EVALUACIÓN EMPÍRICA DE DOS MÉTODOS DE EQUIPARACIÓN:
BUSCANDO EQUIDAD EN LA EVALUACIÓN DE PERSONAS CON LIMITACIÓN VISUAL

DIANA RODRÍGUEZ VALBUENA

Universidad Nacional de Colombia
Facultad de Ciencias Humanas
Departamento de Psicología
Maestría en Psicología
Bogotá, Colombia
2016

EVALUACIÓN EMPÍRICA DE DOS MÉTODOS DE EQUIPARACIÓN:

BUSCANDO EQUIDAD EN LA EVALUACIÓN DE PERSONAS CON LIMITACIÓN VISUAL

DIANA RODRÍGUEZ VALBUENA
Código: 2661106

Tesis de grado para optar al título de: Magister en Psicología

Directora: Ph.D., Aura Nidia Herrera Rojas

Línea de Investigación: Métodos e Instrumentos para la investigación en Ciencias del Comportamiento

Universidad Nacional de Colombia
Maestría en Psicología
Bogotá, Colombia
2016

Agradecimientos

A mi familia, por serlo todo.

A la Universidad Nacional de Colombia por dejarme vivir y aprender durante 12 años, gracias por enseñarme a ser la profesional que quiero ser.

Al Instituto Colombiano para la evaluación de la educación-ICFES y a la División de Investigación de la sede Bogotá de la Universidad Nacional– DIB, por respaldar esta investigación por medio de la Convocatoria para grupos de investigación en 2013 y la convocatoria del programa nacional de proyectos para el fortalecimiento de la investigación, la creación y la innovación en posgrados 2013-2015.

A los integrantes (los de antes, los de ahora y los de siempre) del grupo de investigación Métodos e Instrumentos para la investigación en Ciencias del Comportamiento por dedicar parte de su tiempo a escuchar el proyecto y apoyarme con sus sugerencias para que este trabajo fuese lo mejor posible.

A los amigos que no se cansaron de preguntar "¿Cómo va la tesis?"

A la profesora Aura Nidia Herrera por presentarme el tema y por enseñarme lo que significa ser investigador.

Al profesor Vicente Ponsoda por aceptar revisar mi trabajo y darme la oportunidad de conocer otras formas de enseñanza en la Universidad Autónoma de Madrid.

TABLA DE CONTENIDO

RESUMEN

El proceso de comprensión de lectura es diferente entre las personas con y sin limitación visual debido en parte a que los estudiantes con limitación visual necesitan un mayor tiempo para decodificar y reconocer las palabras cuando están leyendo. Este tipo de diferencias afecta la medición de dicho atributo en pruebas de aplicación masiva, especialmente en programas de evaluación que producen grandes cantidades de pruebas para evaluar muchas personas al mismo tiempo.

Este estudio tuvo como objetivo comparar dos métodos de equiparación utilizando las bases de datos de las puntuaciones en la prueba de lenguaje del examen de estado SABER 11 presentada por estudiantes con y sin limitación visual en el segundo semestre del año 2013. En total se analizaron las respuestas de 518453 estudiantes sin limitación visual y 104 con limitación Debido a las diferencias en el procesamiento de la información entre las dos poblaciones de estudiantes, se considera que las pruebas no son las mismas para los dos grupos, a pesar de que el contenido de los ítems sea el mismo para todos los estudiantes. Así pues, el uso de métodos de equiparación de puntuaciones se presenta como una alternativa para hacer comparables los resultados de las dos poblaciones. En el presente estudio se buscó evaluar el funcionamiento de los métodos Tucker y Media-sigma para grupos no equivalentes con test de anclaje.

Para la evaluación de los métodos se implementó el análisis del error de equiparación de los métodos, así como una comparación del funcionamiento de los mismos cuando las pruebas son más cortas, para esto se eliminaron de la calificación de la prueba de cada grupo los ítems que presentan funcionamiento diferencial..

Los resultados señalaron que el método Media-sigma presenta menor error de equiparación que el método lineal, así como que el error de equiparación es menor para ambos métodos cuando los ítems que presentan funcionamiento diferencial no se incluyen en la calificación.

Palabras clave: Equiparación, Comprensión de lectura, Tucker, Media-sigma, Funcionamiento Diferencial de los Ítems.

ABSTRACT

Title: Empirical assessment of two equating methods. Searching equity in the assessment of students with visual impairment.

Reading comprehension process has some different and relevant aspects between students with and without visual impairment. For instance blind students need more time to decode and recognise words when they are reading. Those differences make an impact in the reading comprehension assessment for those two groups, especially in testing programs that administer various test forms in a massive amount of students at the same time.

Scores of 518453 students without visual impairment y 104 students with visual impairment were used in this study, which main goal is to compare two equating methods using the scores the students in the reading comprehension proficiency exam from the SABER 11 test used in the second semester of 2013.

However, because the differences between students with visual impairment and students without visual impairment, the test seems not be the same for the both groups despite that items have the exact same content for them. So to adjust test scores and compare effectively the measurement in these two groups it's plausible to use equating methods; in this case Tucker and Mean-sigma methods with non-equivalent groups with anchor test design were used.

The assessment of the equating methods was made identifying the equating error for those methods and includes a comparison between the Tucker and Mean-sigma methods when the test are complete and when test were shorter cause items identified with differential functioning for one of the groups were eliminated in the total score.

Results shown that mean-sigma method presents less equating error than linear equating. Another finding in this study is that equating error decreases for both methods when items with differential functioning are left out of total test score.

Key words: Equating, Reading Comprehension, Tucker method, Mean-sigma method, Differential Item Functioning, DIF.

INTRODUCCIÓN

En Colombia los mecanismos de evaluación de la calidad de la educación incluyen la aplicación de pruebas masivas a personas que se encuentran último grado de educación media, siendo ésta una exigencia para ingresar a la educación superior. Dado que el primer requisito para presentar la prueba es tener un grado de escolaridad determinado, cabría esperarse que las posibles diferencias en el desempeño de los evaluados sean atribuibles únicamente a sus niveles de habilidad o desempeño; sin embargo, se debe tener en cuenta que estas personas tienen diferentes condiciones económicas y sociales e importantes diferencias individuales, entre ellas diversas características físicas, como la capacidad visual.

El plan decenal de educación 2006-2016 plantea que el Estado Colombiano debe responder por la equidad en la educación, es decir, debe garantizar el derecho y el acceso a un sistema educativo público adecuado y de calidad; además señala la necesidad de garantizar los apoyos pedagógicos y tecnológicos para minimizar las barreras en el aprendizaje de la población vulnerable o con necesidades educativas especiales, ya sea por discapacidad o por el talento que posee la persona.

La entidad encargada de la realización y aplicación de esta evaluación es el Instituto Colombiano para la evaluación de la educación – ICFES que, según el artículo 5 de la Resolución 092 de 2008 debe organizar y brindar las condiciones especiales para la población discapacitada, siempre y cuando se suministre por su parte la información requerida para este fin. Además, el ICFES debe autorizar el ingreso de apoyo necesario de acuerdo con la información previamente suministrada por el usuario sobre su discapacidad.

En el caso de las personas invidentes o con baja visión, el apoyo que brinda el ICFES para la presentación de la prueba SABER es el acompañamiento de un lector, previamente capacitado por el Instituto Nacional para Ciegos – INCI, cuya función principal es leer las preguntas del cuadernillo y registrar la opción elegida por el evaluado en la hoja de respuestas. Aunque para el momento de la evaluación se busca garantizar que se presenten las condiciones necesarias para que los evaluados con discapacidad no tengan desventaja, es importante considerar si el presentar la misma prueba por medios diferentes (leer por los

propios medios o escuchar la lectura de un tercero, en el caso de las personas con limitación visual) es equitativo o al menos equivalente. Es necesario entonces plantear preguntas importantes en torno a la evaluación, considerando las diferencias individuales, por ejemplo: ¿La presentación de pruebas homogéneas por parte de todos los subgrupos de personas, garantizan una medida invariante entre los grupos? ¿Se deben aplicar pruebas diferentes a grupo distintos para evaluar el mismo constructo?

En investigaciones recientes (Herrera, Soler, Espinosa, Lancheros, y Jiménez, 2012), se realizaron análisis de los puntajes obtenidos en las pruebas SABER 11 de comprensión de lectura por personas con y sin limitación visual. Los resultados evidenciaron que las personas con limitación visual se ubicaban en los niveles bajos de habilidad y que algunos de los ítems que componen la prueba presentan funcionamiento diferencial (DIF - por las iniciales de Differential Item Functioning) al comparar los resultados de los dos grupos; dichas diferencias cuestionan la equivalencia entre las pruebas aplicadas a diferentes poblaciones, ya que, según Santana (2009), se asume que un ítem presenta funcionamiento diferencial si las funciones de respuesta a dicho reactivo no son iguales entre los grupos.

Una de las razones que podrían dar cuenta de la presencia del funcionamiento diferencial en los ítems que componen la prueba de comprensión de lectura, en los grupos de personas con y sin limitación visual, es la modalidad sensorial por la cual las personas acceden a la información. Mientras las personas sin limitación pueden acceder al texto de manera visual directa, las personas invidentes deben acceder al texto de forma auditiva o táctil; en el caso de la prueba SABER 11, las personas con limitación visual acceden a la información de la prueba por el canal auditivo. Como lo señalan diferentes autores (González (2004), Soler (2013)), la comprensión de lectura por parte de las personas invidentes requiere un mayor uso de memoria de trabajo y concentración en comparación con las personas sin limitación visual, ya que para los primeros los estímulos requieren un mayor nivel de discriminación para eliminar lo que no sirve, utilizar lo que sí e identificar esa información de forma completa. De lo anterior se puede desprender que al realizar la evaluación con el mismo instrumento en estos dos grupos se están desconociendo otras variables que pueden estar interviniendo y que afectan directamente la precisión de la medición para uno de los grupos y la comparabilidad de los resultados de ambos grupos.

Teniendo en cuenta estas circunstancias Lancheros (2013) consideró aplicar los métodos de equiparación de puntuaciones con el fin de hacer comparables las calificaciones de los dos grupos; para esto realizó simulaciones de distribuciones de puntajes de los grupos de personas con y sin limitación visual, para realizar una evaluación del funcionamiento de tres procedimientos de equiparación diferentes (Método lineal, Tucker y Media/sigma). El objetivo de este tipo de procedimientos es hacer intercambiables los puntajes de distintas pruebas, o los resultados de la misma prueba en diferentes grupos, con el fin de facilitar las comparaciones necesarias que permitan determinar con precisión el nivel de atributo independientemente de la prueba presentada o, en este caso, la limitación visual de los estudiantes.

Como resultado de esta investigación se encontró que dos de los tres métodos evaluados presentan resultados satisfactorios en la situación planteada, así el método de Media/sigma (Marco (1977), Kolen y Brennan (2014)) presenta menor error de equiparación, mientras el método Tucker (Angoff y Modu (1973), Dorans (2010)) mostró mayor estabilidad entre replicas (Herrera et al (2012) y Lancheros (2013)). Se debe subrayar que con los datos obtenidos en las diferentes aplicaciones de prueba, el ICFES realiza equiparaciones utilizando métodos basados en la teoría de respuesta al ítem bajo el modelo de Rasch, sin embargo estas ecuaciones de equiparación se utilizan entre aplicaciones, es decir, se realizan con base en los resultados de cada uno de los grupos que tomaron la prueba en el año para evaluar el rendimiento (se compara el resultado del primer semestre y el del segundo semestre, por ejemplo), más no se realiza la equiparación entre grupos dentro de la misma aplicación (Lancheros, 2013). Es importante evaluar el comportamiento de los procedimientos de equiparación en este tipo de aplicaciones (Pruebas SABER entre grupos de la misma aplicación) porque al tener múltiples formas de prueba para la evaluación de la educación en distintas aplicaciones y en diversos grupos, se debe aprovechar o al menos evaluar la pertinencia de algunos procedimientos estadísticos que pueden dar cuenta de las diferencias entre formas de prueba y entre grupos, ya que como señala Holland (2007) “la equiparación de puntuaciones es necesaria en cualquier programa de evaluación que produzca diferentes formas de prueba continuamente” (p.22). De igual forma, a nivel experimental es necesario dar cuenta del funcionamiento de los procedimientos de

equiparación en diferentes tamaños de muestra y con distintos tipos de test de anclaje, de los usados hasta el momento (Suh, Mroch, Kane y Ripkey, 2009). Responder a este tipo de cuestiones no es simplemente un problema técnico o de método, ya que como señala Mislevy (1992) en su reporte sobre el enlace de puntuaciones en el ámbito educativo, las cuestiones técnicas no surgen ni pueden ser contestadas sin dirigirse directamente a los propósitos y consecuencias de la evaluación.

Estableciendo como punto de partida para el presente estudio los avances realizados en los trabajos previos relacionados con la evaluación de comprensión lectora en pruebas de aplicación masiva, los resultados encontrados con los métodos de equiparación con datos simulados, y considerando que los procedimientos de equiparación en los puntajes obtenidos plantean variados desafíos metodológicos, desde la elección adecuada del procedimiento teniendo en cuenta las diferencias en los tamaños de las muestras y las distribuciones de los datos, hasta las decisiones sobre la inclusión en la calificación de los ítems que presentan funcionamiento diferencial; el objetivo principal de esta investigación es evaluar el funcionamiento de los métodos de equiparación Tucker y Media-sigma con los datos empíricos en pruebas masivas con el propósito de establecer si, en el caso de la prueba de comprensión lectora del examen SABER 11, los procedimientos de equiparación Tucker y Media-sigma podrían ser utilizados como herramienta para garantizar equivalencia en la evaluación entre estudiantes con y sin limitación visual.

Objetivos específicos

1. Establecer si existen diferencias en el desempeño de los estudiantes con y sin discapacidad visual en las pruebas de lectura SABER 11 – 2013, que no puedan ser atribuidas únicamente al nivel de atributo de los evaluados.
2. Evaluar los métodos de equiparación, comparando la medida de error de los procedimientos realizados con datos empíricos y los realizados con datos simulados.

REVISIÓN BIBLIOGRÁFICA

El Instituto Colombiano para la evaluación de la educación – ICFES - tiene como objeto "realizar la evaluación del sistema educativo colombiano y propender por la calidad de dicho sistema, efecto para el cual […] lleva a cabo la aplicación de diferentes instrumentos de evaluación educativa en el país" (Resolución 092 de 2008). El decreto 869 de 2010, especifica que el examen de la educación media ICFES-SABER 11 tiene como objetivos principales comprobar el grado de desarrollo de las competencias de los estudiantes que están por finalizar la educación media (grado 11) y suministrar al estudiante elementos para que pueda realizar una autoevaluación para el desarrollo de su proyecto de vida.

Este último es muy importante porque pone de presente el impacto que pueden tener los resultados obtenidos sobre las decisiones que tomen los estudiantes que aplican a la prueba, lo que plantea la importancia de garantizar que la invarianza en la medición se mantenga en las distintas aplicaciones y en las diferentes formas de prueba y, sobretodo, en lo relativo a la evaluación de la población con necesidades especiales.

Durante mucho tiempo se ha reclamado que el sistema educativo debe ser más justo, por ejemplo Moliner (2008) afirma que si el sistema fuese más inclusivo, se podrían cambiar las actitudes hacia la diferencia, logrando disminuir los procesos de exclusión. En Colombia, se ha establecido que para alcanzar cambios que respondan a las demandas de la sociedad "le corresponde al Estado adoptar políticas sistemáticas que afecten todos los componentes de un sistema educativo, tales como la formación de los profesores, la financiación, la dirección y supervisión escolar, entre otros" (Corte Constitucional, 2012).

Por esto el ICFES ha buscado desarrollar nuevas estrategias para que la población discapacitada tenga un mayor y mejor acceso a los exámenes; un ejemplo de esto es la generación del examen electrónico para las personas con discapacidad auditiva en el que, por medio de una plataforma multimedia, se estandariza la presentación del contenido de la prueba utilizando el lenguaje de señas colombiano. En el caso específico de la prueba de lenguaje se considera que no se deben presentar modificaciones porque al evaluar lectura "no resulta pertinente realizar una interpretación" (ICFES, 2013).

Con respecto a las medidas que se toman en la aplicación de la prueba regularmente a personas invidentes, el Instituto Nacional para Ciegos - INCI manifestó que contar con un lector para las personas con limitación visual, facilita "el acceso a la información de una manera clara, específica y con calidad, para alcanzar la comprensión de las preguntas y por ende la elección de las respuestas" (INCI, 2009, p.10). Por su parte el ICFES señaló en una comunicación citada en la Sentencia T-598/13 que se suministra apoyo de un lector para el manejo del cuadernillo y la hoja de respuestas, y en caso de ser necesario, el ingreso de materiales adicionales para las personas que presentan baja visión. A pesar de que al hacer uso de materiales adicionales para responder la prueba puede implicar mayor tiempo en la presentación de la misma, el ICFES ha manifestado que "la única consideración que al respecto tiene el ICFES con la población discapacitada es en cuanto a que los salones donde están ubicadas las personas con alguna discapacidad, son los últimos donde se recogen los cuadernillos" (comunicación ICFES citada en la Sentencia T-598/13), ya que de hacerlo de otra manera, señala el instituto, sería injusto con los demás evaluados.

Con base en estas circunstancias, el Estado colombiano resolvió, en Sentencia T-598/13, que tanto el ICFES como el Ministerio de Educación Nacional deben desarrollar políticas que permitan el acceso "a las personas que se encuentren en situación de discapacidad, teniendo en cuenta su circunstancia específica, las herramientas o apoyos que requieran para presentar en condiciones dignas y de igualdad los exámenes de Estado"; así mismo se solicita a los distintos encargados de la educación en todo el país (secretarías de educación), que se realice un seguimiento a la presentación de las pruebas "por parte de las personas en situación de discapacidad, para coadyuvar a la materialización real de su desarrollo integral, como obligación constitucional".

Esto es de gran importancia, ya que como lo consagra el artículo 23 de la ley 361 de 1997, el Servicio Nacional de Aprendizaje – SENA- debe encargarse de promover cursos para la población con diferentes tipos de limitación y garantizará su acceso a los programas de formación establecidos. Sin embargo, si la valoración y evaluación de las habilidades de las personas con algún tipo de discapacidad se ve afectada por los factores externos descritos con anterioridad (como manejo de tiempo, acceso a la información presentada,

entre otros), entonces se afectaría de manera directa el acceso a la educación superior de las personas con limitación visual, minando el derecho a la igualdad de oportunidades.

Sin embargo, las consideraciones que ha establecido el ICFES pueden quedarse cortas en algunos casos específicos, sobre todo si se tiene en cuenta el atributo a evaluar y la discapacidad específica, por ejemplo ¿Cómo realizar la prueba de lenguaje del examen ICFES-SABER 11 2013, siendo el objetivo de esta evaluar la competencia de lectura cuando el evaluado no puede tener contacto directo e independiente con el texto?

Para determinar si la medición del rasgo evaluado es invariante entre los grupos, recientemente se han realizado estudios sobre el funcionamiento de las pruebas SABER cuándo se evalúa competencia lectora, en estos se identificaron los ítems de la subprueba de Lenguaje en las cohortes 2008 y 2009 que presentaron funcionamiento diferencial entre los grupos de personas con y sin limitación visual (Herrera, Soler, Espinosa, Lancheros y Jiménez, 2012). Con este resultado, tales ítems se sometieron a un análisis de contenido para encontrar las posibles razones por las que se presenta el DIF, entre las cuales se destaca el uso de textos largos y complejos o preguntas que implican "que los evaluados deban volver a leer el texto cada vez que respondan a una pregunta ligada a ese texto" (Espinosa Garzón, 2013, p. 69).

Evaluación de la comprensión de lectura y población invidente

Para el INCI la discapacidad visual no solo se refiere a las personas ciegas (quienes no perciben la luz), sino también a las personas que perciben poca luz y "que aún con el uso de gafas tienen una pérdida visual que dificulta sus actividades diarias" (INCI, sin fecha,p.2). Según el registro para la Localización y Caracterización de Personas con Discapacidad del Ministerio de Salud y Protección Social de las 882.232 personas que se han reconocido como discapacitadas, 368.865 declaran presentar algún tipo de limitación visual. Cifras como éstas muestran la importancia de tener en cuenta las necesidades de esta población, para garantizar equidad con respecto a las personas que no presentan discapacidad.

La lectura es considerada como una capacidad intelectual superior que va más allá de la decodificación de las palabras (Soler, 2013). La comprensión de textos es una actividad cognitiva compleja en la que se mueven tanto la interacción entre las características del texto, como el estilo de pensamiento del lector. Los procesos cognitivos involucrados en esta tarea permiten integrar la información contenida en el texto con información previamente adquirida a través de un proceso inferencial (González, 2004). De acuerdo con este autor, cuando la lectura se da a través del canal visual sigue un procesamiento en el que se agrupan varias letras de acuerdo con el nivel de habilidad del lector, mientras que a través del canal táctil, se sigue un procesamiento secuencial que implica el uso de la memoria de trabajo para almacenar letras individuales hasta que se llegue al volumen de letras requerido para identificar toda la palabra.

Ochaíta (citado en Soler, 2013) plantea que el acceso a la información a través de sentidos como el tacto o el oído es más difícil, comparada con el acceso realizado por medio visual, ya que se requiere mayor tiempo para decodificar y reconocer las palabras al realizar la lectura. Si bien el desarrollo cognitivo se forma de manera semejante en niños con y sin limitación visual, los niños ciegos tienen más dificultades para definir el significado de palabras del vocabulario, muestra de ello, es el hecho de procesar auditivamente y realizar una verbalización de palabras, ya que desde el punto de vista cognitivo esto no significa necesariamente que las palabras sean totalmente comprendidas por el individuo (Barraga, 1978).

En cuanto a la comprensión lectora por medio auditivo (en este caso, el medio de presentación del examen SABER 11), Soler (2013) plantea que "la persona que escucha pierde la concentración rápidamente, lo que provoca que no pueda responder apropiadamente; en muchas ocasiones, el mensaje oral va acompañado de una serie de estímulos que interfieren con el mensaje" (p.22), por lo que debe ser más cuidadosa en la selección de la información verdaderamente relevante para la efectiva comprensión de un texto.

Rosa y Ochaíta (citados por Soler, 2013) manifiestan que a pesar de no encontrarse diferencias entre personas con y sin limitación visual en la ejecución de tareas como la discriminación de fonemas, la interpretación de estos datos es difícil cuando se requiere que

los sujetos realicen tareas con requisitos cognitivos superiores, como lo es la comprensión de lectura. Por su parte, Hoover y Gough (1990) afirman que la modalidad sensorial de cada una de las vías de acceso de información, visual, táctil y auditiva, supone una serie de diferencias dadas no sólo por el tipo de estímulo objeto, sino también por la velocidad de procesamiento que cada uno de estos procesos implica.

Teniendo en cuenta estas diferencias, en muchos casos se sugiere el uso de formas paralelas de evaluación para medir el mismo constructo, sin embargo, este tipo de procedimientos plantea la necesidad de usar técnicas que permitan la comparación apropiada de los resultados obtenidos cuando las pruebas sean distintas.

En este caso se habla de una comparación de estas pruebas por medio de la equiparación y no una comparación de modalidades de prueba debido justamente al constructo; ya que, como lo señalan Cirino, Romain, Barth, Tolar, Fletcher y Vaughn (2012), desde ciertas teorías la comprensión de lectura puede ser entendida como el producto de componentes relacionados con la decodificación y la comprensión de escucha. Así pues podría considerarse que la evaluación de la comprensión de lectura no se convierte en una evaluación de la "comprensión de escucha" simplemente por el medio en que la persona percibe/accede a la información.

En el caso de la prueba de lenguaje del examen SABER 11, si se tiene en cuenta que la población invidente debe acceder al texto por medio de la lectura que hace un tercero, se puede considerar que los dos grupos de evaluados no están, de hecho, respondiendo a la misma prueba a pesar que los ítems que la componen sean los mismos; ya que mientras el estudiante sin discapacidad visual es evaluado mediante la aplicación de la prueba escrita, las personas invidentes o con baja visión no tienen acceso directo y autónomo al texto, lo que según lo encontrado en la literatura, implicaría procesos cognitivos distintos en los grupos.

Aunque en la prueba Saber 11 de 2013-2 se denominaba a esta parte del test como prueba de lenguaje, el modelo que ésta sigue evalúa principalmente la comprensión lectora debido a que los componentes y las competencias definidos para esta se centran en la evaluación de la capacidad de entender un texto o hacer inferencias a partir del mismo.

Como se puede extraer de la tabla 1 (tomada de Casas Hernández, 2016), los tres componentes de función semántica de los elementos locales, configuración del sentido global del texto y de intertextualidad se combinan con las competencias (interpretativa, argumentativa y propositiva) que se relacionan directamente con el manejo que el evaluado pueda alcanzar de los textos presentados en la evaluación.

Así pues, el nivel más básico de comprensión de los textos hace referencia al reconocimiento y uso adecuado de las palabras y el significado de las mismas dentro de un texto; un nivel más avanzado requerirá del estudiante integrar los significados y dar un sentido general del texto, lo cual genera una postura u opinión propia frente al mismo; finalmente, se espera que un estudiante con un nivel de comprensión lectora alto, no sólo comprenda los textos sino que sea capaz de relacionarlo con otros contextos y de establecer relaciones del mismo en distintas situaciones.

Componente	Competencia / Proceso		
	Interpretativa/ Recordar	Argumentativa/ Recordar y Analizar - Sintetizar	Propositiva/ Recordar, analizar, sintetizar y aplicar
Función semántica de los elementos locales	Está relacionada de forma directa con aspectos como el vocabulario, la idea principal, los detalles, las secuencias, las inferencias y la estructura del texto, que le permiten al lector interpretar el texto a partir de un nivel de comprensión literal. Busca determinar la identificación del significado de palabras y oraciones, recordar o reconocer pasajes del texto, entre otras.	Identificar la explicación correcta para un elemento específico del texto que refleja una idea contenida en el texto completo. Indagar por el sentido que puedan tener las palabras del título en la construcción e intención textual de la lectura. Analizar una expresión y encontrar el argumento que complete el enunciado.	Identificar la importancia de los diferentes elementos que componen el texto: palabras, frases, párrafos, y la relación de estos con la comprensión literal del texto. Capacidad para introducir una idea nueva que guarde relación con lo contenido en el texto sin cambiar el sentido, una idea que quepa perfectamente.

Componente	Competencia / Proceso		
	Interpretativa/ Recordar	Argumentativa/ Recordar y Analizar - Sintetizar	Propositiva/ Recordar, analizar, sintetizar y aplicar
Configuración del sentido global del texto	Identificar los elementos del texto para darle un sentido global. A partir de la recuperación o identificación de elementos del texto la persona debe realizar una integración de la información. Tomar elementos claves del texto e interpretarlos de manera global.	Sustentar información, dar explicaciones, plantear juicios de valor y relaciones con conocimientos previamente adquiridos. Incluye la formación de puntos de vista propios del lector a partir del texto y sus conocimientos previos, plantear posiciones acerca del contenido de un texto, distinguir un hecho de una opinión, captar sentidos implícitos, etc.	Inferir conclusiones, el lector debe hacer relaciones y proponer un marco temático general del texto. Por ejemplo: Atribuir el título general de la lectura a partir de la temática que se plantea.
Intertextualidad: del sentido del texto hacia otros textos	Comprender lo descrito en el texto para relacionarlo con una idea ajena al éste pero que guarde relación con la idea del texto. Por ejemplo: Identificar la frase que mejor refleja lo descrito en el texto.	Establecer una relación entre conocimientos previos y las ideas planteadas por el autor para llegar a la comprensión global del texto. Identificar la explicación más adecuada y la pertinencia para usar cierto tipo de textos en contextos específicos.	Sustentar y dar explicaciones, plantear juicios de valor y relaciones con conocimientos previamente adquiridos. Incluye la formación de puntos de vista propios a partir del texto y sus conocimientos previos e inferir información que no se presenta en el texto.

Tabla 1. Componentes y competencias evaluadas en la subprueba de Lenguaje SABER 11 2013-2.

Indudablemente la identificación y definición de los procesos posiblemente involucrados es una cuestión que ameritaría un estudio completo y complejo, sin embargo, una forma de evidenciar el posible efecto de interacción entre variables asociadas al canal de entrada de la información y la habilidad en comprensión de texto, son los estudios de Funcionamiento Diferencial de los Ítems que permiten identificar tareas o preguntas en las cuales videntes y no videntes que se suponen con la misma habilidad podrían tener diferente probabilidad de acierto en los ítems que conforman el test presentado.

Funcionamiento Diferencial de los Ítems

Teóricamente se maneja el término de unidimensionalidad para hacer referencia a la garantía de que un ítem o una prueba midan solo un atributo y que no se presenten interferencia con alguna otra variable o rasgo, sin embargo, como señala Herrera (2005) "técnicamente no es posible que un instrumento de medición psicológica evalúe una única dimensión exactamente identificable" (p.55), por lo que se suele hablar de un factor o atributo que se ve mejor representado en la medición y la evaluación.

La expresión funcionamiento diferencial de los ítems (DIF) hace referencia a la diferencia en la probabilidad de responder un ítem o pregunta entre examinados o grupos diferentes que se suponen tienen el mismo nivel del atributo evaluado (Khalid y Glas, 2013). Este funcionamiento diferencial puede presentarse por varias razones, entre ellas se destaca la posibilidad de que el ítem mida algo más que lo que se quiere evaluar y que ese algo sea irrelevante para la medición del atributo (Herrera, 2005).

Si se define el DIF al comparar grupos con el mismo nivel de habilidad se requiere un método para comparar los grupos con el mínimo sesgo en la medición, así pues se suelen igualar según los puntajes totales de la prueba o la estimación de la magnitud de atributo a partir de las respuestas a los ítems que la componen. Este procedimiento se realiza en dos fases, en la primera se igualan los grupos con la totalidad de los ítems en la prueba, y luego se realiza un procedimiento similar, con la única diferencia que se excluyen los ítems que según el primer paso presentan DIF. En la teoría de respuesta al ítem (IRT, por sus siglas en inglés) se asume que un ítem presenta funcionamiento diferencial si las funciones de

respuesta a dicho ítem no son iguales entre los grupos (Herrera, 2005; Santana, 2009) por lo que se requiere que las estimaciones de los distintos grupos se encuentren la misma escala (Prieto Adánez y Dias Velasco, 2003)

En el caso de la prueba de comprensión de lectura en la población invidente, es posible que se presenten ítems que funcionen de manera diferencial debido al procesamiento de la información que deben realizar y los procesos cognitivos involucrados en el procesamiento de la información, ya sea por el canal auditivo o por medio visual. El estudio de Soler (2013) con la prueba SABER 11 encontró que algunos ítems funcionan de manera diferencial, sin embargo el canal de entrada auditivo para procesar la información por parte de la población invidente no es una variable nueva que sea introducida en la evaluación, ya que las personas con limitación visual están acostumbradas a recibir la información por medios auditivos, por lo que plantea que algunos aspectos propios de las prueba influyen en el desempeño y que "el diseño de éstas no estaría contemplando este mecanismo de recepción de la información, el cual requiere por parte de quien responde, un mayor uso de recursos cognitivos como memoria y atención" (p.). Al no tener en cuenta este tipo de consideraciones en la construcción de los ítems para la evaluación, puede introducirse variables que influyan en una disminución de la precisión de la medición, lo que ocasionaría un funcionamiento diferencial en los ítems que componen la prueba.

Igualmente para el caso de la lectura en voz alta, también pueden considerarse la influencia que tiene en la precisión de la medida las diferencias en la entonación de los lectores, ya que ellos son quienes manipulan principalmente el texto; sin embargo se busca controlar este tipo de variabilidad con el entrenamiento que reciben los lectores de esta prueba por parte del ICFES.

Este funcionamiento diferencial tiene implicaciones directas sobre la precisión de los resultados de la evaluación y la calificación de la misma porque introduce dentro de la medición de un constructo otras variables que no están contempladas en este, por lo que es necesario hacer uso de métodos estadísticos que permitan hacer comparables los puntajes de las pruebas de estos grupos (personas con y sin limitación visual) de forma precisa buscando garantizar una calificación más precisa; en este caso, la equiparación de

puntuaciones que permite hacer comparaciones entre diferentes grupos o pruebas, haciendo intercambiables entre si las puntuaciones de las dos pruebas.

Equiparación de puntuaciones

Los diferentes tipos de procedimientos de equiparación de puntuaciones han sido presentados por diferentes autores como métodos que buscan garantizar la equidad, mejorar la comparación al asegurar que los puntajes de test, o formas paralelas del mismo, sean comparables e intercambiables, porque las pruebas miden el mismo constructo en niveles de habilidad similares y con la misma confiabilidad (Zhu (1998), Dorans y Holland (2000), Dorans (2007), Dorans, Moses y Eignor (2010), Puhan (2012)).

Este tipo de procedimientos permite hacer mediciones equitativas de diferentes grupos porque posibilitan comparar resultados de manera precisa a pesar de que cada grupo responda distintas formas de una prueba, como por ejemplo, test en diferentes idiomas o pruebas paralelas aplicadas en diferentes fechas, etcétera. Como señala Dorans (2007) la equiparación busca eliminar los efectos de la dificultad relativa diferencial que puede estar presente en las diferentes versiones de prueba, y hacer que las comparaciones entre los resultados de quienes respondieron tales versiones tengan un significado; Holland (2013) respalda esta afirmación, señalando que se corrigen los puntajes de las inevitables diferencias en la dificultad de las pruebas.

Sin embargo, se han planteado algunas discusiones en las que se señala que la equiparación puede no ser la mejor opción para realizar las comparaciones entre grupos; por ejemplo Van Der Linden (2013) afirma que una de las dificultades es que las puntuaciones equiparadas de una persona dependen del desempeño total del grupo que le fue asignado, ya que la equiparación se realiza con base en las distribuciones de puntaje observado de los grupos totales; igualmente señala que en la equiparación se suelen confundir dos factores, la dificultad de las formas de prueba y las diferencias en la habilidad de los grupos de examinados que las responden, por lo que sugiere que deben hacerse otro tipo de transformaciones para ajustar las dificultades de cualquiera de las formas de prueba, teniendo en cuenta las diferencias en la habilidad de las personas.

Antes estas críticas, autores como Dorans (2013) y Holland (2013) señalan que la confusión puede darse sobre algunos términos que deben ser revisados con detenimiento (diferencia entre linking –enlace- y equiparación, por ejemplo), sobre todo cuando se trata de la equiparación de puntaje verdadero. Esto se debe a que cuando se trabaja con las puntuaciones verdaderas se consideran equivalentes las puntuaciones de dos test cuando estas corresponden al mismo nivel estimado de habilidad, es decir, se aplica el método de equiparación a las distribuciones que se generan teóricamente y no a las observadas (Navas, 1996; Kolen y Brenan, 2004, citados en Lancheros, 2013).

En la equiparación de puntuaciones se distinguen cuatro fases principales, que no necesariamente son consecutivas:

1. Elección de un diseño para la recolección de datos y determinar en qué momento se pueden equiparar.
2. Seleccionar el procedimiento estadístico para equiparar las puntuaciones. En general estos métodos estadísticos suelen clasificarse en dos grupos, los basados en la teoría clásica de los test (TCT) y los que utilizan la teoría de respuesta al ítem (IRT).
3. Formulación de la ecuación de equivalencia y tabla de conversión o de equivalencia entre puntuaciones o entre parámetros, según el caso.
4. Evaluación del proceso realizado, donde se revisan los efectos de los diferentes tipos de error presentados para determinar si la equiparación fue exitosa y los resultados son realmente comparables. (Pacheco, 2006, p.17)

Los tres diseños más utilizados son: Diseño de un solo grupo, diseño de grupos equivalentes y diseño con test de anclaje (Zhu (1998), Dorans et al (2010)). El primero de estos diseños hace referencia a que el mismo grupo de personas responda a las dos pruebas y de estos resultados obtener la ecuación de equiparación; una de las ventajas de este diseño es que presenta menor nivel de error porque solo se manejan los datos de un grupo, aunque debido a esto también se pueden presentar efectos de fatiga y práctica (Zhu, 1998).

En el diseño de grupos equivalentes, se tienen tantos subgrupos como pruebas a comparar, por lo que cada grupo responde a una de las pruebas (por ejemplo, subpoblación 1 toma la prueba X, subpoblación 2 toma la prueba Y, etc). Aunque en este diseño se eliminan los efectos de fatiga, Zhu (1998) señala que al tener varios grupos se introduce

sesgo, debido a que generalmente los subgrupos no presentan la misma distribución a lo largo del continuo de habilidad.

Finalmente el diseño con test de anclaje, consiste en que dos pruebas diferentes comparten algunos de sus ítems y son aplicadas a grupos de evaluados diferentes; así, a pesar de la variabilidad que puede presentarse debido a las diferencias entre los grupos, no es necesario que estos tengan la misma distribución de habilidad, ya que la función de los ítems comunes es facilitar el ajuste de las diferencias encontradas en los grupos, con base en los estadísticos de dichos ítems. El diseño de anclaje también puede realizarse con personas, es decir, hay participantes que hacen parte de los dos grupos y con base en sus resultados, se realizan las estimaciones necesarias.

Los ítems de anclaje son seleccionados por ser representativos, tanto por su contenido como por sus especificaciones estadísticas porque se utilizan para comparar los puntajes de estos en las dos formas de prueba (Kane (2009), Raykov (2010)); en la práctica, dichos ítems se sitúan en el mismo orden en las dos pruebas, para poder controlar los posibles efectos de la práctica y el cansancio, además estos ítems ayudan a disminuir el sesgo que puede presentar la función de equiparación cuando los grupos no son equivalentes (Dorans et al, 2010), por esto se sugiere utilizar el diseño de grupos no equivalentes con test de anclaje – NEAT (Non Equivalent Anchor Test).

El anclaje por ítems puede ser interno o externo. En el anclaje interno, los ítems comunes hacen parte de la calificación total de las pruebas, mientras que en el anclaje externo el desempeño en estos ítems no influye en la calificación final del test (Kane, 2009). Debido a esto el anclaje interno puede presentarse como un tipo de anclaje más confiable o estable porque al ser parte de la calificación total de la prueba, su correlación con las mismas es mayor, lo que genera estabilidad a través de su uso (Dorans et al, 2010), sin embargo este aumento en la correlación puede resultar espuria justamente por la presencia de ítems comunes.

Para Elousa y López-Jáuregui (2008) la equiparación de pruebas adaptadas usando el diseño de test de anclaje se puede resumir en dos grandes etapas: En la primera se debe analizar la equivalencia psicométrica entre las pruebas y en la segunda se escogen los ítems

que definitivamente harán parte del test de anclaje, a partir del cual se definirá la función de equiparación.

Generalmente se espera que los ítems de anclaje conformen una versión pequeña de la prueba total en la que se garantiza, como plantean Dorans, Kubiak, & Melican, 1998, (citado en Sinharay y Holland (2007)), que la media y la variación de dificultad de los ítems sea similar a la del test completo, sin embargo en la práctica puede ser difícil garantizar este último requerimiento, ya que en la prueba es más fácil encontrar ítems con dificultad media. Así pues, en este caso se utilizaría lo que Sinharay y Holland (2006) denominan un "miditest" que consiste en utilizar ítems de dificultad media; ya que como lo plantean los mismos autores en 2007, que no se ha establecido una manera de elegir un miditest cuando los ítems de anclaje están basados en estímulos compartidos (como el mismo texto para diferentes preguntas de comprensión lectora, por ejemplo).

Para la primera etapa, Holland y Dorans en 2006 (citados en Dorans et al, 2010) plantean 5 requisitos que deben tener las pruebas para poder desarrollar la equiparación entre dos pruebas (X y Y), estos son:

1. Mismo constructo: Las dos pruebas deben medir el mismo atributo.

2. Mismo nivel de confiabilidad: Los test deben tener niveles de confiabilidad similares.

3. Simetría: La ecuación de equiparación para transformar los puntajes de la prueba *X* a los de la prueba *Y*, debe ser inversa a la ecuación para los transformar los puntajes de *Y* a *X*.

4. Equidad: No debe importar cuál de las pruebas responda el examinado, el resultado debe ser el mismo.

5. Invarianza poblacional: La función de equiparación para los puntajes de las pruebas *X* y *Y* debe mantenerse, sin importar la subpoblación sobre la cual se haya derivado.

Con respecto a estos requisitos Lord (citado por Dorans, 2013) planteó que son tan difíciles de alcanzar que es prácticamente imposible equiparar, o que si se alcanzan no es necesaria la equiparación porque las pruebas serias completamente paralelas. En el caso de la prueba de lenguaje del examen ICFES -SABER 11, se tiene exactamente los mismos ítems construidos bajos las mismas especificaciones para los grupos de examinados.

Sin embargo es importante destacar las diferencias en el canal de entrada de información, y esto es lo que las hace diferentes y apropiadas para realizar procedimientos

de equiparación. Solo después de tener los datos de la aplicación de la prueba se podrá determinar si se cumplen adecuadamente los demás requisitos, ya que si se violan estos supuestos los métodos de equiparación funcionarán de manera deficiente (Puhan, 2010).

Métodos de equiparación

Los métodos de equiparación suelen clasificarse en dos grandes grupos: los basados en la teoría clásica de los test (TCT) y los que se basan en la teoría de respuesta al ítem (IRT).

Según Pacheco (2006) existen tres grandes posibilidades para establecer una métrica común entre las puntuaciones de dos tests, aunque las dos primeras pueden reunirse en lo que se han llamado métodos basados en la TCT. De acuerdo con la clasificación de Pacheco (2006), las transformaciones pueden ser: a) Transformaciones centiles, en las puntuaciones equiparadas sean aquellas que corresponden al mismo centil, b) transformaciones lineales en las que se consideran puntuaciones equiparadas las que corresponden a la misma puntuación típica y c) transformaciones por nivel en las que se considera que las puntuaciones son equiparables cuando estas corresponden a un nivel estimado del rasgo evaluado.

Entre los métodos que realizan transformaciones centiles se encuentra el método equipercentil. Según Han, Kolen y Pohlmann (1997) en este se utilizan los puntajes brutos o las distribuciones de los puntajes observados, es decir, se tabulan las distribuciones de frecuencia relativa acumulada para las dos formas, y luego se obtienen los puntajes equiparados a partir de dichas distribuciones.

En los métodos que utilizan transformaciones lineales, las puntuaciones de un test Y equivalentes a las de un test X vienen dadas por una transformación lineal de las puntuaciones de Y, siendo función de la media y desviación típica de ambas pruebas. La forma de determinar el valor de estas medias y desviaciones típicas es diferente según sea el diseño de equiparación utilizado para la recolección de datos.

Finalmente las transformaciones por nivel de habilidad se realizan con métodos basados en la teoría de respuesta al ítem, en estas se deben determinar las constantes de

transformación lineal que relaciona las estimaciones obtenidas para los parámetros de los individuos y de los ítems en dos ocasiones distintas (Pacheco, 2006).

En su investigación Lancheros (2013) comparó 3 métodos de equiparación, dos que realizan la transformación lineal (método Tucker y método Levine) y uno basado en IRT; (Media/Sigma). Estos procedimientos se realizaron con base en las estimaciones de los parámetros de la prueba de lectura del examen SABER del año 2008; se simularon 1000 matrices de respuesta iguales en longitud de prueba y estimación de parámetro de dificultad, para analizar el comportamiento de los métodos de equiparación a lo largo de las mismas. Como resultado se encontró que el método de Media/ Sigma muestra menor error de equiparación y el método de Tucker es más estable a través de las réplicas realizadas.

Teniendo como punto de partida esta investigación, se hará énfasis en los métodos de Tucker y Media/Sigma para evaluar su desempeño al equiparar los puntajes empíricos de los grupos de personas con y sin limitación visual que aplicaron a la prueba de lenguaje ICFES – SABER 11.

Método Tucker

El método de Tucker, hace parte de los procedimientos de equiparación basados en la teoría clásica de los test – TCT, es un método lineal de puntaje observado que se suele utilizar con diseño NEAT (Pacheco (2006), Zu (2012)), debido a que se basa en la sustitución de parámetros. Entre sus supuestos principales se encuentra que los parámetros de la ecuación de regresión de X y de Y sobre el test de anclaje (A) son idénticos en las subpoblaciones 1 y 2 (grupos que toman la prueba X y Y, respectivamente) y que la varianza de los errores es idéntica en los grupos. Así mismo, se asume que la distribución condicional de X dado A y de Y dado A en la población objetivo es la misma para cualquier población, por lo que se usa el puntaje del test de anclaje para relacionarlo con los puntajes totales (Dorans et al (2010), Puhan (2010)).

Dado que la equiparación realiza un ajuste a la dificultad, es decir, se estima el puntaje equiparado a partir de medidas que den cuenta de la dificultad de las pruebas es importante señalar que la equiparación lineal NO considera que las diferencias de dificultad entre las

dos pruebas se mantengan constantes a lo largo de los puntajes, sino por el contrario, permite su variación dentro de la escala (Kolen y Brennan, 2014).

Livingston (2014) señala que el método Tucker es solo un método lineal básico ajustado, en el que se sustituye las estimaciones de media y desviación estándar, por los estimadores de la "población objetivo" para las dos formas de prueba. En este tipo de métodos no solo se asume la invarianza poblacional, también se asume que la relación para cada muestra a equiparar en la población objetivo es una relación condicional (Dorans et al., 2010), por lo que se busca que las puntuaciones z de las dos formas sean iguales, según la siguiente fórmula (von Davier, A. y Kong, 2005):

$$\frac{x-\mu_{XT}}{\sigma_{XT}} = \frac{y-\mu_{YT}}{\sigma_{YT}} \quad (1)$$

Si se despeja *y*, el resultado es la fórmula de la función de equiparación lineal para determinar la puntuación de la prueba Y en una población T es:

$$Lin_{XY;T}\,(X) = \mu_{YT} + \sigma_{YT}\ (\,(x - \mu_{XT})/\sigma_{XT}\,) \quad (2)$$

Donde se denominan X y Y a las pruebas a equiparar, por tanto *x* y *y* representan las puntuaciones en cada una de estas pruebas. μ_{XT} y μ_{YT} son las medias de las pruebas X y Y en una población T y σ_{XT} y σ_{YT} son las desviaciones de las pruebas X y Y en una población T.

Así pues, lo que se enuncia en la ecuación 1 es que la distancia entre la puntuación de la prueba y la media de esta en el test en la población T, sobre la varianza de la prueba se asumen iguales para poder realizar la equiparación lineal. Para el caso del método Tucker cuando se tiene un diseño de anclaje, las ecuaciones para obtener la media y la desviación estándar de las pruebas, se pueden expresar como lo señalan Kolen y Brennan (2014)

$$\mu_2(X) = \mu_1(X) - \alpha_1\,(X/A)(\mu_1(A) - \mu_2(A)) \quad (3)$$

$$\mu_1(Y) = \mu_2(Y) - \alpha_2\,(Y/A)(\mu_1(A) - \mu_2(A)) \quad (4)$$

$$\sigma_2^2(x) = \sigma_1^2(X) - \alpha_1^2\,(X/A)(\sigma_1^2(A) - \sigma_2^2(A)) \quad (5)$$

$$\sigma_1^2(y) = \sigma_2^2(Y) - \alpha_2^2\,(Y/A)(\sigma_1^2(A) - \sigma_2^2(A)) \quad (6)$$

Utilizando los términos definidos anteriormente, las ecuaciones 3 y 4 expresan el modo en que se estima la media X y Y que se utiliza en la ecuación de equiparación, usando la media de X y Y en el test de anclaje (A), la desviación estándar de X o Y en ese mismo anclaje y la media de la prueba que se desea estimar. Para estimar la desviación estándar para X y Y para la equiparación (ecuaciones 5 y 6, respectivamente) se sigue un procedimiento similar, pero en este caso no se utiliza la media de X o Y, sino la varianza de estas pruebas.

Para cada una de estas ecuaciones se observa que, al ser un método lineal de estimación de puntajes, es necesario contar con una pendiente (α) de transformación. Utilizando los términos definidos anteriormente, se presentan las ecuaciones 7 y 8 en las cuales se observa que la pendiente depende de las desviaciones estándar de cada una de las pruebas en el test de anclaje.

$$\alpha_1\,(X/A) = \frac{\sigma_1(X,A)}{\sigma_1^2(A)} = \frac{\sigma_2(X,A)}{\sigma_2^2(A)} \quad (7)$$

$$\alpha_2\,(Y/A) = \frac{\sigma_1(Y,A)}{\sigma_1^2(A)} = \frac{\sigma_2(Y,A)}{\sigma_2^2(A)} \quad (8)$$

Este método puede presentar sesgo en la equiparación cuando las subpoblaciones difieren en nivel de habilidad (Puhan, 2010), porque si se da el caso en que la relación entre el puntaje de la prueba de anclaje y la prueba total sea baja, el método asumirá que las diferencias se deben únicamente a la dificultad de las pruebas. Una estrategia es la conformación de una población sintética conformada por la combinación ponderada de las dos subpoblaciones, de manera que la media de población sintética se calcula con a partir de las medias de las subpoblaciones ponderadas (con pesos que sumen 1) dependiendo del tamaño de cada una, así "cuanto más débil la correlación, más se acercará la media de la población sintética a la media de las muestras observadas" (Puhan, 2012, p. 317). Sin

embargo, según Van Der Linden (2013) trabajar con poblaciones sintéticas es una de las principales fuentes de sesgo en la equiparación.

Lancheros (2013) con el uso de datos simulados, encontró que el método Tucker presentaba el mejor funcionamiento dentro de su estudio comparativo porque presentó menor varianza a lo largo de las réplicas; sin embargo Suh et al. (2009) señalan que el método Tucker suele presentar mejor comportamiento cuando los grupos son similares y las pruebas son diferentes. Así pues, será interesante analizar el funcionamiento empírico de este método con los datos de la prueba de comprensión de lectura de SABER, ya que las pruebas contienen los mismos ítems, pero son los grupos y la manera en que procesan la información de los textos los que hacen la diferencia.

Método Media/Sigma

El método media/sigma hace parte del grupo de métodos basados en la IRT denominados métodos basados en los momentos (Battauz, 2013), en estos se deriva el valor de las constantes β y α - necesarias para obtener los parámetros de dificultad y discriminación del ítem - a partir de los primeros momentos de las distribuciones de las estimaciones de los parámetros (Pacheco, 2006), es decir, se estima el parámetro de dificultad de los ítems que componen la prueba y, con base en estos, se plantea la ecuación de equiparación.

Kolen y Brennan (2014) afirman que la equiparación basada en IRT, requiere de al menos los tres pasos siguientes: a) Estimar los parámetros de los ítems, b) escalamiento de los parámetros en una escala IRT usando una transformación lineal (sustitución de parámetros), es decir, se sustituyen las medias y las desviaciones estándar de los parámetros estimados de los ítems a los parámetros de los ítems comunes en el diseño NEAT (Kolen y Brennan, 2014, p. 183), y c) convertir los puntajes de la forma nueva de prueba, a los de la forma anterior. Por ejemplo, si se usa el puntaje por cantidad de respuestas correctas en la nueva forma, estos deben convertirse a la escala utilizada en la forma antigua.

Según Guemin y Fitzpatrick (2008) cuando se mantiene el modelo IRT, los parámetros estimados en distintas calibraciones están linealmente relacionados, así pues una ecuación

lineal puede convertir la estimación de los parámetros en otra métrica sin cambiar la probabilidad de acierto en modelos IRT.

El método media/sigma en diseño de anclaje NEAT suele ser el más usado porque las estimaciones del parámetro de dificultad son más estables, por lo que se utilizan las medias y las desviaciones estándar de la dificultad de los ítems comunes, en lugar de las estimaciones de los parámetros (Kolen y Brennan, 2014). Así lo primero que se debe hacer es estimar la dificultad de los ítems del test de anclaje en cada grupo de forma independiente, por lo que estos ítems tienen dos estimaciones distintas de su dificultad.

Si las dos estimaciones están linealmente correlacionadas se tendrá una ecuación como la siguiente (Pacheco, 2006)

$$b_y = \frac{S_y(A)}{S_x(A)} b_x + \left(\overline{Y}(A) - \frac{S_y(A)}{S_x(A)} \overline{X}(A) \right) \tag{9}$$

Donde b_x y b_y son los valores estimados de la dificultad de los ítems para las pruebas X y Y; $S_y(A)$ y $S_x(A)$ son las estimaciones de la desviación estándar de los test y finalmente $\overline{Y}(A)$ y $\overline{X}(A)$ son las estimaciones de las medias de las dificultades para el anclaje.

En la ecuación 9 se expresan los valores que se utilizan para definir el valor equivalente para la dificultad de cada uno de los ítems en la prueba Y (b_y). Se observa que, de manera similar al método Tucker, se trabaja principalmente con las medias ($\overline{Y}(A)$ y $\overline{X}(A)$) y desviaciones estándar ($S_x y S_y$) pero a diferencia del método clásico, no se utiliza la calificación directa sino la estimación de las dificultades de los ítems que hacen parte de la prueba.

Lancheros (2013) señala que dada la robustez del modelo de Rasch, es frecuente encontrar que en los procedimientos de equiparación de puntuaciones primero se calibran los ítems con dicho modelo, para luego realizar el escalamiento, así las constantes se podrían establecer con base en el siguiente modelo

$$\alpha = \frac{S_y(\theta)}{S_x(\theta)} \tag{10}$$

$$\beta = \mu_y(b) - \frac{S_y(\theta)}{S_x(\theta)}\mu_x(\theta) \tag{11}$$

Donde μ_x y μ_y representan la media estimada de la dificultad de los ítems para el grupo 1 y para el grupo 2, respectivamente, y θ hace referencia al nivel de habilidad de los evaluados dentro de la IRT.

Luego de realizar la equiparación de las estimaciones de los parámetros y del nivel de habilidad, puede usarse la métrica de las puntuaciones verdaderas o la métrica de las puntuaciones observadas, esto debido a que la métrica de θ de habilidad es más difícil de interpretar.

Como resultado de la aplicación de los métodos de equiparación basados en la TCT se obtiene una tabla de equivalencia por medio de la cual se presenta los puntajes correspondientes a la prueba X en la prueba Y, mientras que en los métodos basados en la teoría de respuesta al ítems, la tabla de equivalencia expresa la correspondencia en términos de dificultad de los ítems o en términos de magnitud de atributo.

En la investigación desarrollada por Lancheros (2013) se concluyó que el método *Media/sigma* presentó menor error de equiparación en cada una de las matrices de respuestas simuladas en relación con los métodos basados en la teoría clásica de los test (Métodos lineales Tucker y Levine). Sin embargo, uno de los inconvenientes que tiene este método es que solo se tiene en cuenta la información del parámetro de dificultad de los ítems, por lo que se deja influenciar por las diferencias en este parámetro, dejando de lado la información que pueden brindar los otros parámetros (Pacheco, 2006; Kolen y Brennan, 2014). Igualmente, en un estudio presentado por Li, Jiang y von Davier (2012), se encontró que este método presenta mayor sesgo y varianza muestral, comparado con otros métodos que se basan en IRT, sobretodo en los puntajes extremos.

Así pues, para establecer si los métodos funcionan de manera adecuada y si los resultados son estables, es necesario evaluar el funcionamiento de la ecuación de equiparación dentro de la muestra.

Evaluación de la equiparación

Dorans y Holland (2000) afirman que "los métodos de equiparación pueden realizarse, pero muy rara vez los datos muestran indicadores que alerten que se hizo algo mal. Esta falta de advertencia basada en los datos en la equiparación mal realizada, es un problema que subyace en todos los tipos de enlace". (p. 283)

Puhan (2009) señala dos tipos de error en la equiparación, el error aleatorio depende del tipo de muestra utilizada para realizar la equiparación (si en verdad es representativa de la población a la que pertenece), mientras que el error sistemático puede resultar de diversos factores como el uso de un método inadecuado para el caso específico, utilizar ítems de anclaje que no sean representativos. "Cuando se minimiza el error sistemático, el error de la equiparación es resultado de la variabilidad en el muestreo."(p.80)

El error estándar de equiparación da como resultado un estimado del error debido a la muestra y da cuenta de la variabilidad de los resultados obtenidos (Liu, Schulz y Yu, 2008), sin embargo Kolen y Brennan (2014) manifiestan que aún no hay un procedimiento analítico general para estimar el error sistemático en este tipo de procedimientos, mientras autores como Van Der Linden (2013) afirman que el error de equiparación tiende a disminuirse si la longitud o la confiabilidad de las pruebas incrementa, más no necesariamente ocurre lo mismo al aumentar el número de examinados.

Así, el error aleatorio es la "desviación estándar de los puntajes verdaderos equiparados con IRT sobre las réplicas hipotéticas del proceso de equiparación de puntaje verdadero en muestras extraídas de una población" (Liu et al, 2008, p. 263).

El método bootstrap ha sido, uno de los métodos más famosos para estimar el error estándar de equiparación; en este procedimiento se extraen múltiples muestras con reposición para calcular repetidamente las puntuaciones equiparadas y estimar el error típico de equiparación (Lancheros (2013), Kolen y Brennan (2014)).

Según Kolen y Brennan (2014) el error estándar usando el bootstrap se calcula por medio de los siguientes pasos (p. 250)

1. De un grupo de datos tamaño *N*, se toma una muestra con reposición de tamaño *n*.

2. Se calculan los estadísticos de interés de cada muestra.

3. El proceso se repite siguiendo los pasos anteriores *R* veces

4. Calcular la desviación estándar de los estadísticos de interés sobre las *R* muestras. Esta desviación estándar es el error estándar "bootstrap" estimado del estadístico.

Así pues, para el caso específico de la equiparación, el error utilizando el método bootstrap, tomaría la siguiente forma (Kolen y Brennan, 2014, p. 252)

$$se = \sqrt{\frac{\sum(e_{yr}(x_i) - e_y(x_i))^2}{R-1}} \quad (12)$$

Donde

$e_y(x_i) = \frac{\sum e_{yr}(x_i)}{R}$, valor estimado a partir de R muestras.

Y $e_{yr}(x_i)$ representa el estimado equivalente en x_i usando los datos de los remuestreos realizados, lo que indica que la estimación del error con el método bootstrap es simplemente un cálculo de la diferencia entre los resultados obtenidos en cada muestreo y el resultado esperado de la equiparación.

En su momento Tsai et al (2001) planteaban que se usaba el método bootstrap para estimar el error en la equiparación IRT porque no se habían desarrollado ecuaciones apropiadas para dicha estimación en ese tipo de equiparación.

Sin embargo, a pesar de sus ventajas, autores como Hutchison (2010) aseguran que este método es muy costoso, por lo que sería preferible utilizar otras medidas de error, sin embargo estas no han sido lo suficientemente trabajadas con los datos empíricos por lo que en esta investigación se prefiere el método de remuestreo (bootstrap). Además este método, debido en parte a su sencillez, permite hacer las comparaciones entre métodos que

tienen fundamentos teóricos diferentes, como es el caso de los métodos basados en la teórica clásica y los basados en la teoría de respuesta al ítem.

Como se ha evidenciado a lo largo de la revisión, la equiparación es un procedimiento que no se encuentra lejos de la controversia, no solo por la variedad de tipos de equiparación que se pueden encontrar en la literatura, sino por el tipo de uso que se les ha dado a estos métodos en diferentes circunstancias, según las características de las pruebas, la muestra, el manejo que se dé a los datos e incluso el tipo de problema que se espera resolver. Ya que el uso más común de los métodos de equiparación es en los grandes centros de evaluación y en pruebas de aplicación masiva, es necesario evaluar algunas circunstancias específicas en las que estos puedan funcionar, buscando aprovechar al máximo los recursos disponibles de tal forma que se pueda garantizar la medición y evaluación adecuada, a pesar de las diferentes características que puedan tener los grupos de evaluados.

MÉTODO

Teniendo en cuenta que el presente trabajo tiene como objetivo la evaluación de los métodos de equiparación, es necesario conservar algunas de las variables de estudios anteriores para comparar la estabilidad de los procedimientos con datos de distintas aplicaciones, específicamente la investigación realizada por Lancheros (2013). En esta investigación se utilizaron entonces los datos de la aplicación de la ICFES-SABER 11 de lectura, aplicación del segundo semestre del 2013.

Debido a que se quiere hacer comparables los puntajes de prueba de dos grupos distintos (personas con y sin limitación visual) se utilizó un diseño de equiparación de grupos no equivalentes con test de anclaje – NEAT. Además de ser el diseño más adecuado para las condiciones en las que se desarrolló esta investigación, permite comparar los resultados que se obtengan en el presente trabajo con las investigaciones anteriores en este tipo de población.

Instrumento

La prueba SABER 11 es el instrumento utilizado por el Estado Colombiano como parte de la evaluación de la calidad de la educación media en el país y para el año 2013 estaba compuesta por varias subpruebas como: Lenguaje, matemáticas, biología, física, química, filosofía, ciencias sociales e inglés. Todas ellas aplicadas en dos sesiones, una en la mañana y otra en la tarde del mismo día simultáneamente en todas las ciudades del país, proceso que se repite dos veces en el año.

La subprueba de lectura del examen SABER 11 2013-2 está compuesta por 24 ítems de opción múltiple con única respuesta; según la descripción del ICFES (2013b) está estructurada en cinco competencias básicas: (1) identificación o repetición de lo que dice el texto, (2) resumen, (3) información previa, (4) gramática y (5) pragmática; divididas en tres tipos de lectura, así el evaluado debe ser capaz de realizar lecturas literales, inferenciales y críticas. Dentro de esta subprueba también se propone una competencia comunicativa, entendida como la capacidad de entender y la habilidad de producir textos manejando tres dimensiones de lenguaje: sintáctica, semántica y pragmática, siguiendo los lineamientos establecidos por el Ministerio de Educación Nacional (ICFES, 2013b).

En la segunda aplicación de 2013, utilizada en el presente trabajo, se aplicaron dos formas de prueba diferentes de 24 ítems cada una con 12 ítems comunes en las dos formas de tal manera que se contó con información de 36 ítems aplicados, como se muestra en la figura 1.

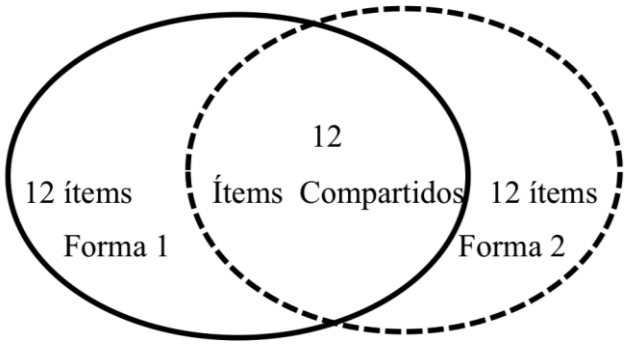

Figura 1. Representación de la composición de la subprueba de lenguaje de la prueba SABER 11- segunda aplicación de 2013

Bases de datos

La base de datos utilizada en el presente estudio fue proporcionada directamente por el ICFES y contenía las respuestas de 537837 personas a la subprueba de Lenguaje por parte de los examinados en el segundo semestre de 2013, tanto videntes como personas que reportaron limitación visual. Se utilizaron muestras aleatorias de examinados videntes y la población total de personas con limitación visual, como se describe más adelante.

La población de videntes se definió con dos criterios: 1) personas que no declararon ante el ICFES necesitar la ayuda de un lector para presentar la prueba y 2) personas que hubieran respondido a la totalidad de los ítems que componen la subprueba de lenguaje. La población de examinados con limitación visual estuvo conformada por la totalidad de examinados que: 1) reportaron necesitar ayuda de un lector debido a una deficiencia visual y 2) hubieran contestado al menos 20 de los 24 ítems de la prueba de Lenguaje. En total la base de datos para estudio quedó conformada por las respuestas de 518557 examinados sin limitación visual y 104 personas con limitación visual, distribuidos como se muestra en la tabla 2.

Tabla 2. Número de examinados por forma de prueba

Estudiantes	Forma 1	Forma 2	Total
Con limitación visual	46	58	104
Sin limitación visual	230400	288053	518453
Total	230446	288111	518557

Finalmente, la base de datos que contenía las respuestas a las formas de la prueba se transformó en tabla de datos en dicótoma que contenía 1 o 0; es decir, comparando las respuestas de los examinados con los vectores de claves, para cada uno de los ítems se asignó el valor de 1 si la respuesta era correcta y 0 si era incorrecta.

Procedimiento

En primer lugar se procedió a la depuración de la base de datos de acuerdo con los criterios mencionados anteriormente y a la recopilación de la información sobre la estructura y especificaciones de prueba. Una vez conformada la base de datos depurada se procedió a seleccionar las muestras de la población de examinados videntes y con base en las mismas se estimó el parámetro de dificultad de los ítems, se identificaron aquellos que presentaban posible funcionamiento diferencial y se obtuvieron las tablas de equiparación mediante los dos métodos de interés: Tucker y Media/sigma. El criterio de comparación de los mismos fue el error de equiparación.

Muestreo

Los muestreos se realizaron con dos objetivos: fijar la razón de tamaños de muestra *con limitación visual: videntes* y estimar el error de equiparación mediante el procedimiento de bootstraping. Se decidió mantener la razón de tamaño en 1:25; es decir, por cada persona con limitación visual se seleccionaron 25 videntes, ya que se ha encontrado que un aumento en la razón de tamaños de muestra puede disminuir la potencia de los estadísticos y generar error tipo I, en particular cuando se trata de detectar ítems con posible funcionamiento diferencial (Arias Patiño (2008), Berrio Beltrán (2008), Herrera & Gómez (2008), Santana (2009),). Para mantener esta razón de tamaños, se realizaron muestreos de la población de examinados videntes de cada una de las dos formas de prueba, seleccionando 25 personas sin limitación visual, por cada persona invidente.

Se extrajeron muestras aleatorias de videntes con tamaño de muestra entre 1150 y 1450, y se conservó siempre la población total de 104 examinados con limitación visual. Para la forma 1 cada muestreo estuvo compuesto por cerca de 1150 datos (0,49% del total de datos para esta forma), mientras que para cada muestreo de la forma de prueba 2 se extrajeron el 0,50% de los datos (cerca de 1450 datos).

Identificación de ítems con DIF

A partir de esta base modificada y por medio del programa estadístico Winsteps, se estimó la dificultad de cada uno de los ítems para los dos grupos y de acuerdo a esta, se identificaron los ítems que presentaron funcionamiento diferencial en cada una de las formas de prueba teniendo en cuenta las estimaciones realizadas con métodos Mantel-Haenszel (desarrollado por Mantel y Haenszel en 1959 y utilizado para la detección de DIF por Holland y Thayer, (Arias Patiño, 2008)) y por medio de la prueba t de Welch.

Se consideraron ítems con DIF aquellos que fueran detectados por los dos procedimientos en las primeras 25 muestras extraídas, para no generar mayores costos computacionales ya que a lo largo de estas se encontró que los resultados fueron suficientemente estables. Se declararon con DIF los ítems que resultaron identificados en al menos 15 de las 25 réplicas según la significancia de las prueba chi cuadrado de MH y la t de diferencia de medias. La identificación del grupo desfavorecido obedeció a los valores observados en el estadístico; esto es, valores de MH mayores que 1 y valores positivos en la prueba t indican que el ítem favorece al grupo de personas sin limitación visual mientras que valores de MH menores que 1 o valores de la prueba t negativos, indican que el ítem favorece a las personas con limitación visual.

A partir de esta identificación, se creó una base de datos adicional en la cual se eliminaron estos ítems con el fin de evaluar los métodos de equiparación cuando estos ítems no son calificados. Es importante realizar este tipo de comparaciones para tener información con cuál de las dos opciones (mantener los ítems con DIF o no) la equiparación es más precisa, ya que, como señalan Chu y Kamata (2005), usualmente es más simple eliminar de la equiparación los ítems que presentan funcionamiento diferencial, sin considerar otras posibilidades.

De esta manera se tuvieron pares de bases de datos por cada forma de prueba, una que contenía las respuestas de los 24 ítems para los dos grupos de estudiantes y una que contenía los ítems que no presentaron funcionamiento diferencial.

.

Equiparación

Al trabajar con un diseño de equiparación NEAT se determinó que los ítems de anclaje para la equiparación fuesen aquellos que tenían niveles de dificultad estimada por medio del modelo de Rasch similar para los dos grupos. Teniendo en cuenta que para que las estimaciones obtenidas a partir de los ítems de anclaje sean lo más informativas posible sobre el comportamiento de las pruebas totales, Angoff (citado por Gau (2004)) sugirió que el anclaje debía ser de cerca del 20% de la prueba total por lo que en este caso se seleccionaron 5 ítems de cada forma de prueba (que conforman el 20,83% de la subprueba) para el anclaje. Teniendo en cuenta que es una prueba pequeña, compuesta solo por 24 ítems, los ítems de anclaje se incluyeron en la calificación total de la prueba conformando así un test de anclaje interno (Dorans, Moses y Eignor,, 2011)

Para el procedimiento lineal Tucker se calificó cada una de las formas de prueba según la cantidad de respuestas correctas de cada uno de los estudiantes y a partir de esta calificación se obtuvo la media y desviación estándar de cada forma para cada uno de los grupos (personas con y sin limitación visual). Este procedimiento se repitió para cada uno de los pares de muestras donde se realizó la estimación de media y desviación estándar, tanto para la calificación de la prueba total, como para cuando se omite de la calificación los ítems con funcionamiento diferencial, en las dos formas de prueba.

En el caso del método Media-Sigma se utilizó la media y desviación estándar de las medidas de dificultad estimada por medio del modelo de Rasch, para poder así hacer la equiparación siguiendo este método de teoría de respuesta al ítem. Las estimaciones de dificultad de los ítems, así como las del funcionamiento diferencial, se realizaron por medio del programa Winsteps, a partir de las cuales se estimó la media y desviación estándar de las mismas, tanto para la prueba con los 24 ítems, como para la prueba cuando no se tenían en cuenta los ítems con funcionamiento diferencial.

Estimación del error de equiparación

Con esos resultados se hicieron las tablas de equivalencia para cada forma de prueba, contra las cuales se compararon los resultados de cada muestro para estimar cuál de los métodos es más preciso. El error de equiparación se calculó por medio del método bootstrap, en el cual el resultado de la ecuación de equiparación en cada uno de los muestreos realizados fue comparado con la tabla de equivalencia establecida para la forma de prueba respectiva por medio de la diferencia entre el resultado esperado (tabla de equivalencia de forma de prueba) y el resultado encontrado (resultado de cada uno de los muestreos). Para el método de media sigma, se calculó el error de estimación de los parámetros de dificultad a través de las muestras.

Para determinar cuál método es más estable entre los muestreos se utilizó el coeficiente de variación, el cual se calcula con base en la media y desviación estándar de los errores encontrados para cada método en las formas de prueba. El objetivo del uso de este coeficiente para este caso es comparar la estabilidad de esos errores a lo largo de las muestras, más no comparar el contenido de los mismos; se reportan además, los errores máximos y mínimos de cada método entre los muestreos.

RESULTADOS

En primer lugar, las distribuciones de puntajes brutos en la subprueba de lenguaje para las poblaciones de examinados con y sin limitación visual se muestran en las figuras 1 y 2. En estas se aprecia que si bien para ninguna de las dos poblaciones se encuentran puntajes de prueba mayores a 22, se puede observar que para los estudiantes con limitación visual el puntaje máximo es de 18 para la forma 1 y 17 para la forma 2. También se observan algunas diferencias en las distribuciones de puntajes entre las formas 1 y 2 para las personas con limitación visual, diferencias que no se observan en el grupo de examinados sin limitación visual.

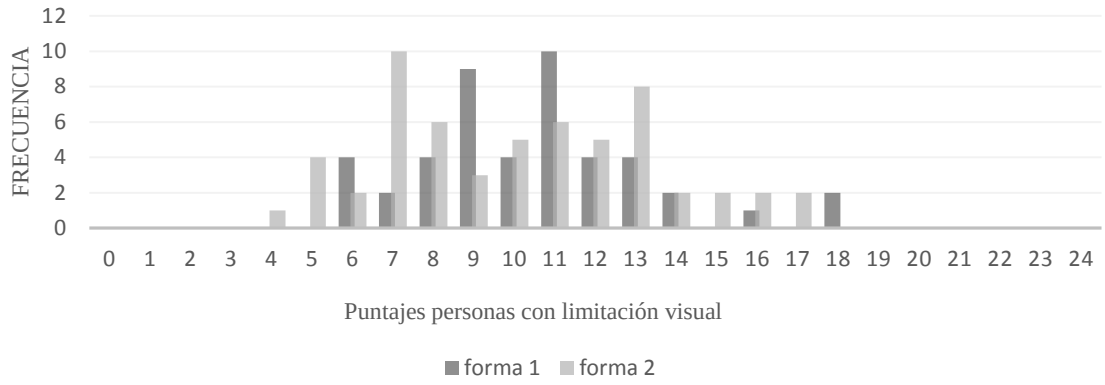

Figura 1. Distribuciones de puntajes de personas con limitación visual en las dos formas de prueba

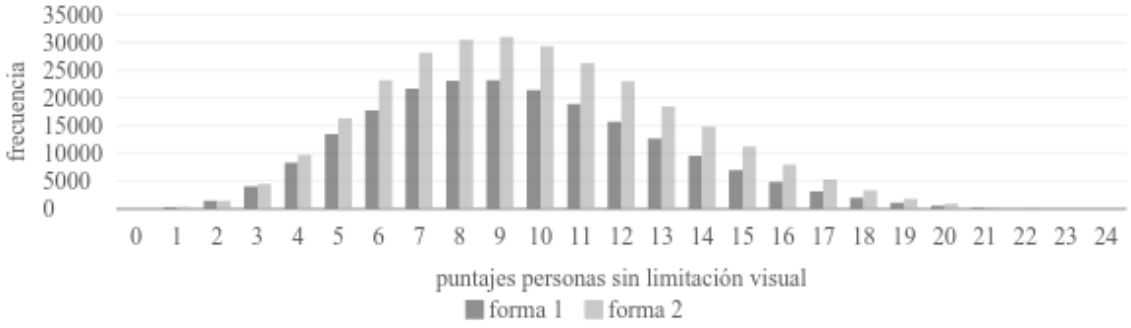

Figura 2. Distribuciones de puntajes de personas sin limitación visual en las dos formas de prueba

Conformación del test de anclaje

El anexo 1 muestra las estimaciones de dificultad para los 36 ítems analizados para las poblaciones de examinados con y sin limitación visual. De acuerdo con las estimaciones de la dificultad, se seleccionaron los cinco ítems que se utilizaron como test de anclaje. En la tabla 3 aparecen los ítems del test de anclaje para las dos formas de la prueba con sus respectivas estimaciones de dificultad.

Tabla 3. .Dificultad de ítems de anclaje para cada grupo en las formas de prueba

Forma de prueba	Estimaciones de Dificultad		
	Ítem	Grupo sin limitación visual	Grupo con limitación visual
Forma 1	10	0,10	0,24
	16	-1,13	-1,01
	18	0,01	0,14
	32	-0,26	-0,15
	36	0,55	0,66
Forma 2	7	0,52	0,58
	20	0,28	0,24
	21	0,03	0,00
	30	0,21	0,16
	33	-0,49	-0,46

Identificación de ítems con DIF

De los 36 ítems analizados, se identificaron 7 ítems con funcionamiento diferencial, en la tabla 4 se presentan los promedios de dificultad para estos ítems en los primeros 25 muestreos. Se puede observar que se detectaron ítems con funcionamiento diferencial a favor de ambos grupos, si bien cuatro ítems parecen desfavorecer a las personas con limitación visual, otros 3 ítems desfavorecen a los estudiantes sin limitación. Los ítems 23 de la forma 1, 4, 12 y 25 de la forma 2 mostraron funcionamiento diferencial en contra del grupo de examinados sin limitación visual.

Tabla 4. Ítems identificados con funcionamiento diferencial

Forma de prueba	Ítem	Prueba t	Valor p	χ^2 de Mantel-Hanszel	Prob.
Forma 1	2	2,46	0,01	4,36	0,03
	23	-2,22	0,02	4,02	0,04
	31	2,19	0,03	7,41	0,006
Forma 2	3	2,27	0,02	3,62	0,05
	4	-2,6	0,01	6,09	0,01
	12	-2,03	0,04	4,45	0,03
	25	-2,97	0,004	9,91	0,001

De acuerdo con estos resultados se recalificaron las pruebas excluyendo los ítems 2 y 31 de la forma 1 para los examinados sin limitación visual y el ítem 23 de la misma forma para las muestras de examinados con limitación visual. La forma 2 se recalificó excluyendo el ítem 3 para las personas sin limitación visual y el 4, 12 y 25 para las muestras de examinados con limitación visual.

Tablas de equiparación

Con los resultados obtenidos hasta aquí, se obtuvieron ocho tablas de equivalencia para los puntajes de los grupos, es decir se tienen cuatro tablas de equivalencia por forma de prueba, de la siguiente manera: 1) Tabla para el método Tucker con la calificación total, 2) Tabla para el método Tucker eliminando de la calificación los ítems con funcionamiento diferencial, 3) Tabla para el método Media-Sigma con los 24 ítems que componen la prueba y 4) Tabla para el método Media-Sigma al eliminar los ítems con funcionamiento diferencial. La información resultante se resume a continuación.

En la tabla 5 se presentan la tabla de equivalencia para la forma 1 obtenida mediante el método de equiparación lineal Tucker, y el error estimado para cada una de las posibles puntuaciones de esa forma de prueba. En la tabla 6 se presentan los resultados del mismo método para la forma 2. En la primera columna aparecen los puntajes brutos posibles para el grupo de personas sin limitación visual y en la segunda columna el respectivo puntaje estimado para examinados sin limitación visual. En las dos últimas columnas se presentan los puntajes estimados para las pruebas cuando se eliminan de la calificación los ítems con funcionamiento diferencial, por esto el puntaje máximo deja de ser 24. Es importante señalar que estas tablas y las presentadas para el método media-sigma contienen los datos promedio de todas las muestras utilizadas.

Tabla 5. Puntajes estimados y error promedio para la forma 1

Puntaje	Estimado Tucker	Error Tucker	Estimado Tucker sin DIF	Error Tucker sin DIF
0	4,57	0,07	4,17	0,07
1	5,17	0,05	4,81	0,05
2	5,76	0,04	5,49	0,03
3	6,36	0,03	6,16	0,02
4	6,96	0,03	6,84	0,02
5	7,55	0,04	7,51	0,01
6	8,15	0,05	8,19	0,01
7	8,74	0,07	8,86	0,01
8	9,34	0,09	9,54	0,01
9	9,94	0,13	10,21	0,01
10	10,53	0,16	10,89	0,02
11	11,13	0,21	11,56	0,03
12	11,72	0,26	12,24	0,04
13	12,32	0,32	12,91	0,05
14	12,92	0,38	13,59	0,07
15	13,51	0,45	14,27	0,09
16	14,11	0,53	14,94	0,11
17	14,70	0,62	15,62	0,13
18	15,30	0,71	16,29	0,16
19	15,90	0,81	16,97	0,19
20	16,49	0,91	17,64	0,22
21	17,09	1,03	18,32	0,25
22	17,68	1,15	18,99	0,29
23	18,28	1,28	-----	----
24	18,88	1,42	------	----

Los resultados mostraron que para la forma 2 el error es menor para el método lineal Tucker comparado con los resultados obtenidos en la forma 1, así mismo las diferencias entre los puntajes estimados para el grupo de personas con limitación visual y el puntaje en la prueba de personas sin limitación es menor que los presentados en la otra forma de prueba, esto puede deberse a la distribución de los puntajes originales en las respectivas formas de prueba.

Igualmente se obtuvo que para la forma 1 de la prueba, el error de equiparación disminuye cuando se hace la equiparación de las pruebas que no incluyen los ítems con funcionamiento diferencial, sin embargo ocurre lo contrario con la segunda forma de prueba.

Tabla 6. Puntajes estimados y error promedio en para la forma 2

Puntaje	Estimado Tucker	Error Tucker	Estimado Tucker sin DIF	Error Tucker sin DIF
0	1,673	0,053	1,315	0,04
1	2,541	0,039	2,245	0,02
2	3,409	0,027	3,176	0,01
3	4,277	0,018	4,106	0,02
4	5,146	0,011	5,036	0,04
5	6,014	0,006	5,967	0,07
6	6,882	0,004	6,897	0,11
7	7,750	0,004	7,827	0,17
8	8,618	0,006	8,758	0,24
9	9,486	0,011	9,688	0,32
10	10,354	0,019	10,618	0,42
11	11,223	0,029	11,548	0,52
12	12,091	0,041	12,479	0,65
13	12,959	0,055	13,409	0,78
14	13,827	0,072	14,339	0,93
15	14,695	0,092	15,270	1,09
16	15,563	0,114	16,200	1,27
17	16,431	0,138	17,130	1,45
18	17,299	0,165	18,061	1,65
19	18,168	0,194	18,991	1,87
20	19,036	0,225	19,921	2,09
21	19,904	0,259	20,852	2,33
22	20,772	0,295	21,782	2,59
23	21,640	0,334	22,712	2,85
24	22,508	0,375	-----	-------

En los resultados obtenidos se observa que los puntajes bajos de la prueba de personas con limitación visual son mayores a los de las personas sin limitación visual, sin embargo este resultado se invierte en los puntajes más altos de la prueba, igualmente se observa que

hay mayor nivel de error en la estimación para los puntajes más altos de la prueba, porque en la calificación original de la prueba las personas con limitación visual no alcanzaron puntajes mayores a 20.

Tabla 7. Dificultad de los ítems y error de equiparación promedio para la Forma 1

Ítem	Equivalencias para las formas completas			Equivalencias para las formas excluyendo los ítems con DIF		
	Dificultad SLV	Dificultad Estimada	Error estimado	Dificultad SLV	Dificultad Estimada	Error estimado
1	0,708	0,844	0,008	0,650	0,836	0,009
2	0,478	0,612	0,012	_____	_____	_____
5	-1,260	-1,134	0,009	-1,315	-1,136	0,0092
6	1,437	1,577	0,023	1,379	1,567	0,023
8	1,896	2,035	0,019	1,836	2,023	0,019
9	-0,606	-0,476	0,006	-0,662	-0,480	0,005
10	0,098	0,229	0,003	0,042	0,224	0,003
12	0,492	0,626	0,008	0,433	0,617	0,008
13	0,443	0,577	0,007	0,385	0,569	0,007
15	0,807	0,944	0,010	0,749	0,936	0,010
16	-1,134	-1,005	0,001	-1,190	-1,008	0,001
18	0,020	0,152	0,002	-0,036	0,148	0,003
19	0,014	0,145	0,006	-0,043	0,140	0,006
21	0,029	0,159	0,004	-0,028	0,153	0,004
23	0,488	0,623	0,008	0,430	0,615	0,008
24	-0,020	0,113	0,005	-0,075	0,109	0,0053
25	-0,351	-0,220	0,007	-0,409	-0,227	0,015
26	-0,661	-0,531	0,003	-0,716	-0,534	0,005
28	-0,628	-0,500	0,006	-0,684	-0,503	0,006
29	-0,177	-0,044	0,005	-0,234	-0,051	0,017
31	-1,733	-1,609	0,010	_____	_____	_____
32	-0,281	-0,150	0,005	-0,336	-0,153	0,004
34	-0,584	-0,456	0,007	-0,639	-0,459	0,007
36	0,526	0,658	0,002	0,468	0,650	0,002

Los resultados para el método media-sigma se presentan en las tablas 7 y 8. Se puede observar que para los 24 ítems de la forma 2 completa la dificultad estimada para el grupo de personas con limitación visual es mayor a la calculada para la prueba presentada por las

personas sin limitación; esto indica que la prueba es más difícil para las personas invidentes, a pesar de que el contenido de los ítems es el mismo para los dos grupos.

Aunque el error de equiparación es similar para la mayor parte de los ítems de la forma 2 en los dos procedimientos, cuando si se presentan diferencias, como por ejemplo en los ítems 7, 9 y 15, el error es mayor para las formas de prueba modificada, es decir, se encuentran mayores errores cuando se excluye de la calificación los ítems con DIF.

Tabla 8. Dificultad de los ítems y error de equiparación promedio para la Forma 2

Ítem	Equivalencias para las formas completas			Equivalencias cuando se excluyen ítems con DIF		
	Dificultad SLV	Dificultad Estimada	Error estimado	Dificultad SLV	Dificultad Estimada	Error estimado
2	0,493	0,516	0,009	0,472	0,658	0,010
3	-0,595	-0,621	0,001	_____	_____	_____
4	1,312	1,368	0,024	1,283	1,558	0,025
5	-1,233	-1,291	0,011	-1,255	-1,269	0,048
7	0,492	0,510	0,000	0,468	0,648	0,012
9	-0,598	-0,627	0,003	-0,623	-0,566	0,036
10	0,108	0,110	0,001	0,083	0,219	0,020
11	0,688	0,719	0,007	0,662	0,869	0,011
12	0,510	0,531	0,003	0,487	0,672	0,013
14	-0,065	-0,068	0,004	-0,090	0,032	0,014
15	0,840	0,871	0,002	0,813	1,028	0,016
16	-1,097	-1,148	0,008	-1,120	-1,119	0,045
17	-1,753	-1,833	0,011	-1,782	-1,854	0,061
20	0,270	0,281	0,002	0,245	0,403	0,009
21	0,020	0,015	0,006	-0,003	0,119	0,030
22	-0,320	-0,336	0,001	-0,347	-0,255	0,011
24	-0,023	-0,027	0,003	-0,048	0,075	0,012
25	-0,488	-0,511	0,001	-0,513	-0,441	0,018
26	-0,643	-0,673	0,003	-0,670	-0,616	0,028
27	1,722	1,800	0,029	1,697	2,024	0,029
30	0,198	0,204	0,002	0,172	0,319	0,008
33	-0,470	-0,490	0,001	-0,497	-0,419	0,013
35	0,035	0,036	0,001	0,005	0,136	0,013
36	0,588	0,613	0,010	0,568	0,763	0,015

Comparación del error de equiparación para los dos procedimientos

En general el error de equiparación es menor que para el método Tucker (tanto en la forma completa como la forma modificada en la que se eliminan los ítems con funcionamiento diferencial para ambas formas), estos resultados son consistentes con los encontrados en las investigaciones anteriores (Lancheros, 2013). De otra parte, las diferencias entre el nivel de error entre las formas son menores para el media/sigma que en el método clásico; al igual que las diferencias de error entre las formas completas y las modificadas son mínimas.

Aunque, como se mencionó anteriormente, estos valores no son directamente comparables, es destacable que en ambos casos se observa mayor magnitud de error en los puntajes bajos y altos de la distribución, ya que dentro de la base de datos las personas con limitación visual no alcanzaban estos puntajes. El hecho de no tener puntajes altos o bajos (o tener pocos) aumenta la magnitud del error de equiparación estimado para estas puntuaciones porque al realizar los muestreos utilizados para calcular el bootstrap, se disminuye la probabilidad de que estos puntajes extremos se vean representados adecuadamente, de forma tal que la diferencia entre el puntaje equiparado estimado para la muestra y la estimación total va a ser mayor.

Al realizar la comparación del funcionamiento de los métodos de equiparación cuando se utilizan con la calificación completa y la calificación sin los ítems con DIF para cada forma de prueba se encontró que el método media-sigma presenta menor error de equiparación. En la forma 2 se observó que el método lineal y el método media sigma con la prueba completa presentan menor error de equiparación que con la forma en la que se eliminan los ítems con DIF; este comportamiento difiere con respecto a los resultados obtenidos para la forma 1.

Finalmente, los resultados del coeficiente de variación de los errores calculado sobre las muestras para cada uno de los métodos, mostraron que el método lineal presenta mayor variabilidad, mientras que el método media-sigma presenta un funcionamiento similar para ambas formas en los dos escenarios planteados. Estos resultados se observan en la tabla 9.

Tabla 9. Descriptivos de las medidas de error encontradas para los métodos de equiparación

Forma	Método	Máximo	Mínimo	Media	Desviación	Coeficiente variación
Forma 1	Tucker	4,46	0,005	0,41	0,68	1,65
	Tucker/Sin DIF	0,86	0,03	0,23	0,16	0,70
	Media-Sigma	0,13	0,05	0,09	0,01	0,21
	Media-Sigma/Sin DIF	0,17	0,05	0,09	0,02	0,28
Forma 2	Tucker	0,51	0,018	0,19	0,10	0,52
	Tucker/Sin DIF	0,42	0,013	0,18	0,10	0,57
	Media-Sigma	0,23	0,11	0,14	0,04	0,29
	Media-Sigma/Sin DIF	0,40	0,21	0,28	0,07	0,27

DISCUSIÓN Y CONCLUSIONES

El objetivo principal de esta investigación fue comparar dos procedimientos de equiparación cuando se trata de hacer equivalentes las calificaciones de grupos no equivalentes: Examinados con limitación visual y sin limitación visual evaluados con la subprueba de Lenguaje de la prueba SABER 11. La comparación de los métodos se basó en la variabilidad de los errores de equiparación para las dos formas de la prueba cuando se equiparan las formas completas y cuando se excluyen los ítems identificados con Funcionamiento Diferencial.

La presencia de ítems con DIF se tuvo en cuenta porque, además de ser una variable importante que puede afectar la calidad de la equiparación, permitió comparar el desempeño de las formas de prueba para los dos grupos de interés. A partir de los descriptivos observados para las dos formas de la prueba y los resultados del análisis de DIF, se puede considerar que efectivamente se presentan diferencias en el funcionamiento para los grupos de estudiantes con y sin limitación visual. Así pues, ha de considerarse que en la medición de la comprensión de lectura por medio de "pruebas de lápiz y papel" para personas con limitación visual intervienen otras variables que no están asociadas directamente a la definición del constructo utilizada para la evaluación, lo que afecta directamente el desempeño y la calificación de estas personas frente a otras personas sin limitación visual.

Esto se evidencia, además con los resultados de las estimaciones de dificultad de los ítems de las dos formas de prueba para los dos grupos, las dificultades estimadas de los ítems en la equiparación suelen ser mayores a las estimadas para el grupo de invidentes, lo que indicaría que la prueba es efectivamente más difícil para este grupo.

Si bien es cierto que en las dos formas de prueba se encuentran ítems con funcionamiento diferencial que desfavorecen tanto al grupo de estudiantes con limitación visual, como a aquellos que no la tienen se puede asegurar que esto no es de ninguna manera una medida compensatoria; es decir, los resultados evidencian que el hecho de tener ítems que funcionen de manera diferencial para los dos grupos afecta directamente la

precisión de la medición y la medida de este efecto no es dada exclusivamente por la cantidad de ítems que desfavorecen a un grupo u otro.

En lo referente a las tablas de equiparación obtenidas para los dos formas de prueba mediante los dos procedimiento, pudo observarse que los dos métodos presentan funcionamiento similares (como se observa en la tabla 9 del apartado de resultados) en las formas 1 y 2, lo que puede considerarse como una evidencia que estas formas de pruebas funcionan de manera similar para evaluar el constructo de comprensión lectora.

Sin embargo, en el método Tucker se observa una inversión en los puntajes de prueba, es decir los puntajes equivalentes calculados son mayores para los puntajes menores a 6, pero menores para los puntajes mayores a 15, esto puede deberse a que en la calificación los estudiantes con limitación visual no tienen puntajes en los extremos (ni bajos, ni altos) de la calificación luego las equivalencias en puntajes altos conlleva errores importantes.

Un resultado que pone en duda la equivalencia de las formas de prueba es que la equiparación de la forma 1 presentó mayor error, comparado con la forma 2. Sin embargo, este resultado puede deberse a la diferencia en la variabilidad de los dos grupos, ya que para la forma 2 se presentó menor variación de puntaje en los muestreos, es decir, pocas veces se presentan resultados extremos o cambios significativos en la media estimada de cada muestreo para el grupo de personas sin limitación visual, cambios que si se presentaron en los muestreos de la forma 1.

Cuando se recalificaron las pruebas sin considerar los ítems detectados con DIF y se equipararon pruebas de diferente longitud, los resultados cambiaron ligeramente. Para la forma 2 el error por ítem es mayor que el error por puntaje cuando se eliminaron de la calificación los ítems con funcionamiento diferencial, esto posiblemente se deba a que en la forma 2 se eliminan de la calificación una mayor cantidad de ítems que en la forma 1 (4 ítems en total). Debe tenerse en cuenta que 3 o 4 ítems representan alrededor del 15% para una prueba de una longitud tan corta; esto sugeriría que, aunque es recomendable siempre revisar los ítems con funcionamiento diferencial, puede que si estos representan cierto porcentaje dentro de la prueba total, quizá la mejor decisión sea hacer un análisis de contenido y de identificación de sesgo antes de tomar la decisión de eliminarlos.

Referente al contenido de los ítems, una limitación importante de la presente investigación se deriva de la imposibilidad de conocer el contenido de los ítems por razones de confidencialidad del banco de preguntas del ICFES, lo que impide realizar un análisis a profundidad de los mismos. Sin embargo, los hallazgos indicarían la necesidad de revisar algunos ítems detectados con DIF con el fin de tomar decisiones frente a su pertinencia para la medición de comprensión lectora ya que aparentemente están midiendo un factor irrelevante respecto al constructo de interés por lo que presentan funcionamiento diferencial entre los grupos; algunos de los ítems que se encontraban en las dos formas de prueba (es decir los ítems compartidos), presentan funcionamiento diferencial significativo en una de las formas de prueba pero no en la otra. Teniendo en cuenta que no se conoce el contenido de los ítems, se podría especular que esta diferencia podría deberse a algún contenido o longitud en las otras preguntas de la forma específica, que pudo inducir a distracciones o respuestas que no necesariamente reflejaban el nivel de conocimientos o habilidad de los evaluados.

Debido a esa misma imposibilidad, los ítems de anclaje se eligieron considerando únicamente la información de la dificultad estimada de los mismos, sin embargo, para que los test de anclaje cumplan con los supuestos que plantean autores como Gau (2004), Dorans et al (2011) y Kolen y Brennan (2014), estos deben representar significativamente tanto el contenido de la prueba como la distribución de puntajes de la prueba total por lo que para futuras investigaciones se esperaría que los ítems de anclaje representen también las dimensiones de la comprensión de lectura que pretende evaluar la prueba total. En síntesis, para futuras investigaciones es necesario hacer también el análisis del funcionamiento de la equiparación en función del contenido particular de los ítems y de las dimensiones que estos representan dentro de las pruebas y el constructo definido.

En cuanto al error de equiparación por medio del bootstrap, se encuentra que en el método lineal Tucker se deja afectar fácilmente por los puntajes poco probables como los de los extremos de la distribución, por ejemplo, el error de estimación es mayor en los puntajes poco frecuentes en la muestras, mientras que la estimación es más estable para los puntajes de los que se tienen más datos.

Por otra parte, para el método media-sigma se encuentra que la cantidad de error es menor, esto en parte se debe al tipo de datos utilizados en este método ya que las estimaciones de dificultad de los ítems por medio del modelo de Rasch (y en general los modelos basados en IRT) se caracterizan por su estabilidad en diferentes muestras.

En síntesis, uno de los objetivos de la equiparación es hacer comparables las medidas de dos pruebas diferentes que miden el mismo constructo, por lo tanto es un procedimiento que debe evaluarse a profundidad ya que su uso permitiría medidas más equitativas sin necesidad de aumentar los costos o poner en riesgo la confidencialidad de los bancos de ítems; en este caso se utilizaron los métodos de equiparación para hacer comparables dos pruebas cuyo contenido era idéntico, pero que dadas las características de presentación de la prueba hacen que dentro de este estudio, estas sean consideradas diferentes.

De acuerdo con los resultados de este estudio, los métodos de equiparación efectivamente pueden contribuir a subsanar la diferencia que existe en la calificación de la comprensión lectora en estudiantes con limitación visual, en relación con los estudiantes que no sufren esta limitación. Si bien puede considerarse que las diferencias en cuanto a cantidad de puntaje (o puntos de diferencia entre los grupos) son pequeñas, es importante resaltar que el objetivo es hacer la medición más precisa, sobretodo en este tipo de pruebas de aplicación masiva y reguladas por el Estado Colombiano ya que estas tienen efectos reales en la población (facilitando el acceso a la educación superior, por ejemplo), por lo que la precisión de la medición es un compromiso que las entidades deben garantizar a todos sus evaluados.

En cuanto al funcionamiento de los métodos utilizados, se observa que los resultados obtenidos soportan las conclusiones de investigaciones anteriores (como la de Lancheros en 2013); así pues el método media-sigma haya mostrado ser más estable entre las diferentes réplicas, lo que apoyaría lo encontrado en la literatura sobre la robustez de este método aun cuando no se cumplen todos los supuestos del modelo de Rasch (Kolen y Brennan, 2014), ya que la prueba utilizada en este caso no es unidimensional (Herrera, Barajas, Casas, Rodríguez y Jiménez (2015), Barajas (2016)) y la cantidad de personas del grupo de estudiantes con limitación visual en mucho menor de lo que se suele utilizar en los análisis realizados con base en la teoría de respuesta al ítem. Además los resultados encontrados

con respecto al método media-sigma, pueden servir de argumento para contestar a la pregunta sobre por qué es necesaria la equiparación aun cuando se utilice el modelo de Rasch, ya que evidencia que la medida de dificultad de los ítems requieren ser corregidas cuando se presentan diferencias importantes entre los grupos.

Se evidenció que el funcionamiento de los métodos lineales de equiparación y los resultados de medidas de error por medio del método bootstrap depende en gran medida de las distribuciones de puntaje obtenido por los grupos que presentan las pruebas a equiparar, especialmente en los métodos clásicos de equiparación, ya que si las medias de las muestras difieren significativamente de la media y desviación de la muestra total, las estimaciones de los parámetros con los que se calcula el puntaje equiparado se ven directamente afectados.

En este tipo de situaciones se deben tener en cuenta varios factores para decidir cuál de los métodos es más adecuado utilizar ya que, aunque se encontró que el error del método Media-sigma es más estable entre réplicas, también se debe tener en consideración los supuestos que según la teoría de respuesta a los ítems se deben cumplir para utilizarlo de manera adecuada. Así pues, el método lineal Tucker puede ser útil cuando se tienen muestras pequeñas y cuando las pruebas a equiparar no necesariamente se basen en la medición de un constructo unidimensional, ya que el cálculo de la ecuación de la equiparación en este método es bastante sencillo al solo depender de la media y la desviación estándar; por su parte el método Media-sigma puede ser mejor aprovechado en muestras grandes donde se requiere un medida más estable del desempeño de los evaluados.

Finalmente, los métodos de equiparación se mostraron como una alternativa eficiente para mejorar la calificación haciéndola más precisa cuando, por diferentes motivos, dos grupos de evaluados no pueden presentar exactamente la misma prueba. Con los resultados encontrados se empiezan a ver sugerencias de posibles formas de hacer más precisos los procedimientos de equiparación, ya que lo que inició como el proceso de evaluación de los procedimientos de equiparación con ítems que presentan funcionamiento diferencial, se mostró como una alternativa para hacer estos procedimientos más precisos, según el caso.

Así pues, este trabajo constituye un aporte sobre los métodos a utilizar para una adecuada medición y evaluación de procesos psicológicos al mostrar el comportamiento de la equiparación en diversidad de condiciones, en este caso cuando se tuvo un alto porcentaje de ítems con DIF y desigualdad en los tamaños de muestra de los dos grupos de interés. Otro aporte destacable de este trabajo es resaltar la importancia de mejorar la precisión en la medición de comprensión lectora, ya que como proceso transversal, su evaluación es fundamental para la adecuada medición y evaluación de otras áreas de conocimiento como matemáticas o ciencias, en personas con limitación visual. Sin embargo, dado que no se realizó un análisis de contenido de los ítems, para futuras investigaciones se sugiere que se realicen estudios complementarios en los que se analice el DIF en función del contenido de los ítems, con el fin de tomar las decisiones en cuanto a la eliminación o no de algunos ítems en la calificación con fines de equiparación. Igualmente, podría considerarse la opción de evaluar el uso de los métodos equiparación de puntuaciones con más de dos grupos en los que, por ejemplo, se tenga en cuenta también la población con limitación auditiva.

REFERENCIAS

Angoff, W.H.,Modu, C,C. (1973) Equating The Scales Of The Spanish-Language Prueba De Aptitud Académica And The English-Language Scholastic Aptitude Test Of The College Entrance Examination Board. *ETS Research Bulletin Series*. doi: 10.1002/j.2333-8504.1973.tb00200.x

Arias Patiño, E. (2008) *Efecto de la razón de tamaño y el ajuste del modelo sobre el estadístico Mantel-Haenszel y su métrica delta en la detección del DIF.* Tesis inédita de maestría en psicología. Universidad Nacional de Colombia, Bogotá.

Barajas, R. (2016). *Validez en Test Adaptativos Informatizados (TAI): Evidencia en un TAI diseñado para evaluar comprensión lectora en personas con y sin limitación visual*. Tesis inédita de Maestría en Psicología. Universidad Nacional de Colombia, Bogotá.

Barraga, N. (1978). Disminuidos visuales y aprendizaje (enfoque evolutivo). España: Organización Nacional de Ciegos de España ONCE. http://sid.usal.es/idocs/F8/FDO23237/disminuidos_visuales_y_aprendizaje.pdf

Battauz, M (2013) IRT Test Equating In Complex Linkage Plans. *Psychometrika, 78 (3)*, 464-480. doi:10.1007/s11336-012-9316-y

Berrio Beltrán, Ángela Iannine (2008). *La razón de tamaños de muestra y desajustes del modelo en la detección de ítems con funcionamiento diferencial mediante el procedimiento de diferencia de dificultad.* Tesis inédita de maestría en psicología. Universidad Nacional de Colombia, Bogotá.

Casas Hérnandez, Maritza (2016) *Acomodaciones Computarizadas para la Evaluación de Comprensión Lectora en Estudiantes con y sin Limitación Visual*. Tesis inédita de Maestría en Psicología. Universidad Nacional de Colombia, Bogotá.

Chu, K. Kamata, A (2005). Test equating in the presence of DIF items. *Journal of Applied Measurement, 6 (3)*, 342-354.

Cirino, P.T., Romain, M.A., Barth, A.E., Tolar, T.D., Fletcher, J.M. y Vaughn, S. (2012). Reading skill components and impairments in middle school struggling readers. *Read Writ, 26(7), 1059–1086* DOI 10.1007/s11145-012-9406-3

Congreso de la República de Colombia (1997). Ley 361 de 1997. Por la cual se establecen mecanismos de integración social de la personas con limitación y se dictan otras disposiciones. Recuperado de http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=343

Corte Constitucional (2012). *Sentencia T-495/12. Proteccion Constitucional Reforzada De Niños Con Discapacidad*. Recuperado de http://www.corteconstitucional.gov.co/relatoria/2012/t-495-12.htm.

Corte Constitucional (2013). Sentencia T-598/13. Personas Con Discapacidad Como Sujetos De Especial Protección Constitucional-Protección nacional e internacional. Recuperado de http://www.corteconstitucional.gov.co/relatoria/2013/T-598-13.htm.

Dorans, N.J. (2007). Linking scores from multiple health outcome instruments. *Qual Life, 16,* 85–94. doi:10.1007/s11136-006-9155-3

Dorans, N.J. (2013). On Attempting to Do What Lord Said Was Impossible: Commentary on van der Linden's "Some Conceptual Issues in Observed-Score Equating" *Journal of Educational Measurement, 50 (3),* 304–314. doi: 10.1111/jedm.12017

Dorans, N.J. & Holland, P (2000) Population Invariance and the Equatability of Test: Basic Theory and the Linear Case. *Journal of Educational Measurement, 37 (4),* 281–306. doi:10.1111/j.1745-3984.2000.tb01088.x

Dorans, N.J, Moses, T., Eignor, D. (2010) *Principles and Practices of Tests Score Equating* (Reporte de investigación - ETS RR-10-29). Princeton, Educational System Service. Recuperado de http://www.ets.org/Media/Research/pdf/RR-10-29.pdf.

Dorans, N.J, Moses, T., Eignor, D. (2011) *Equating Test Scores: Toward Best Practices*. En Von Davier, A. A (Ed) (2011).Statistical Models for Test Equating, Scaling, and

Linking. Springer Books. Recuperado de http://link.springer.com.ezproxy.unal.edu.co/chapter/10.1007/978-0-387-98138-3_2

Elousa, P., López-Jáuregui A. (2008) "Equating between linguistically different test: Consequences for Assessment". *The Journal of Experimental Education, 76(4*), 387-402. doi:10.3200/JEXE.76.4.387-402

Espinosa Garzón, A.M. (2013). *Evaluación objetiva de los procesos cognitivos involucrados en la comprensión de lectura*. (Tesis inédita de maestría en psicología). Universidad Nacional de Colombia, Bogotá.

Gau, H. (2004). *The effect of different anchor tests on the accuracy of test equating for test adaptation*. Disertation. Ohio: The Faculty of de College of Education. https://etd.ohiolink.edu/!etd.send_file?accession=ohiou1089917802&disposition=inline

González, L. (2004). Assessment of Text Reading Comprehension by Spanish-speaking Blind Persons. British Journal of Visual Impairment, 22, 4-12. doi: 10.1177/026461960402200102

Guemin, L., Fitzpatrick, A. (2008). A New Approach to Test Score Equating Using Item Response Theory with Fixed C-Parameters. *Asia Pacific Education Review, 9 (3*). 248-261. doi: 10.1007/BF03026714

Han, T., Kolen, M., Pohlmann, J. (1997) "A Comparison Among IRT True – and Observed – Score Equatings and Traditional Equipercentile Equating". *Applied Measurement in Education, 10(2)*, 105-121. doi: 10.1207/s15324818ame1002_1

Herrera, A. N. (2005). *Efecto del tamaño de muestra y la razón de tamaños de muestra en la detección del funcionamiento diferencial de los ítems*. Tesis doctoral no publicada, Universidad de Barcelona, Barcelona.

Herrera, A. N. Gomez, J (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality & Quantity, 42 (6). 739-755*. http://dx.doi.org/10.1007/s11135-006-9065-z

Herrera, A. N., Lancheros, L. C. Jiménez, G. J. (2012). *Métodos de equiparación en pruebas SABER 11º: Comparación de procedimientos en personas videntes e invidentes*. Ponencia publicada en http://www.icfes.gov.co/investigacion/component/docman/doc_download/138-aura-nidia-herrera-catheryne-lancheros-javier-jimenez-metodos-de-equiparacion-en-pruebas-saber-11o?Itemid

Herrera, A. N., Soler, M. P., Espinosa, A.M., Lancheros, L.C. Jiménez, G. J. (2012). Procedimiento para establecer equivalencia en las puntuaciones de pruebas de aplicación masiva, en personas con y sin limitación visual. Informe técnico de investigación. Inédito. Bogotá: ICFES.

Herrera, A. N., Barajas, R., Casas, M., Rodríguez, D., 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.

Holland, P. (2007). A Framework and History for Score Linking. En Dorans, Pommerich y Holland (Eds). *Linking and Aligning Scores and Scales* (5-29). Springer

Holland, P. (2013). Comments on van der Linden's Critique and Proposal for Equating. *Journal of Educational Measurement, 50 (3),* 286-294. doi: 10.1111/jedm.12015

Hoover, W. & Gough, P. (1990). The Simple view of reading. *Reading and writing: An interdisciplinary Journal. 2*, 127-160. Recuperado de http://homepage.psy.utexas.edu/HomePage/Class/Psy338K/Gough/Chapter7/simple_view.pdf

Hutchison, D. (2010). The Standard Error of Moving Average Smoothed Equipercentile Equating. *Qual Quant, 44*. 783-791. doi: 10.1007/s11135-009-9231-1

Instituto Colombiano para la Evaluación de la Educación - ICFES (2008). Resolución 092 del 22 de febrero de 2008: Por la cual se expide reglamentación de los procedimientos para registro, inscripción, citación y presentación de exámenes ante el ICFES y se deroga la Resolución 256 de 2006. Bogotá.

Instituto Colombiano para la Evaluación de la Educación – ICFES (2013). Resolución 286 del 21 de mayo de 2013: Por la cual se reglamenta la presentación de la prueba electrónica SABER 11° Calendario A para la población con discapacidad auditiva en el 2013. Bogotá.

Instituto Colombiano para la Evaluación de la Educación – ICFES (2013b). *Sistema Nacional de Evaluación Estandarizada de la Educación. Alineación del examen SABER 11°*.Bogotá. http://www.icfes.gov.co/examenes/component/docman/doc_view/775-alineacion-del-examen-saber-11?Itemid=

Instituto Nacional para Ciegos -INCI (Sin fecha) La Inclusión Social De La Niñez Con Discapacidad Visual. Recuperado de www.inci.gov.co%2Fobservatoriosocial%2Fanalisissituacional%2Fsocial%3Fdownload%3D50%3Aninezcondiscapacidadvisual&ei=CbnlUrSBCdSsQT24ILoBQ&usg=AFQjCNGRdXMt9y3yGQwJBzrakAE3eSJQRg&bvm=bv.59930103,d.cWc

Instituto Nacional para Ciegos -INCI y Universidad Pedagógica Nacional - UPN. (2009). *"Miradas Valiosas" Lectores para personas con limitación visual más que una oportunidad*. Bogotá. Recuperado de: http://www.inci.gov.co/centro_documentacion/cartillas/Miradas_valiosas.pdf

Khalid M.N., Glas C.A.W. (2013) A scale purification procedure for evaluation of differential item functioning. *Measurement, 50*. 186-197. http://dx.doi.org/10.1016/j.measurement.2013.12.019.

Kolen, M; Brennan, R (2014) *Test Equating, Scaling, and Linking. Methods and Practices*. Tercera Edición. Springer. doi:10.1007/978-1-4939-0317-7

Lancheros Florian, L. C. (2013). *Métodos de Equiparación de Puntuaciones: Los exámenes de estado en población con y sin limitación visual.* (Tesis inédita de maestría en psicología). Universidad Nacional de Colombia, Bogotá.

Li, D; Jiang, Y; von Davier, A. (2012) The Accuracy and Consistency of a Series of IRT True Score Equatings. *Journal of Educational Measurement, 49 (2),* 167–189. doi: 10.1111/j.1745-3984.2012.00167.x

Liu, Y., Schulz, M., Yu, L (2008). Standard Error Estimation of 3PL IRT True Score Equating with an MCMC Method. *Journal of Educational and Behavioral Statistics, 33 (3),* 257-278. doi: 10.1111/j.1745-3984.2012.00167.x

Livingston, S. (2014). *Equating Test Scores (without IRT).* Segunda Edición. Educational Testing Service.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement,* 14, 139–160. http://www.jstor.org/stable/1434012

Ministerio de Educación Nacional. República de Colombia (2010). Decreto 869 del 17 de Marzo de 2010: Por la cual se reglamenta el Examen de Estado de la Educación Media, ICFES – SABER 11. Bogotá.

Ministerio de Salud y Protección Social (2012). *Discapacidad en Colombia. Registro para Localización y Caracterización de Personas con Discapacidad*. Bogotá. Recuperado de http://www.minsalud.gov.co/Documentos%20y%20Publicaciones/Cifras%20Registro%20de%20discapacidad%20(oct%202012).pdf.

Mislevy, R, J. (1992) Linking Educational Assessment: Concepts, Issues, Methods and Prospects. Educational Testing Service. Princeton.

Moliner, O. (2008). Condiciones, Procesos y Circunstancias que Permiten Avanzar Hacia la Inclusión Educativa: Retomando las Aportaciones de la Experiencia Canadiense. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 6*(2), pp. 27-44. http://www.rinace.net/arts/vol6num2/art3.pdf. Consultado el 03 de Enero de 2014

Pacheco, J. S. (2006). *Equiparación de puntuaciones. Revisión conceptual y metodológica" Tesis de grado profesional*. (Tesis inédita de pregrado en psicología). Universidad Nacional de Colombia, sede Bogotá.

Plan decenal de educación 2006 -2016 (2006). *Lineamientos en TIC*. Bogotá. Recuperado de http://www.plandecenal.edu.co/html/1726/articles-166057_TICS.pdf

Prieto Adánez, G., Dias Velasco, A. (2003) Uso del modelo de Rasch para poner en la misma escala las puntuaciones de distintos test. *Actualidades en Psicología. 19 (106)*, 8 -28. Recuperado de http://www.redalyc.org/pdf/1332/133217953001.pdf.

Puhan, G. (2009). Detecting and Correcting Scale Drift in Test Equating: An Illustration from a Large Scale Testing Program. *Applied Measurement In Education, 22*, 79 – 103. doi:10.1080/08957340802558391

Puhan, G. (2010). A comparison of chained linear and post-stratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*, 54–75. doi:10.1111/j.1745-3984.2009.00099.x

Puhan, G. (2012). Choosing Among Tucker or Chained Linear Equating in Two Testing Situations: Rater Comparability Scoring and Randomly Equivalent Groups With an Anchor. *Journal of Educational Measurement, 49(3*), 312-329. doi: 10.1111/j.1745-3984.2012.00177.x

Raykov, T (2010) Test Equating Under the NEAT Design: A Necessary Condition for Anchor Items. *Measurement, 8*, 16–20. doi: 10.1080/15366361003684687

Santana, Ana Cristina (2009). *Efecto de la Razón de Tamaños Sobre la Detección del Funcionamiento Diferencial del Ítem Mediante Regresión Logística*. (Tesis inédita de maestría en psicología). Universidad Nacional de Colombia, sede Bogotá.

Sinharay, S., Holland, P (2007). Is It Necessary to Make Anchor Tests Mini-Versions of the Tests Being Equated or Can Some Restrictions Be Relaxed? *Journal of Educational Measurement, 44(3), 249–275*. Recuperado de http://edmeasurement.net/Equating/Sinharay%20Holland%202007%20anchor%20test.pdf

Soler M. P. (2013). Diseño *de un Banco de Ítems para evaluar Comprensión de Lectura en personas con Limitación Visual*. (Tesis inédita de maestría en psicología). Universidad Nacional de Colombia, Sede Bogotá.

Suh, Y., Mroch, A., Kane M., Ripkey, D. (2009). An Empirical Comparison of Five Linear Equating Methods for the NEAT Design. *Measurement, 7,* 147-173. doi:10.1080/15366360903418048

Tsai, T., Hanson, B., Kolen, M., Forsyth, R. (2001). A Comparison of Bootstrap Standard Errors of IRT Equating Methods for the Common-Item Nonequivalent Groups Design. *Applied Measurement in Education, 14 (1),* 17-30. doi:10.1207/S15324818AME1401_03

Van Der Linden, W, J. (2013). Some Conceptual Issues in Observed-Score Equating. *Journal of Educational Measurement, 50 (3),* 249 – 285. doi:10.1111/jedm.12014

Von Davier, A. A., Kong, N. (2005). A Unified Approach to Linear Equating for the Nonequivalent Groups Design. *Journal of Educational and Behavioral Statistics, 30(3),* 313-342. doi: 10.3102/10769986030003313

Zhu, W. (1998). Test Equating: What, Why, How?, *Research Quarterly for Exercise and Sport, 69:1*,11-23, doi:10.1080/02701367.1998.10607662

Zu, J; Yuan, K. (2012). Standard Error of Linear Observed-Score Equating for the NEAT Design With Nonnormally Distributed Data. *Journal of Educational Measurement 49(2*), 190–213. doi:10.1111/j.1745-3984.2012.00168.x

ANEXO 1

Estimaciones de dificultad de los ítems que conformaron las dos formas de la Subprueba de Lenguaje

Ítem	Forma	Estimación de dificultad	
		Grupo sin limitación visual	Grupo con limitación visual
1	1	0,72	0,14
2	1/2	0,49/0,50	1,50/0,67
3	2	-0,59	0,00
4	2	1,25	0,49
5	1/2	-1,25/-1,22	-1,69/-1,45
6	1	1,4	1,18
7	2	0,52	0,58
8	1	1,86	2,48
9	1/2	-0,61/-0,59	-1,11/-0,31
10	1/2	0,10/0,13	0,24/0,32
11	2	0,73	0,97
12	1/2	0,5/0,51	0,04/-0,08
13	1	0,44	0,24
14	2	-0,05	-0,31
15	1/2	0,81/0,83	0,55/0,97
16	1/2	-1,13/-1,08	-1,01/-1,01
17	2	-1,75	-1,36
18	1	0,01	0,14
19	1	0,02	-0,15
20	2	0,28	0,24
21	1/2	0,02/0,03	-0,24/ 0,00
22	2	-0,31	0,00
23	1	0,52	-0,15
24	1/2	-0,05/-0,02	-0,24/0,24
25	1/2	-0,39/-0,45	0,34/-1,42
26	1/2	-0,66/-0,65	-1,22/-0,57
27	2	1,63	1,76
28	1	-0,62	-0,05
29	1	-0,19	-0,05
30	2	0,21	0,16
31	1	-1,72	-1,11
32	1	-0,26	-0,15
33	2	-0,49	-0,46
34	1	-0,57	-0,34
35	2	0,05	-0,39
36	1/2	0,55/0,56	0,66/0,97