVA	VERSIÓN VISUAL
¿La voz del aplicativo fue clara, presentaba una entonación apropiada? ¿Tuvo inconvenientes relacionados con la calidad del sonido?	¿La letra del aplicativo fue clara, presentaba tamaño y color apropiados? ¿Tuvo inconvenientes relacionados con la calidad de la imagen?
Respuesta	Respuesta

¿Considera que esta prueba tiene aspectos positivos? / *Si responde afirmativamente, preguntar* ¿Cuáles?

¿Tiene alguna observación adicional sobre la prueba?

Carrera 30 No.45-03, **EDIFICIO AULAS DE CIENCIAS HUMANAS**, 2° piso Oficina 227
Teléfono: (57-1) 316 5000 Ext. 16378 - 16347 - 16311
Correo electrónico: labpsico_fchbog@unal.edu.co
Bogotá, Colombia, Sur América