Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Equiparación de puntuaciones de las versiones alternativas de la prueba de Matemáticas del Examen de Ingreso a la Educación Superior

TESIS

Que para obtener el grado de

MAESTRA EN CIENCIAS EDUCATIVAS

Presenta

Erika Álvarez Álvarez

Director de tesis

Dr. Joaquín Caso Niebla

Ensenada, B.C., México, diciembre de 2018

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO EDUCATIVO

Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Maestría en Ciencias Educativas

Equiparación de puntuaciones de las versiones alternativas de la prueba de Matemáticas del Examen de Ingreso a la Educación Superior

TESIS

Que para obtener el grado de

MAESTRA EN CIENCIAS EDUCATIVAS

Presenta

Erika Álvarez Álvarez

APROBADO POR:

Dr. Joaquín Caso Niebla
Director de tesis

Mtro. Carlos David Díaz López
Sinodal

Dr. Juan Carlos Pérez Morán
Sinodal

Maestría en Ciencias Educativas

Ensenada, B.C., a 15 de noviembre de 2018

ASUNTO: Voto aprobatorio sobre trabajo de tesis de grado de Maestría.

Dr. José Alfonso Martínez Jiménez
Coordinador de la Maestría en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. ERIKA ÁLVAREZ ÁLVAREZ** para poder presentar la defensa de su examen y obtener el grado de Maestría en Ciencias Educativas, me permito comunicarle que he dado mi VOTO APROBATORIO, sobre su trabajo intitulado:

Equiparación de puntuaciones de las versiones alternativas de la prueba de Matemáticas del Examen de Ingreso a la Educación Superior

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dr. Joaquín Caso Niebla

Ensenada, B.C., a 15 de noviembre de 2018

ASUNTO: Voto aprobatorio sobre trabajo de tesis de grado de Maestría.

Dr. José Alfonso Martínez Jiménez
Coordinador de la Maestría en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. ERIKA ÁLVAREZ ÁLVAREZ** para poder presentar la defensa de su examen y obtener el grado de Maestría en Ciencias Educativas, me permito comunicarle que he dado mi VOTO APROBATORIO, sobre su trabajo intitulado:

Equiparación de puntuaciones de las versiones alternativas de la prueba de Matemáticas del Examen de Ingreso a la Educación Superior

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Mtro. Carlos David Díaz López

Ensenada, B.C., a 15 de noviembre de 2018

ASUNTO: Voto aprobatorio sobre trabajo de tesis de grado de Maestría.

Dr. José Alfonso Martínez Jiménez
Coordinador de la Maestría en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. ERIKA ÁLVAREZ ÁLVAREZ** para poder presentar la defensa de su examen y obtener el grado de Maestría en Ciencias Educativas, me permito comunicarle que he dado mi VOTO APROBATORIO, sobre su trabajo intitulado:

Equiparación de puntuaciones de las versiones alternativas de la prueba de Matemáticas del Examen de Ingreso a la Educación Superior

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dr. Juan Carlos Pérez Morán

A Tavo

No podré rendirme nunca, porque desde el cielo él me mira

Agradecimientos

Gracias al Dr. Joaquín Caso Niebla, director de esta tesis, por la orientación, el apoyo y la guía en todo momento.

Al Mtro. Carlos David Díaz López, por la orientación constante y la disposición para apoyar con su conocimiento y experiencia durante el desarrollo de la presente investigación.

A la Dra. Rosario Martínez Arias, por sus aportaciones, observaciones y sugerencias que fueron determinantes para la construcción de la tesis.

Al Dr. Juan Carlos Pérez Morán, por la orientación y apoyo para lograr este proyecto.

A los todos los docentes e investigadores del Instituto de Investigación y Desarrollo Educativo (IIDE), en especial a la Dra. Alicia Chaparro, al Dr. Javier Organista, la Dra. Guadalupe Tinajero, la Dra. Graciela Cordero, la Dra. Edna Luna, la Dra. Maricela López, y al Dr. Alfonso Jiménez, por compartir su experiencia, conocimiento y tiempo y con ello enriquecer mi formación académica.

Al personal del IIDE, por facilitar cada proceso y por alegrar el día con su presencia.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo financiero.

Índice de contenido

Resumen 1
Introducción 2
Planteamiento del problema 4
Objetivos 7
Objetivo general. 7
Objetivos específicos: 7
Justificación 8
Marco teórico 10
Exámenes de selección en el ingreso a la Educación Superior 10
Equiparación de puntuaciones 14
Diseños de recopilación de datos en procedimientos de equiparación. 18
Procedimientos de equiparación 23
El Examen de Ingreso a la Educación Superior (ExIES) 31
Método 36
Participantes 36
Instrumento 37
Análisis de datos 38
Fase 1. Análisis de la calidad métrica de los ítems. 38
Fase 2. Equiparación 41
Resultados 44
Análisis de la calidad métrica de los ítems 44
Análisis de la equiparación de puntuaciones 49
Discusión 56
Referencias 64
Apéndices 72

Índice de tablas

Tabla 1. Promedio de los índices de validez y confiabilidad del ExIES 34

Tabla 2. Distribución de ítems en el ExIES..... 35

Tabla 3. Distribución de frecuencias por sexo 36

Tabla 4. Proporción de ítems de la prueba de Matemáticas del ExIES..... 37

Tabla 5. Criterios de interpretación para el coeficiente de discriminación 39

Tabla 6. Propiedades psicométricas de la prueba de Matemáticas..... 45

Tabla 7. Coeficiente de discriminación de los ítes de anclaje 46

Tabla 8. Índice de dificultad de los ítems de anclaje..... 47

Tabla 9. Ítems que presentan DIF en función del sexo..... 49

Tabla 10. Paramétro de dificultad e índice de ajuste de ítems ancla bajo el modelo de Rasch 50

Tabla 11. Proporción de ítems ancla e ítems únicos por nivel de dificultad y modelo de estimación..... 51

Tabla 12. Estadísticos y prueba t de muestras relacionadas. 54

Tabla 13. Clasificación de aspirantes por cuartiles en función del modelo de estimación . 55

Índice de figuras

Figura 1. Diseños de recopilación de datos en procedimientos de equiparación, basado en la clasificación de Kolen y Brennan (2014). 22

Figura 2. Evidencia de unidimensionalidad por versión, mediante el análisis paralelo..... 48

Figura 3. Distribución de las puntuaciones bajo el modelo de la TCT y la TRI por sujeto. 52

Figura 4. Distribución de la puntuación de los sujetos por modelo de estimación. 53

Resumen

El objetivo de la presente investigación fue equiparar las puntuaciones de la prueba de matemáticas de las versiones alternativas del Examen de Ingreso a la Educación superior (ExIES). El proceso metodológico fue con base en los criterios para equiparación de puntuaciones propuestos por Kolen y Brennan (2014). Para cumplir dicho objetivo se trabajó con los resultados de las cuatro versiones de la prueba de matemáticas del ExIES, aplicadas a los 2898 aspirantes a los programas de licenciatura de la Universidad Autónoma de Baja California en la convocatoria 2018-1. El diseño para la selección de la muestra fue de grupos no equivalentes con ítems de anclaje y se aplicó el procedimiento de calibración concurrente para equiparar las puntuaciones. El análisis de datos se dividió en dos fases, la primera consistió en estimar las propiedades de calidad psicométrica de cada versión, así como la unidimensionalidad y la propiedad de invarianza de los ítems. En la segunda fase, se estimó el parámetro de dificultad con TRI y se compararon los resultados de las estimaciones con TCT y las obtenidas con TRI. Los resultados de la investigación destacan la importancia de equiparar las puntuaciones de las diferentes versiones, para asegurar una selección justa y equitativa de los aspirantes.

Palabras clave: Equiparación, calibración concurrente, ExIES, versiones alternativas y selección de aspirantes.

Introducción

Los exámenes de selección a las instituciones de educación superior son considerados instrumentos de alto impacto, por lo que su diseño, desarrollo y validación debe realizarse en apego a los estándares de calidad para pruebas educativas y psicológicas que dictan diferentes organismos internacionales. En este sentido, el examen de selección a los programas de licenciatura de la Universidad Autónoma de Baja California presenta cuatro diferentes versiones, de manera que compromete la aplicación de procedimientos de equiparación de los puntajes para asegurar una evaluación justa a los aspirantes. Por tal motivo, el objetivo de este trabajo de investigación fue equiparar los puntajes de la prueba de matemáticas de las cuatro versiones alternativas del Examen de Ingreso a la Educación Superior (ExIES).

Para documentar el desarrollo de la investigación se presenta el siguiente documento, organizado de la siguiente manera:

En el primer apartado se describe el problema de investigación que fundamenta el presente estudio, así como los objetivos y la justificación para llevarlo a cabo.

Posteriormente, se presenta el marco teórico que sustenta la investigación. En la primera parte se describe el uso de los exámenes de selección a las instituciones de educación superior (IES). Posteriormente, se describe el concepto, los diseños y los procedimientos de equiparación de puntuaciones; y finalmente se presentan las generalidades del ExIES.

En el siguiente apartado se presenta el método, en el cual se describen los participantes, el instrumento y el procedimiento para analizar los datos a fin de cumplir el objetivo planteado.

Después del método, se describen los resultados de los análisis de datos. Primero se presentan los resultados de la calidad métrica de los ítems para cada versión del examen. Posteriormente, se analiza la equiparación de las puntuaciones y al final se comparan ambas estimaciones.

En el último apartado se presenta una discusión a los resultados obtenidos, así como la mención de las limitaciones y las recomendaciones asociadas a los mismos; y finalmente, se adjuntan los apéndices que presentan los resultados de los análisis estadísticos y complementan la presentación de dicho apartado.

Planteamiento del problema

La selección de aspirantes para el ingreso a los programas educativos de la mayoría de las Instituciones de Educación Superior (IES) está basada en distintos mecanismos de admisión que se apoyan, principalmente, en la aplicación de exámenes estandarizados y a gran escala. Dichos mecanismos tienen como objetivo asegurar que los espacios disponibles en las IES sean ocupados por los aspirantes que demuestren mayor probabilidad de éxito al cursar la carrera (Backhoff, 2001).

En el caso particular de México, los principales exámenes de admisión a las IES son el Examen Nacional de Ingreso a la Educación Media Superior (EXANI-II) y el Examen de Habilidades y Conocimientos Básicos (EXHCOBA). El EXANI-II evalúa conocimientos y habilidades numéricas, verbales y no verbales en las áreas de pensamiento matemático, pensamiento analítico, estructura de la lengua y comprensión lectora. Se considera una prueba de alto valor predictivo sobre el desempleo académico del sustentante en el primer ciclo de la educación superior (CENEVAL, 2017).

Por su parte, el EXHCOBA es un examen en formato computarizado que evalúa habilidades y conocimientos básicos en diferentes áreas de conocimiento: económico-administrativas, químico-biológicas, salud, ingeniería, físico-matemáticos, humanidades y ciencias sociales. El principal objetivo del EXHCOBA es seleccionar a los mejores estudiantes para ingresar al nivel medio superior y superior (Métrica Educativa, 2017).

Tanto el EXANI- II como el EXHCOBA se caracterizan por ser instrumentos estandarizados que cumplen con los criterios de confiabilidad y validez recomendados por estándares propios de la evaluación educativa. Asimismo, suelen clasificarse como exámenes

de alto impacto, puesto que el uso de sus puntuaciones conlleva a decisiones positivas o negativas sobre el futuro profesional de los examinados.

En este sentido, las evaluaciones de esta naturaleza suelen aplicar distintas versiones del mismo examen con el fin de asegurar el contenido de la prueba a través de diversos diseños. En el caso del EXANI-II, anualmente se elaboran más de 70 versiones, compuestas cada una por distintas preguntas y opciones de respuesta, esto con la finalidad de cuidar que los sustentantes que responden el examen en diferentes aplicaciones no memoricen las preguntas (CENEVAL, 2017). En tanto, en el EXHCOBA cada aspirante responde una versión distinta de la prueba a través del empleo de un generador automático de ítems, de manera que en cada aplicación se generan cientos de versiones diferentes del examen (Antillón, Larrazolo y Backhoff, 2008; Métrica Educativa, 2018).

Si bien la aplicación de diferentes versiones de un examen contribuye a lograr un mejor funcionamiento del instrumento y permite evaluar a los aspirantes sin necesidad de aplicar a todos los mismos ítems o de evaluarlos en distintos momentos, las puntuaciones obtenidas deben expresarse en una misma escala, por lo tanto, deben emplearse técnicas estadísticas que, bajo un modelo psicométrico determinado, generen un puntaje equivalente y comparable para todas las versiones (ITC, 2013; Kolen y Brennan, 2014; Lozzia, Abal, Blum, Aguerri, Galibert y Attorresi, 2015).

En este sentido, la International Test Commission (ITC, 2013) señala que cuando las puntuaciones se obtienen de diferentes versiones de una misma prueba es necesario equiparar los puntajes, es decir, ubicarlos en una misma escala. Para lograr la equiparación se debe aplicar un conjunto de procedimiento estadísticos con el objetivo de ajustar las puntuaciones de las diferentes versiones para comparar los puntajes de manera indistinta, ya que se

encuentran midiendo lo mismo (González y Wiberg, 2017; Kolen y Brennan, 2014). Existen diversos procedimientos para lograr la equiparación de puntuaciones, cuya aplicación dependerá de los elementos técnicos asociados al diseño de selección de la muestra y a los procedimientos estadísticos para equiparar las versiones (Gempp, 2010).

Con respecto a los análisis de equiparación de los exámenes antes mencionados, los reportes técnicos documentan evidencia de su calidad psicométrica, más no documentan información acerca de los procedimientos de equiparación de las puntuaciones de sus diferentes versiones, de manera que sea posible asegurar justicia y equidad a los examinados.

De manera particular, la Universidad Autónoma de Baja California (UABC) desarrolló un nuevo examen de selección de aspirantes para sus programas de licenciatura, denominado Examen de Ingreso a la Educación Superior (ExIES), el cual está compuesto por tres pruebas: lengua, lengua escrita y matemáticas. Este examen tiene como objetivo medir la capacidad que tienen los aspirantes para aplicar los conocimientos y habilidades que poseen y que serán requeridos para atender con éxito las demandas propias de su formación universitaria (Caso, Díaz, Castro y Martínez, 2017). Al igual que los exámenes anteriores, el ExIES es una prueba de alto impacto, y emplea un diseño de aplicación de distintas versiones construidas con el mismo contenido, especificaciones y niveles de demanda cognitiva.

La aplicación de diferentes versiones de este examen exigirá del empleo de métodos estadísticos para equiparar los puntajes y con ello proporcionar evidencia de que las distintas versiones del ExIES se encuentran midiendo lo mismo, asegurando con ello equidad y justicia para los examinados.

Objetivos

Objetivo general.

Equiparar las puntuaciones de la prueba de Matemáticas de las versiones alternativas del Examen de Ingreso a la Educación Superior.

Objetivos específicos:

1. Analizar las propiedades psicométricas de la prueba de Matemáticas de cada versión del examen, bajo el modelo de la Teoría Clásica de los Test (TCT).
2. Obtener evidencias de validez de la estructura interna de la prueba de Matemáticas de cada versión del examen.
3. Analizar la invarianza poblacional de los ítems de la prueba de Matemáticas en cada versión del examen.
4. Calibrar el parámetro de dificultad de los ítems de las cuatro versiones bajo el modelo de Rasch y estimar la puntuación en la prueba.
5. Comparar los puntajes calibrados con los analizados en TCT y Teoría de Respuesta al Ítem (TRI).

Justificación

Uno de los principales problemas al que se enfrentan las IES está asociado con la incapacidad de atender la creciente demanda de estudiantes que buscan un espacio para realizar sus estudios universitarios, por lo que los exámenes de selección representan una alternativa que proporciona certidumbre, transparencia e imparcialidad a este proceso (González, 2007). Los resultados de la investigación en esta materia refieren que los puntajes en estos exámenes se asocian al rendimiento académico futuro, lo que sugiere que una selección adecuada de aspirantes traerá consigo bajos índices de abandono escolar y, por lo tanto, reducción de los costos sociales asociados a dicha deserción (González, 2007), además, los resultados que obtienen los aspirantes en los exámenes de selección determinan, en muchos casos, la toma de decisiones de las autoridades educativas en torno al cupo de la institución, la oportunidad de obtener becas, el diagnóstico de los individuos en función de sus habilidades, aptitudes y conocimientos, entre otras (Pacheco-Villamil, 2007).

Por lo tanto, contar con exámenes de admisión válidos y confiables para evaluar las habilidades y competencias de los aspirantes resulta de gran relevancia, pues las decisiones que de estos se desprenden tienen un impacto en los aspirantes, la institución y en los programas educativos (Antillón, Larrazolo y Backhoff, 2008). En este sentido, el desarrollo del ExIES debe apegarse a estándares de calidad establecidos por organismos como la APA, la AERA, el NCME y la ITC, de manera que compromete el diseño y desarrollo de cuatro versiones diferentes.

El aplicar diferentes versiones del examen, genera beneficios a la validez de este, ya que proporciona mayor seguridad del contenido al controlar la ventaja que puede tener un aspirante que responde el examen en dos o más ocasiones, se disminuye la posibilidad de copia en la misma aplicación, asimismo, posibilitan el seguimiento de los estudiantes a largo plazo y facilitan la aplicación de las pruebas en diferentes días (Kolen y Brennan, 2014).

Así la aplicación de diferentes versiones conduce a la necesidad de ajustar los puntajes de cada una, de manera que el contenido y la dificultad de este sea similar en todas. Para lograr ajustar las posibles diferencias se aplican procedimientos de equiparación de las puntuaciones para igualar con éxito las pruebas y utilizarlas indistintamente (Kolen y Brennan, 2014). En este sentido, el asegurar la equivalencia métrica de las distintas versiones del ExIES, forma parte de las exigencias técnicas del examen y de los mecanismos que proporcionan justicia y equidad a los examinados, garantizando que todos hayan sido evaluados de la misma forma. Aunado a ello, los análisis de calidad métrica que compromete el procedimiento de equiparación también aportan evidencia de validez al examen.

La ITC (2013) señala que se debe asegurar que el diseño y los procedimientos de equiparación de puntuaciones se han realizado correctamente, además de documentar el proceso, los efectos de los ítems y la consistencia de las puntuaciones en todas sus versiones. De manera que, el análisis de las condiciones previas de equiparación generara, además, información sobre la calidad métrica de cada versión. Por lo tanto, en el presente estudio se documenta la aplicación del procedimiento de equiparación del ExIES a fin de aportar evidencia que sustente la importancia de este tipo de prácticas en los procesos de selección de aspirantes a la educación superior.

Marco teórico

En el siguiente apartado se presenta una descripción del uso de los exámenes de selección en los procesos de admisión, principalmente en las Instituciones de Educación Superior (IES), así como los principales exámenes de admisión aplicados en México. Después se describen los conceptos asociados, sus principales características, los diseños empleados para la selección de la muestra y los diferentes procedimientos de equiparación. Finalmente, se presenta el Examen de Ingreso a la Educación Superior (ExIES), sus etapas de desarrollo, los resultados del análisis primario de sus ítems y el diseño de sus diferentes versiones.

Exámenes de selección en el ingreso a la Educación Superior

En la actualidad, la gran mayoría de las IES han desarrollado mecanismos específicos de selección que representan un filtro para elegir a los estudiantes que ingresan a este nivel educativo. Estos mecanismos incluyen una serie de factores que las universidades utilizan como criterios de admisión, entre otros factores se encuentran, el promedio del bachillerato, entrevistas, cartas de recomendación, actividades extracurriculares y exámenes de selección (Noble y Wayne, 2003).

Los exámenes de selección son aplicados, principalmente, en los procesos de ingreso a los programas educativos del nivel medio superior y superior, y tienen como principal objetivo medir los conocimientos que tiene un individuo para cursar una carrera profesional con éxito (Kolen y Brennan, 2014). Los *Standards for Educational and Psychological Testing* (APA, AERA, NCME, 2014) definen los exámenes de selección como pruebas que evalúan un determinado dominio a partir de las puntuaciones obtenidas, por medio de un

proceso estandarizado para la selección o rechazo de aspirantes. Los resultados de este tipo de pruebas, tienen consecuencias importantes y directas para los individuos, los programas o las instituciones (Sánchez-Mendiola y Delgado-Maldonado, 2016), siendo útiles como mecanismo de predicción del rendimiento académico posterior al primer año de la universidad (Zwick, 2007).

Entre las principales ventajas de aplicar exámenes de selección se encuentran las siguientes: permiten discriminar a los aspirantes con base en distintos niveles de desempeño, permiten identificar diferencias en los currículos de procedencia, ayudan a obtener información sobre la capacidad que denota el aspirante y su distancia con respecto a los estándares exigidos por la institución, y fomentan la transparencia asociada con los procesos de selección (Edwards, Coates y Friedman, 2012). Por el contrario, de acuerdo con Sackett, (2005, en Edwards, Coates y Friedman, 2012), entre las principales objeciones para aplicar exámenes de selección se encuentran las siguientes: insuficiente evidencia de validez de las pruebas, sesgo cultural o de contenido, carga administrativa considerable para las instituciones implicadas, y gastos adicionales para los aspirantes.

Existe una lista considerable de exámenes utilizados por diferentes universidades para el ingreso a sus programas educativos. En Estados Unidos se aplica el *Scholastic Aptitude Test* (SAT), el cual evalúa lectura, matemáticas y escritura (College Board, 2016), y el *American College Testing* (ACT), que evalúa el dominio de inglés, matemáticas, lectura, ciencia y escritura (ACT, 2016). Los resultados de ambos exámenes son aceptados por todas las universidades en Estados Unidos.

Específicamente, en México se identifican un conjunto de exámenes que son aplicados por diferentes universidades como criterio de selección de ingreso. El Examen

Nacional de Ingreso a la Educación Superior (EXANI-II) evalúa aptitudes y competencias disciplinares. Este examen se compone de dos partes, la primera compuesta por 110 reactivos que evalúan pensamiento matemático, pensamiento analítico, estructura de la lengua y comprensión lectura. La segunda parte del examen agrupa 90 reactivos que permite realizar un diagnóstico de las competencias disciplinares del aspirante según el área profesional al que desea ingresar (CENEVAL, 2017).

Otro de los exámenes de mayor relevancia a nivel nacional es el Examen de Habilidades y Conocimientos Básicos (EXHCOBA), este examen tiene como objetivo seleccionar a los mejores estudiantes a ingresar al nivel medio superior y superior, diagnosticar las habilidades y conocimientos básicos del estudiante y detectar problemas en su formación básica (Métrica Educativa, 2017). En el caso del examen aplicado a nivel superior, este, evalúa siete áreas de conocimiento: Económico-administrativo, químico-biológico, salud, ingeniería, físico-matemático, humanidades y ciencias sociales.

En el caso particular de la Universidad Autónoma de Baja California (UABC), se diseñó y desarrolló un examen de selección, que es utilizado en sus procesos de admisión desde el año 2017. El objetivo del Examen de Ingreso a la Educación Superior (ExIES) es medir la capacidad que tienen los aspirantes para aplicar los conocimientos y habilidades que tienen y que serán requerido para atender con éxito las demandas propias de su formación universitaria (Caso, Díaz, Castro y Martínez, 2017).

Los exámenes antes mencionados comparten algunas características importantes. Los tres son considerados exámenes estandarizados de alto impacto, evalúan habilidades y conocimientos en diferentes áreas, sus puntuaciones son normativas y se aplican diferentes versiones de cada prueba. En lo relativo a esta última característica, y en consonancia con los

estándares de calidad de las pruebas psicológicas y educativas, cuando se aplican diferentes versiones del mismo examen es necesario asegurar la equivalencia entre dichas versiones (Kolen y Brennan, 2014; ITC, 2013). Para ello resulta necesario aplicar procedimientos de equiparación de puntuaciones, de manera que los puntajes de los examinados puedan considerarse de manera indistinta como si procedieran de una misma prueba (González y Wiberg, 2017).

La importancia de diseñar diferentes versiones de una misa prueba adquiere mayor relevancia en situaciones en las que los exámenes de selección para el ingreso a la universidad se aplican en un periodo de varios días o se registran varias aplicaciones por día, lo que pudiera presentar ciertas ventajas a algunos aspirantes por encima de otros. Por tal motivo, el contenido del examen y sus ítems se encuentran expuestos a que el examinado informe a otros sobre el contenido de estos, o bien que el sustentante obtenga ventaja sobre otros al aplicar el examen en varias ocasiones, cuestiones que atentan contra los mecanismos de seguridad requeridos en este tipo de pruebas (Kolen y Brennan, 2014).

Para corregir este tipo de problemas es común que en el diseño de los exámenes se considere el desarrollo de distintas versiones orientadas a medir el mismo atributo. De esta manera, al aplicar distintas versiones es común observar diferencias en las propiedades psicométricas, particularmente en lo relativo al índice de dificultad, lo que puede ocasionar que, al momento de interpretar las puntuaciones de los grupos que han respondido diferentes versiones del mismo examen, una versión resulte más fácil o difícil que otra, lo que sin duda pondría en ventaja a un grupo sobre otro, lo cual resultaría injusto (González y Wiberg, 2017).

En virtud de lo anterior, resulta importante resaltar la necesidad de aplicar procedimientos de equiparación de puntuaciones cuando se aplican diferentes versiones de una misma prueba, a manera de asegurar que dichas puntuaciones pueden considerarse de manera indistinta, como si procedieran de una misma versión (González y Wiberg, 2017), por lo tanto, a continuación de describe el concepto de equiparación, los diseños y procedimientos estadísticos asociados a este.

Equiparación de puntuaciones

La equiparación pertenece al grupo de métodos asociados al tema de vinculación de puntuaciones, entendiendo el término de vinculación como la transformación de la puntuación de una prueba a la de otra prueba. Estos métodos de vinculación, de acuerdo con Hollands y Dorans (2006) se dividen en tres categorías básicas: predicción, alineación de escala y equiparación.

La predicción es la forma más antigua de vincular puntuaciones, tiene como objetivo predecir información de un examinado a partir de la puntuación obtenida en una prueba, esta información puede ser de tipo demográfica o bien su puntuación en otra prueba, en otras palabras, sería predecir Y a partir de X.

Por su parte, la alineación de escala tiene como objetivo transformar las puntuaciones de dos pruebas distintas en una escala en común y compararlas, es decir, implica una vinculación directa de las puntuaciones en X e Y. La alineación de escala se subdivide en dos tipos diferentes de alineación, uno que es utilizado cuando las pruebas que se vinculan

miden diferentes constructos y otro que es aplicado cuando se miden constructos similares, pero con distintas pruebas.

Finalmente, la equiparación es entendida como el intercambio de las puntuaciones de la prueba X a la prueba Y. El proceso de equiparación resulta un elemento indispensable para garantizar la comparación de las puntuaciones que se obtienen en pruebas diferentes que están midiendo el mismo rasgo o constructo (Navas, 2000). Por tal motivo, este término ocupa un lugar central en el presente estudio y se hacen las siguientes precisiones conceptuales.

De acuerdo con Kolen y Brennan (2014), la equiparación es definida como el conjunto de modelos y procedimientos estadísticos que se aplican para ajustar las puntuaciones de las diferentes versiones de una misma prueba, de manera que dicho puntaje pueda utilizarse de manera intercambiable. Este proceso tiene como principal objetivo ajustar las diferencias de las puntuaciones para que sean similares tanto en contenido como en dificultad (Kolen y Brennan, 2014).

Cabe mencionar que las diferencias de los puntajes de las pruebas no siempre son debido a las diferencias en la dificultad de estas. Es posible que los examinados sean diferentes entre sí. Por ejemplo, en su nivel de habilidad para responder la prueba. Por ello, el proceso de equiparación implica el ajuste de las puntuaciones para que las diferencias de dificultad puedan compensarse, una vez que se ha asegurado que ambas versiones sean paralelas tanto en contenido como en número de ítems (González y Wiberg, 2017).

Al respecto, Martínez, Lloreda y Lloreda (2007) señalaron que, como paso previo al proceso de equiparación de las diferentes versiones de una prueba, se debe asegurar que éstas se hayan construido en forma similar, particularmente en lo relativo al número y formato de

ítems, a la habilidad a evaluar, al nivel de demanda cognitiva para responder, así como a las condiciones que acompañaron su aplicación.

El proceso de equiparación se basa en gran medida en la adecuada administración de las pruebas, por lo tanto, el proceso de diseño y desarrollo, la propia aplicación y los procedimientos estadísticos deben encontrarse coordinados. Al respecto, Kolen y Brennan (2014) indican los siguientes pasos para implementar procesos de equiparación: (a) Definir el objeto de la equiparación; (b) Construir las formas alternativas, bajo el mismo contenido y las mismas especificaciones estadísticas; (c) Seleccionar un diseño para la recopilación de datos; (d) Aplicar el diseño de recopilación de datos, basado en las especificaciones del diseño seleccionado; (e) Elegir una o más definiciones operativas de equiparación; (f) Elegir los métodos de estimación estadística; y (g) Evaluar los resultados del proceso de equiparación, considerando aspectos asociados al diseño y administración de la prueba, a los procedimientos estadísticos empleados y a las propiedades de equivalencia observadas.

Von Davier (2011) señaló que el proceso de equiparación de puntuaciones sigue los mismos pasos que un proceso de modelado estadístico común. Es decir, se comienza con una pregunta de investigación y un conjunto de datos, se delimitan los procedimientos de muestreo, el diseño de recogida de datos y la modelización explícita de las variables. Posteriormente se propone un modelo estadístico y se valora su ajuste a los datos, se calculan los parámetros de dicho modelo, se realizan inferencias basadas en el modelo, y se evalúan los resultados en torno al error y sesgo de muestreo.

Además de los requisitos previos necesarios para aplicar procedimientos de equiparación, existen una serie de precisiones que deben cumplirse para asegurar que las puntuaciones de las pruebas sean equivalentes (Holland y Dorans, 2006; Von Davier, 2011):

a. Mismo constructo. Las versiones alternativas deben medir el mismo constructo. De manera que es necesario presentar evidencia de que han sido construidas con base a la misma tabla de especificaciones y que se encuentran midiendo el mismo rasgo latente (Dorans, Pommerich y Holland, 2007). Para analizar si las versiones presentan esta propiedad se analiza la estructura interna de la prueba, realizando estudios sobre la dimensionalidad del instrumento. Dichos estudios pueden realizarse mediante técnicas factoriales, en las cuales se analiza la relación entre los ítems y los factores o dimensiones subyacentes (Abad, Olea, Ponsoda y García, 2011).

b. Confiabilidad. Las versiones deben tener el mismo nivel de confiabilidad (Dorans, Pommerich y Holland, 2007). Desde el modelo clásico, la confiabilidad puede definirse como la proporción de varianza verdadera con relación a la varianza total observada. En este sentido, la confiabilidad informa cómo varía la puntuación de una persona en distintas aplicaciones (Abad, et al., 2011). El coeficiente de confiabilidad más común es obtenido mediante la estimación del coeficiente alfa, que parte de la matriz de correlaciones o de covarianzas entre ítems (Oliden y Zumbo, 2008).

c. Simetría. Las puntuaciones obtenidas deben ser intercambiables, es decir que la escala de puntuaciones de la prueba X sea la misma en la escala de la prueba Y. Un procedimiento para cumplir esta propiedad es mediante una función lineal estimando la media y la desviación estándar (Von Davier, 2011).

d. Equidad. Las puntuaciones de los examinados no deben depender de la prueba que respondan sino de su nivel de habilidad para responder. La condición de equidad de las versiones garantiza una evaluación justa, para ello es necesario definir correctamente las especificaciones de la prueba en términos de contenido, demanda cognitiva y nivel de dificultad (Von Davier, 2011).

e. Invariancia poblacional. Las puntuaciones de la prueba X deben ser las mismas que las de la prueba Y, independientemente de la población de la que se obtienen. Cuando las pruebas son aplicadas en subpoblaciones se afecta, hasta cierto grado, la función de igualación (Von Davier, 2011). Al respecto de esta propiedad, Abad, et.al (2011), señalan que es preciso comprobar la invarianza factorial de una prueba, cuando esta ha sido aplicada a diferentes muestras. Es procedimiento consiste asegurar que las puntuaciones en los ítems no están sesgadas, es decir, que presentan ausencia de funcionamiento diferencial del ítem (DIF). El análisis del DIF, contribuye a evaluar las diferencias en los resultados de una prueba aplicada a distintos grupos, asegurando que dichos resultados dependen solo del conocimiento o la habilidad del sujeto y no características personales, como etnia o sexo (Silva y Santelices, 2016). Algunas técnicas clásicas para la detección del DIF son: el procedimiento de Mantel-Haenszel, el modelo de regresión logística y el procedimiento de estandarización (González, 2011).

En suma, el proceso de equiparación de pruebas requiere gran compromiso tanto en el diseño y la administración de la prueba, así como en el conocimiento y aplicación de las técnicas estadísticas necesarias, lo que exige de una integración de elementos prácticos y conocimientos estadísticos (Kolen y Brennan, 2014).

Diseños de recopilación de datos en procedimientos de equiparación.

Los diseños que subyacen a los procesos de equiparación son planes para recopilar los datos que tienen como propósito controlar las diferencias de los grupos de examinados que responden las pruebas. Si el mismo grupo respondiera ambas pruebas se lograría un control

directo sobre las diferencias en la capacidad del examinado para responder, pero al no registrar esta condición normalmente se utilizan dos grupos equivalentes de una población común (Dorans, Moses y Eignor, 2010).

Para realizar la equiparación de puntuaciones de dos pruebas, es necesario vincular los resultados de una prueba con la de otra que sirve como referencia. Existen tres procedimientos para obtenerla: aplicar las dos formas a los mismos examinados; obtener los resultados de las dos formas de dos grupos que son igualados respecto a las habilidades que mide la prueba; y ajustar el nivel de habilidad con base en otro tipo de información de los examinados (Livingston, 2014). Estos procedimientos de equiparación orientan a su vez cuatro diseños básicos de recopilación de datos, cuya diferencia principal se encuentra en la manera de administrar las diferentes versiones a los grupos: (a) diseño de grupo único, (b) diseño de grupo único con contrabalanceo, (c) diseño de grupos aleatorios equivalentes, y (d) diseño de grupos no equivalentes con ítems comunes o de anclaje (Ver figura 1).

Diseño de grupo único. Este diseño consiste en aplicar al mismo grupo de examinados la forma X y la forma Y, por lo que suele tratarse del diseño que menos se utiliza en la práctica. Aun cuando su aplicación se aproxima a lo ideal, una desventaja asociada es que se ven implícitos una serie de factores que pueden afectar los resultados de la forma entre los que se encuentran la fatiga o la familiaridad con el examen (González y Wiberg, 2017; Kolen y Brennan, 2014). En la práctica, resulta difícil asumir los efectos de los supuestos implicados (fatiga, familiaridad y aprendizaje), por lo que se ha preferido aplicar el diseño de grupo único con contrabalanceo (Martínez, Lloreda y Lloreda, 2007).

***Diseño de grupo único con contrabalanceo*.** En este diseño se realiza un procedimiento particular que consiste en contrabalancear el orden en que se administran las

pruebas, tomando una muestra de dos grupos independientes que pertenecen a una población común. Se elaboran cuadernillos que contienen tanto la versión X como la Y, mientras que la otra mitad de los cuadernillos primero presentan la forma Y seguida de la X (Kolen y Brennan, 2014). Este diseño puede eliminar el efecto del orden al aleatorizar la muestra, pero puede presentar problemas asociados con la duplicación del tiempo de administración de la prueba y la posibilidad de que los examinados no completen las dos versiones (Martínez, Lloreda y Lloreda, 2007).

***Diseño de grupos aleatorios equivalentes*.** Este diseño consiste en tomar dos grupos independientes de examinados que pertenezcan a una población común y se les asigna de manera aleatoria la versión X o la Y (González y Wiberg, 2017). Funciona como un proceso en espiral, que consiste en entregar al primer examinado la versión X, al segundo la Y, al tercero la X, y así sucesivamente. La diferencia en el nivel de rendimiento entre grupos se considera un indicador directo de la diferencia en la dificultad de las versiones (Kolen y Brennan, 2014). Otra característica importante de este diseño es que se minimiza el tiempo de aplicación de las pruebas ya que cada examinado responde solo una forma. Por lo general, se necesitan muestras grandes de participantes y se requiere que las pruebas sean aplicadas en forma simultánea, cuidando la forma de ubicar a los examinados a fin de evitar problemas de números impares o que una misma forma sea administrada solo a examinados de un mismo sexo (Kolen y Brennan, 2014).

***Diseño de grupos no equivalentes con ítems comunes o de anclaje*.** El diseño con ítems comunes o de anclaje cosiste en que dos versiones diferentes de una misma prueba comparten un número específico de ítems, que son los mismos en cada versión (Rodríguez,

2016). En este diseño la forma X y la forma Y tienen un conjunto de ítems en común y son aplicados a grupos independientes que pueden pertenecer también a diferentes poblaciones. El diseño de grupos no equivalentes con ítems de anclaje es comúnmente aplicado cuando es necesario administrar ambas versiones en momentos distintos. Por ejemplo, se puede aplicar la versión X en un año y la versión Y en otro año (Kolen y Brennan, 2014).

Una característica importante de este diseño tiene relación con el puntaje de los ítems comunes. Cuando la puntuación de estos ítems contribuye al puntaje total del examinado se conoce como conjunto de ítems interno y, por el contrario, cuando no son tomados en cuenta para asignar la puntuación al examinado, se conoce como conjunto de ítems externo (Dorans, Pommerich y Holland, 2007).

Al respecto, Martínez, Lloreda y Lloreda, (2007), señalan los siguientes elementos que deben tomarse en cuenta para seleccionar los ítems comunes:

a. Se debe incluir un número representativo de los ítems de la forma de referencia, en función de la calidad y el número total de ítems.
b. Los ítems deben ser similares en formato y contenido, así como incluir elementos representativos de todos los rangos de dificultad.
c. Los ítems no deben presentar funcionamiento diferencial.
d. No se permite variación en las cuestiones incluidas en las formas y cuando hay varios ítems que pertenecen a un estímulo en común, debe incluirse el bloque completo (ej. un texto).
e. Los ítems deben ubicarse en posiciones similares en todas las formas.

El uso de grupos no equivalentes provee cierta facilidad en la administración de las pruebas, sin embargo, es necesario aplicar análisis estadísticos robustos que permitan separar las diferencias que se deben a los grupos de aquéllas correspondientes a las formas, por lo que los ítems asumen gran importancia pues representan el único vínculo entre los grupos (Kolen y Brennan, 2014).

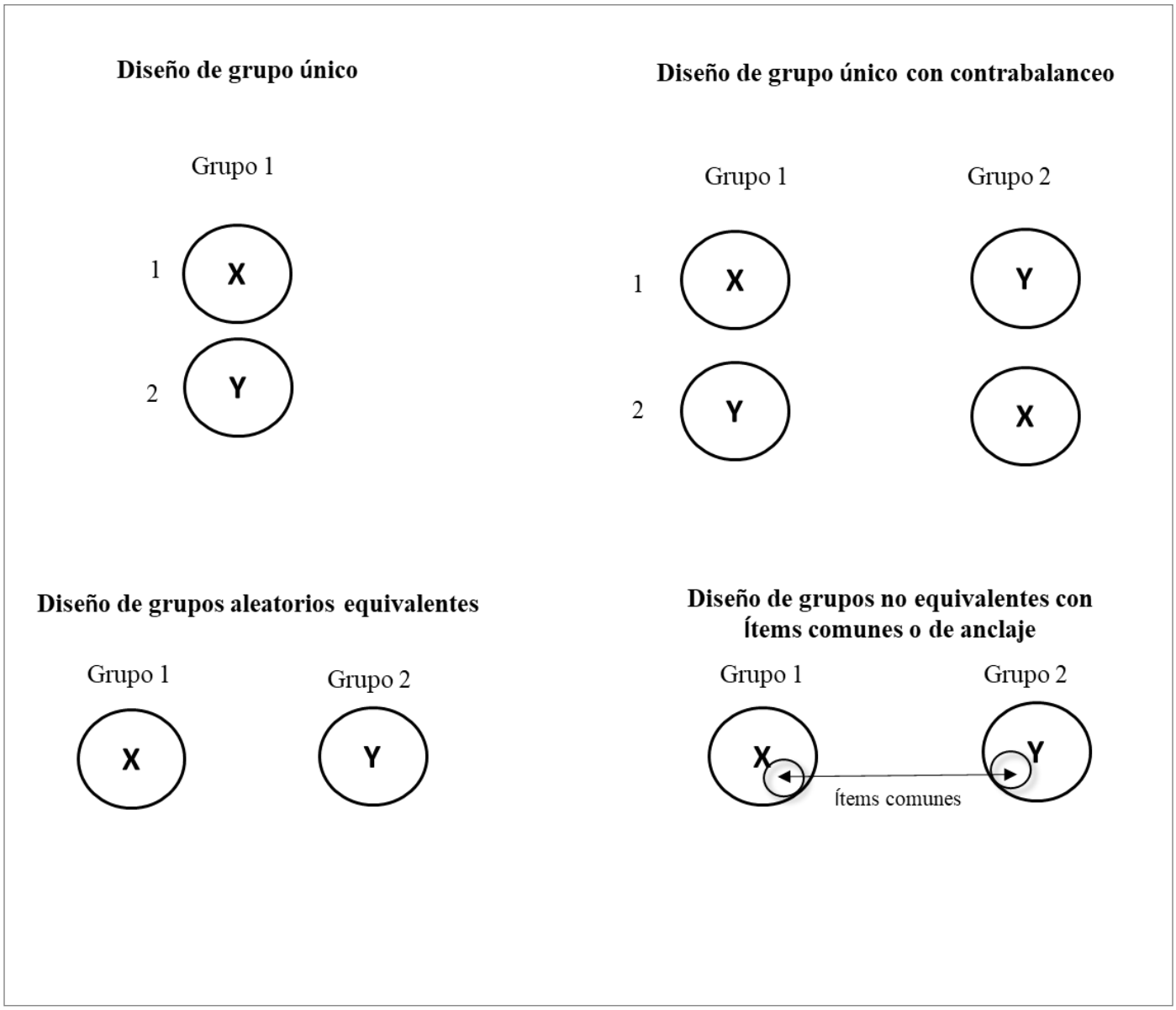

Figura 1. Diseños de recopilación de datos en procedimientos de equiparación, basado en la clasificación de Kolen y Brennan (2014).

Procedimientos de equiparación

Una vez que se ha comprobado que las versiones poseen la misma propiedad de especificaciones, contenido y dificultad es posible equiparar las puntuaciones, para lo cual es necesario aplicar métodos estadísticos que aseguren equivalencia de los puntajes (Kolen y Brennan, 2014). Los procedimientos de equiparación se clasifican de acuerdo con la teoría psicométrica en que están basados y su aplicación depende del diseño de selección de la muestra que se ha utilizado. Kolen y Brennan (2014) los clasifican en procedimientos basados en la Teoría Clásica de los Test (TCT) y los basados en la Teoría de Respuesta al Ítem (TRI).

Procedimientos basados en el Modelo de Teoría Clásica de los Test. Se dividen en dos: el método lineal y el método equipercentil.

Método lineal. Se utiliza el método de equiparación lineal cuando las diferencias entre ambas formas no se consideran constantes, sino que estas diferencias pueden variar a lo largo de la escala de puntuación. En otras palabras, este método permite que la forma X sea más difícil que la forma Y para los examinados de bajo rendimiento, pero más fácil para los de mayor rendimiento (Kolen y Brennan, 2014). En este procedimiento se establece la equiparación en puntuaciones estandarizadas, teniendo en cuenta que pueden ser diferentes tanto en la media como en la desviación estándar (Martínez, Lloreda y Lloreda, 2007).

Específicamente para el diseño de grupos no equivalentes o diseño con ítems de anclaje, el método lineal puede basarse en los datos o en el grupo sintético. En el caso de los métodos basados en los datos, son procedimientos aplicados a un conjunto de datos que fueron obtenidos de manera empírica. Y los métodos basados en el grupo sintético, son

estimaciones de los datos, considerando que existen subpoblaciones con medias, varianzas y distribuciones iguales (Lancheros, 2013). Estos métodos incluyen el método de Tucker, y el método de Levine (Kolen y Brennan, 2014).

*Método de Tucker***.** Este método hace dos tipos de suposiciones para estimar los parámetros que no pueden estimarse directamente. La primera suposición hace referencia a las regresiones del puntaje total del ítem común. Y la segunda, se refiere a las variaciones del total de las puntuaciones dados los puntajes de los ítems comunes (Kolen y Brennan, 2014). De manera general se basa en una ecuación lineal para expresar los puntajes de la versión X en función de la versión Y y se calcula el error estándar, que es desviación estándar de los puntajes igualados con resultados hipotéticos del procedimiento de equiparación (Lancheros, 2013).

Método de Levine. El método de puntuación observada de Levine se caracteriza porque no considera las puntuaciones verdaderas, solo las observadas (Kolen y Brennan, 2014). Este método asume que la correlación entre las puntuaciones verdaderas de las versiones X y Y es igual en los grupos en conjunto con el test de anclaje; que los parámetros de la ecuación de regresión tanto de X como de Y sobre el test de anclaje son iguales en los dos grupos; y que la varianza del error de medida de los test X, Y y los anclas es la misma para los dos grupos (Lancheros, 2013).

*Método equipercentil***.** En este procedimiento se usa una curva para representar las diferencias de dificultad entre las formas (Kolen y Brennan, 2014). En otras palabras, se equiparan los puntajes de las pruebas en términos del rango percentil. Se utiliza una versión de referencia (X) y se ajustan los puntajes de otra versión (Y), transformando cada puntaje de Y al puntaje de X (Livingston, 2014). Con este procedimiento se hacen equiparables las

puntuaciones de ambas versiones en las que los percentiles son similares, de manera que los puntajes ajustados tengan una distribución similar (Lancheros, 2013). El procedimiento para llevar a cabo la equiparación equipercentil también dependerá del diseño utilizado.

La equiparación equipercentil puede realizarse mediante el empleo de procedimientos analíticos o gráficos. En el primer caso se trabaja con fórmulas que proporcionen definiciones más formales de los rangos percentiles, permitiendo equiparar en tiempo real un gran número de formas. Por su parte, los procedimientos gráficos trazan el percentil de cada forma en el mismo gráfico a fin de identificar si el rango percentil de la forma X es similar al de la forma Y, permitiendo generar un marco conceptual posterior a los métodos analíticos (Kolen y Brennan, 2014).

En conclusión, en el caso específico de los diseños de ítems comunes basados en TCT, la selección del procedimiento de equiparación dependerá de la manera en que se utiliza la información de anclaje, lo que exigirá la aplicación de al menos un método para equiparación lineal y otro para equiparación equipercentil. El procedimiento de equiparación resulta más complejo para los diseños de anclaje, pues no es posible suponer que los grupos que responden las dos formas son iguales en cuanto a nivel de conocimiento o habilidad, y la información de anclaje se usa para ajustar estas diferencias (Livingston, 2014).

Procedimientos basados en el Modelo de Teoría de Respuesta al Ítem. Los procedimientos basados en TRI se utilizan principalmente cuando se asume que los grupos que responden versiones alternativas tienen diferente nivel de conocimiento y habilidad y por lo tanto es necesario equiparar las puntuaciones para asegurar que se mide el mismo nivel de atributo (Rodríguez, 2007). La equiparación de puntuaciones bajo este modelo se lleva a cabo después de estimar los parámetros a una escala mediante su transformación lineal y al

escalar los puntajes de la versión Y a los de la versión X (la de referencia) (Kolen y Brennan, 2014).

Al aplicar procedimientos de equiparación con grupos no equivalentes, los parámetros de todas las versiones por analizar deben estar en una métrica común, ya que las responden grupos de sujetos diferentes, de manera que es necesario convertir las estimaciones de cada versión a la misma escala por medio de transformaciones lineales, por lo que los parámetros de los ítems ya transformados se utilizan para establecer la equivalencia de puntuaciones (Martínez, Lloreda y Lloreda, 2007). En este marco, los procedimientos de equiparación basados en TRI se clasifican en: (a) Métodos basados en los momentos; (b) métodos basados en la curva característica (CC), y (c) otros métodos, como el método de la b's fijas (parámetros fijos) y el método de calibración concurrente (Lancheros, 2013).

Métodos basados en los momentos. Se deriva el valor de las constantes de la transformación lineal, relacionando las estimaciones para los parámetros de los sujetos y de los ítems en dos momentos distintos (Lancheros, 2013; Rodríguez, 2016). Los métodos basados en los momentos incluyen los siguientes: (a) el método media/sigma, que estima el valor del parámetro de dificultad de los ítems para generar la ecuación de equiparación. Esta ecuación se deriva del valor de las constantes A y B y se obtiene la estimación del parámetro reescalado (Kolen y Brennan, 2014; Lancheros, 2013; Rodríguez, 2016); (b) método de media y desviación típica robusta; y (c) método iterativo de la media y la desviación típica robusta y ponderada. Estos dos últimos métodos, son estimadores que permiten emplear la totalidad de los valores, ponderando los valores extremos (Atkinson, Ariza y García-Balboa, 2007).

Métodos basados en la curva característica (CC). El principal objetivo de este método es la estimación simultánea de todos los parámetros de los ítems. Para lograr dicha estimación existen métodos como el de Haebara (1980) y el de Stocking y Lord (1983) que se ubican dentro de los métodos de transformación basados en la curva características (CC) (Kolen y Brennan, 2014).

El método de Stocking y Lord se basa en las diferencias al cuadrado entre la CC de las dos versiones en diferente nivel de aptitud, tomando en cuenta la suma de cada nivel de aptitud antes de elevar al cuadrado. En este método se expresa la diferencia al cuadrado entre las CC de la prueba para un ítem determinado y después se suma al puntaje de los examinados (Kolen y Brennan, 2014). Los resultados se basan en una aproximación a las puntuaciones verdaderas, pero al contar con las puntuaciones observadas se sugiere el uso de un programa iterativo para establecer la equivalencia final entre las puntuaciones. Kolen y Brennan (2014) proponen el programa PIE de Hanson y Zeng (2004) o el software POLYST (2014) que implementa esquemas para la suma de los examinados.

En el caso del método de Haebera, en este se usa la suma de las diferencias cuadradas entre las CC de cada ítem para los examinados en un dominio específico (Kim, 2007). Estima los parámetros de las calibraciones como un conjunto separado de habilidades, de manera que sea posible minimizar la función multivariante (Raju y Areson, 2002).

Otros métodos. En la clasificación general de otros métodos de equiparación basados en TRI se encuentran, principalmente, el método de calibración de parámetros fijos y el método de calibración concurrente. La calibración de parámetros fijos propone ajustar los parámetros de los ítems a los ítems comunes, estimando los puntajes de la forma X al calibrar los ítems de la forma Y. Se sugiere su aplicación cuando existen diferencias importantes en

la habilidad de los sujetos de cada grupo, de manera que se estimen los parámetros de los ítems sesgados (Kolen y Brennan, 2014).

En la *calibración concurrente*, los parámetros de todos los ítems se estiman en una misma escala, "0 y 1". En este procedimiento se consideran todos los ítems que no son respondidos por los participantes, generando una cadena de respuesta para cada ítem y se codifican como "no respondido" aquellos ítems de la versión que no ha tomado ese grupo de la muestra (Kim, 2007). Para realizar este procedimiento de calibración concurrente es deseable ejecutar el programa BILOG-MG, para llevar a cabo este procedimiento, se indican los ítems que son comunes en ambas pruebas y se señala el grupo que ha respondido cada versión.

*Procedimientos de equiparación para diseños de grupos no equivalentes con ítems de anclaje***.** Estos diseños se caracterizan por presentar dos componentes importantes, el diseño para recopilar los datos y un modelo estadístico que permita equiparar los puntajes de interés. Este diseño se utiliza comúnmente para evaluar las diferencias en los niveles de habilidad entre los grupos y para estimar distribuciones de puntajes en las dos versiones que se equiparan (Dorans, Pommerich y Holland, 2007).

En los diseños con ítems comunes, es importante mencionar que, las diferentes versiones deben cumplir con ciertas condiciones, entre las que destacan la similitud en el tamaño de la muestra, la similitud de las formas y la relación entre puntajes en cada una de las formas que se igualarán y en los ítems comunes. En otras palabras, se debe asegurar que los ítems son equivalentes en cuanto a su contenido, dificultad y nivel de confiabilidad. Asimismo, es necesario determinar en qué medida los ítems comunes reflejan propiedades de la prueba en su totalidad (Dorans, Pommerich y Holland, 2007).

Los procedimientos de equiparación basados en el diseño de grupos no equivalentes con ítems comunes aplican dos métodos de igualación en el caso de la TCT: la estimación de frecuencias o postestratificación y el método de equivalencia equipercentil encadenado. El primero se basa en la suposición de que la distribución de las puntuaciones de los ítems comunes es invariable a través de las poblaciones. En el caso de la equivalencia equipercentil encadenado primero se iguala X a V y después V a Y (Von Davier, 2011).

Por otro lado, se encuentran los métodos de puntuación verdadera tales como el puntaje real de Levine y los métodos de calibración basados en TRI. Estos últimos se caracterizan por usar la prueba de anclaje para estimar los parámetros de los ítems en las dos formas antes de igualar el puntaje (Dorans, Pommerich y Holland, 2007).

Según Kolen y Brennan (2014), en este tipo de diseños solo se deben estimar los parámetros de la versión X cuando ésta se equipara con la versión Y, puesto que al no ser los mismos sujetos quienes respondieron ambas pruebas, no se consideran equivalentes las muestras y las estimaciones de los parámetros no están en la misma escala, pero hay un conjunto de elementos que es común en ambas formas (ítems de anclaje). Así, las estimaciones de dichos parámetros pueden utilizarse para estimar la transformación de la escala. Como se mencionó anteriormente, la calibración concurrente es la mejor alternativa para estimar los parámetros de la versión X y Y. Kolen y Brennan (2014), sugieren que la calibración concurrente es la mejor alternativa para estimar los parámetros de las diferentes versiones.

Respecto a la importancia de aplicar procedimientos de equiparación a las pruebas, se ha identificado que, puesto que los exámenes de selección para el ingreso a la universidad generalmente se aplican en un periodo de varios días o se registran varias aplicaciones por

día, puede presentar ciertas ventajas a algunos aspirantes por encima de otros. Por tal motivo, el contenido del examen y sus ítems se encuentran expuestos a que el examinado informe a otros sobre el contenido de estos, o bien que el sustentante obtenga ventaja sobre otros al aplicar el examen en varias ocasiones, cuestiones que atentan contra los mecanismos de seguridad requeridos en este tipo de pruebas (Kolen y Brennan, 2014).

Por lo anterior, el diseño de los exámenes considera el desarrollo de distintas versiones orientadas a medir el mismo atributo. Si bien existen procedimientos asociados al diseño de un test que permiten generar versiones paralelas es inevitable observar diferencias considerables en sus propiedades psicométricas, particularmente en lo relativo a sus índices de dificultad. Lo anterior pudiera derivar, al momento de interpretar las puntuaciones de los grupos expuestos a diferentes versiones de una misma prueba, que una versión resulte más fácil o difícil que la otra, lo que sin duda pondría en ventaja a un grupo sobre otro, lo cual resultaría injusto (González y Wiberg, 2017). Por lo tanto, es indispensable realizar procedimientos de equiparación de las formas administradas de un examen, de manera que las puntuaciones de cualquier examinado puedan considerarse de manera indistinta como si procedieran de una misma prueba (González y Wiberg, 2017).

El Examen de Ingreso a la Educación Superior (ExIES)

El ExIES es un examen de admisión diseñado, desarrollado y aplicado por la Universidad Autónoma de Baja California (UABC), utilizado para la selección de aspirantes a sus programas de licenciatura desde el 2017. Tiene como objetivo medir las habilidades y conocimientos de los aspirantes a ingresar a la Educación Superior en los siguientes tres componentes: lectura, lengua escrita y matemáticas (Caso, Díaz, Castro y Martínez, 2017), a continuación, se describe cada uno de los componentes.

Prueba de lectura. Evalúa la capacidad para leer y comprender textos literarios e informativos. Esta prueba está compuesta por 36 preguntas de opción múltiple y están planteadas para evaluar la capacidad de un estudiante para crear conexiones entre los textos y comprender cada texto individualmente.

Prueba de Lengua Escrita. Está compuesta también por 36 preguntas de opción múltiple, que evalúan la capacidad para revisar y editar una gran variedad de textos de naturaleza académica. Estos textos están diseñados de manera que exigen al examinado tomar decisiones de edición y revisión. Las preguntas que conforman la prueba de lengua escrita se enfocan en el reconocimiento, identificación y corrección de errores.

Prueba de Matemáticas. La prueba de matemáticas está compuesta por 50 preguntas de opción múltiple y tiene como objetivo medir la capacidad del aspirante para la aplicación, manejo y comprensión de conceptos matemáticos y la habilidad para resolver problemas, interpretar datos, tablas, cuadros y gráficas.

El ExIES se compone de 122 preguntas y el examinado tiene en total 3 horas para responderlo. Respecto a la puntuación, esta se presenta en una escala de 700 a 1300 puntos,

considerando que todas las preguntas tienen el mismo valor y se califican con un punto por respuesta correcta y cero para pregunta incorrecta o sin responder (Caso et. al. 2017).

Desarrollo del ExIES. De acuerdo con Caso et al. (2017), en el desarrollo del ExIES se consideró la definición de un marco de referencia, la elaboración de especificaciones del contenido, la estipulación de niveles de demanda cognitiva para cada contenido de la prueba y la declaración de especificaciones de orden psicométrico. En lo particular, las especificaciones de contenido se desarrollaron con base en el análisis de las competencias comprometidas en la Educación Media Superior y en las competencias propuestas por la OCDE en el proyecto *Definition and Selection of Competencies* (DeSeCo; OCDE, 2005).

Las especificaciones de contenido permitieron establecer los contenidos y subcontenidos específicos de cada prueba del examen, así como la definición operacional de cada elemento a evaluar:

Prueba de Lectura: Se conforma por las dimensiones, (a) Información e ideas. Evalúa el contenido informativo del texto); (b) Formas discursivas. Analiza la estructura del discurso); y (c) Intertextualidad. Evalúa la síntesis de varias fuentes de información.

Prueba de Lengua Escrita: Se conforma por las dimensiones: (a) Expresión de ideas. Revisión del desarrollo del tema, precisión, lógica, cohesión y uso del lenguaje en un texto; y (b) Cumplimiento de reglas del español escrito. Edición de un texto para asegurar reglas gramaticales, estructura de oraciones, uso y puntuación).

Prueba de Matemáticas: Se conforma de las dimensiones (a) Herramientas algebraicas. Mide la resolución de problemas algebraicos; (b) Problemas, probabilidad y análisis de datos. Mide la creación y análisis de relaciones, representación y análisis de datos

cuantitativos y aplicación de probabilidades; (c) Matemáticas avanzadas. Mide la creación de expresiones algebraicas y uso de gráficas para funciones no lineales o cuadráticas; y (d) Temas adiciones en matemáticas. Mide la solución de problemas sobre área y volumen, aplicación de definiciones y teoremas sobre líneas, ángulos, triángulos y círculos.

Respecto al nivel de demanda cognitiva de sus ítems, se consideraron los niveles de comprensión, aplicación, análisis, síntesis y evaluación declarados en la Taxonomía de Bloom. Así, la configuración de los ítems del examen quedó distribuidos de la siguiente proporción: 8% a nivel de comprensión, 46% a nivel de aplicación, 5% a nivel de análisis, 11% a nivel de síntesis y 30% en el nivel de evaluación.

Una vez elaborados los ítems por un grupo de especialistas en los dominios curriculares que conforman el examen, éstos fueron revisados por expertos que valoraron los siguientes elementos: (a) la correspondencia entre el contenido y el ítem, (b) la correspondencia entre el nivel de demanda cognitiva y el ítem, (c) la redacción de la base del ítem, (d) la redacción de las opciones de respuesta y (e) la equidad en el contenido de los ítems.

En cuanto a las especificaciones de dificultad, discriminación y confiabilidad, el ExIES se diseñó considerando los siguientes criterios; (a) una dificultad media de .60; (b) un rango de dificultad de sus ítems de .20 a .90; (c) un rango de dificultad de cada ítem entre -1.5 a 1.5 lógitos de acuerdo con lo establecido por el modelo logístico de un parámetro de la Teoría de Respuesta al Ítem (TRI); (d) un índice de discriminación o correlación biserial de sus ítems igual o mayor a .20; y (e) índices de consistencia interna mayor a >.70 (Caso et al. 2017).

Tanto la aplicación piloto del periodo 2016-2 (n= 2,402) como la aplicación a gran escala del examen en el periodo 2017-1 (n= 32,388) permitieron documentar las siguientes evidencias de validez y confiabilidad que se observan en la tabla 1.

Tabla 1
Promedio de los índices de validez y confiabilidad del ExIES

	Prueba de lectura	Prueba de Lengua Escrita	Prueba de Matemáticas
No. de ítems	72	72	100
Dificultad promedio	.42	.44	.29
Discriminación promedio	.26	.23	.25
Confiabilidad promedio	.67	.50	.70

Fuente: Caso et al. (2017).

De acuerdo con los criterios de ajuste estadístico, el examen presenta evidencia de unidimensionalidad (INFIT= 1.00, OUTFIT entre 0.99 y 1.02), así como ausencia de sesgo en términos del análisis del funcionamiento diferencial de los ítems, con magnitudes inferiores a .13 que sugieren efectos insignificantes. Respecto al índice de consistencia interna, este fue de .92 para todo el examen.

Versiones alternativas del ExIES. En atención a los estándares de calidad para pruebas psicológicas y educativas (AERA, APA, NCME, 2014, ITC, 2013), el desarrollo del ExIES implicó el diseño y aplicación de cuatro versiones diferentes, que para fines de los análisis psicométricos se han nombrado, versión A, B, C y D, respectivamente. Estas versiones fueron construidas con base en la tabla de especificaciones, cuidando que la proporción de ítems fuera homogénea en relación con el contenido y subcontenido en cada

prueba. Asimismo, se consideraron los resultados de los análisis de calibración, específicamente los índices de dificultad y discriminación.

Para asegurar la equivalencia entre versiones, el equipo técnico encargado del desarrollo del ExIES, diseñó las cuatro versiones con ítems de anclaje, de manera que todas las versiones tienen 10 ítems en común. Los ítems de anclaje representan el 20% del total de ítems en el examen (ver tabla 2), y fueron seleccionados con base en la tabla de especificaciones del examen, son similares en formato, rango de dificultad y nivel discriminación, además no mostraron efecto de funcionamiento diferencial.

Tabla 2
Distribución de ítems en el ExIES

Distribución de ítems			
	Ítems Único	Ítems de anclaje	Total
Lectura	36	10	46
Lengua escrita	36		46
Matemáticas	50		60
Total	122	10	152

Fuente: Elaboración propia

Método

Con base en los criterios para realizar procedimientos de equiparación de puntuaciones, se realizó el diseño y aplicación del Examen de Selección a la Educación Superior (ExIES), para la equiparación de sus puntuaciones se propuso una metodología en dos fases. La primera fase hizo referencia a los análisis psicométricos previos a la equiparación y la segunda fase al procedimiento de equiparación apropiado al diseño de selección de la muestra, en este caso, al diseño de grupos no equivalentes con ítems de anclaje.

Participantes

Participaron 2,898 aspirantes a ingresar a los programas de licenciatura de la UABC, en la convocatoria correspondiente al mes de noviembre de 2017. El sexo de los participantes fue de 49% mujeres y 51% hombres, mismos que se distribuyeron según la versión del examen, de acuerdo como se presenta en la tabla 3.

Tabla 3
Distribución de frecuencias por sexo

		Versión				Total
		A	B	C	D	
Sexo	Mujer	397	351	356	307	1,411
	Hombre	360	400	377	347	1,484
Perdidos						3
Total		757	751	733	654	2,898

Fuente: Elaboración propia

Instrumento

Se utilizaron los datos de la aplicación de las cuatro versiones (A, B, C y D) de la Prueba de Matemáticas del ExIES. La prueba de matemáticas que tiene como objetivo medir la capacidad de los aspirantes para la aplicación, manejo y comprensión de conceptos matemáticos, así como la habilidad para la resolución de problemas y para la interpretación de datos, tablas, cuadros y gráficos (Caso et al. 2017).

Respecto al nivel de demanda cognitiva, el 8% de los ítems corresponde al nivel de comprensión, 22% aplicación, 5% análisis y 6% al nivel de síntesis. En la tabla 4 se presenta la proporción de ítems de acuerdo a la tabla de especificaciones de la prueba que considera los siguientes contenidos: herramientas algebraicas; problemas, probabilidad y análisis de datos; matemáticas avanzadas y temas adicionales en matemáticas.

Tabla 4
Proporción de ítems de la prueba de Matemáticas del EXIES

Contenido	Proporción
Herramientas algebraicas	20%
Problemas, probabilidad y análisis de datos	30%
Matemáticas avanzadas	30%
Temas adicionales en matemáticas.	20%

Fuente: Caso et al. (2017).

Cada versión está compuesta de 60 ítems, de los cuales 50 son ítems únicos y 10 son ítems de anclaje. Por lo tanto, se analizaron un total de 210 ítems (50 ítems por versión, más 10 ítems comunes para las cuatro versiones), distribuidos homogéneamente en cada versión de examen. Los ítems de anclaje fueron seleccionados con base en el resultado de los análisis psicométricos de la aplicación del examen de ingreso en la convocatoria 2017-2 de la UABC. Su dificultad y discriminación media fue de .40 y .39.

Todos los ítems se calificaron con un punto para la respuesta correcta y cero puntos para la respuesta incorrecta o en blanco, sin embargo, el diseño de ítems comunes se realizó de manera externa, es decir, la puntuación de estos no contribuyó a la puntuación total, por lo que no generó un impacto en la decisión de admisión de los aspirantes.

Análisis de datos

El análisis de datos se realizó en dos fases: (1) análisis de la calidad métrica de los ítems para cada versión de la prueba; (2) Aplicación de la técnica estadística para equiparar los resultados de las cuatro versiones de la prueba.

Fase 1. Análisis de la calidad métrica de los ítems. El objetivo de esta fase fue analizar la calidad métrica de los ítems en cada versión de la prueba. Los análisis realizados en esta fase son parte de los requisitos previos para aplicar los procedimientos de equiparación propuestos por Kolen y Brennan (2014), mismos que de cumplirse permitirán vincular los puntajes de las cuatro versiones.

Se realizaron los siguientes análisis:

Análisis de confiabilidad: Para calcular el índice de consistencia interna de las puntuaciones se estimó el índice alfa ordinal, con base en una matriz de correlaciones tetracóricas. El alfa ordinal se aplica en el caso específico de variables dicotómicas (Freiberg, Stover, De la Iglesia y Fernández, 2013). Para la interpretación del alfa ordinal, se considera que valores entre .70 y .90 representan una buena consistencia interna (González y Pazmiño, 2015). Este análisis se realizó dentro del paquete *psych* en el programa estadístico R studio (Revelle, 2018).

Análisis de discriminación: A partir de la matriz de correlaciones tetracóricas, se calculó el coeficiente de correlación biserial para cada ítem. Este coeficiente estima el grado en que el constructo que mide la prueba también lo mide el ítem (Backhoff, Larrazolo y Rosas, 2000). Como criterio de interpretación se tomó como referencia los valores propuestos por Ebel y Frisbie (1986, en Backhoff, Larrazolo y Rosas, 2000), en los cuales, a partir del índice estimado, se señala la calidad del ítem y las recomendaciones asociadas a dicho valor (ver tabla 5). Se estimó dentro del paquete *psych* en el programa estadístico R studio (Revelle, 2018).

Tabla 5
Criterios de interpretación para el coeficiente de discriminación

Discriminación	Calidad	Recomendación
>. 39	Excelente	Conservar
.30 - .39	Buena	Posibilidades de mejorar
.20 - .29	Regular	Necesidad de revisar
0 - .20	Pobre	Descartar o revisar a profundidad
<.01	Pésima	Descartar Definitivamente

Fuente: Elaboración propia con base en Ebel y Frisbie (1986).

Análisis de dificultad (Dj). La dificultad de cada ítem fue calculada de acuerdo con la proporción de estudiantes que respondió correctamente el ítem, entendiendo que entre mayor es la proporción menor será la dificultad del ítem. La estimación del índice de dificultad se realizó con el programa estadístico para las ciencias sociales (SPSS), V. 21. y se interpretó con base en los criterios propuestos por Abad et al. (2007), quienes señalan tres niveles de dificultad (alto, medio y bajo), considerando que, cuando el Dj se acerca a 0 indica que el ítem ha resultado muy difícil; si se acerca a 1, que ha resultado muy fácil; y si se acerca a .5, presenta dificultad media.

Análisis de unidimensionalidad: Con el fin de examinar la dimensionalidad latente de los ítems, se utilizó la técnica de análisis paralelo modificado (MPA; Drasgow y Lissak 1983) para cada versión de la prueba. El MPA utiliza la matriz de puntuaciones de los sujetos en los ítems para estimar los parámetros iniciales de los ítems. Con base en los parámetros estimados se simulan las respuestas de los mismos sujetos y se estiman nuevos parámetros en un modelo factorial. Finalmente se comparan los autovalores de los datos reales y los datos simulados. La obtención de evidencia de unidimensionalidad se da cuando el conjunto de datos se concentra, principalmente, en el primer factor. El MPA se realizó con el paquete "ltm" del programa R (Rizopoulos, 2006).

Análisis del funcionamiento diferencial del ítem (DIF): Como técnica para analizar la propiedad de invarianza poblacional, se examinó el DIF por sexo para cada versión de la prueba. Se utilizó el método Mantel y Haenszel (1959) mismo que se aplica para comparar la ejecución de un ítem en dos grupos, uno focal y uno de referencia (Guilera, Gómez-Benito, Hidalgo y Sánchez-Meca, 2007). Para la interpretación del efecto del DIF en los ítems se tomó la clasificación de Dorans y Holland, (1993 en Mohahan, McHorney, Stump y Perkins,

2007), la cual establece que los efectos del DIF pueden resultar no significativos (< 1); de magnitud leve o moderada (>1, <1.5) y severos (> 1.5). Este análisis fue realizado en la plataforma para análisis estadísticos R-Studio versión 5.1, con el paquete "difR" (Magis, Beland, Tuerlinckx y De Boeck, 2010).

Fase 2. Equiparación. Se aplicó el procedimiento de calibración concurrente como método de equiparación de puntuaciones. El objetivo de este método fue calibrar en una misma métrica los ítems de todas las versiones (Olea, Abad, Ponsoda y Ximénez, 2004). Los pasos para el análisis de calibración concurrente en el software Winsteps V. 3.81.0 (Linacre, 2013) fueron los siguientes:

1. Se unieron las respuestas de las cuatro versiones de la prueba para todos los examinados, de manera que se consideraron los 210 ítems en una misma cadena.
2. Se estimó el parámetro de todos los ítems en la escala de 1 para la respuesta correcta y 0 para la errónea o no respondida, otorgándose como valor perdido la respuesta a los ítems de las versiones que no fueron respondidas por los demás grupos.
3. Se generó la matriz del conjunto de ítems únicos y comunes para identificar la correspondencia de los ítems compartidos.
4. Se calculó el parámetro de dificultad calibrado desde el modelo de Rasch, por ítem.
5. Se interpretaron los índices de bondad de ajuste interno, INFIT y OUTFIT. Según, Badanes (2009), el criterio de ajuste de estos indicadores se encuentra en -2 a + 2, en los cuales un valor >2 sugiere que el modelo no ajusta.

Para presentar evidencia de la equiparación de puntuaciones se realizaron los siguientes análisis:

- Proporción de ítems por nivel de dificultad. Una vez calibrados los ítems, se comparó el parámetro de dificultad estimado con TCT y con TRI y se presentó la frecuencia de ítems en cada nivel de dificultad, según los criterios señalados para cada modelo. En el caso de la TCT, se tomó como criterio de interpretación los propuestos por Abad et.al (2007): Fácil (0- .30), media (.30 -.70), difícil (.70-1). Como criterio para la estimación con TRI , se considera, según Prieto y Delgado (2003): difícil (-3 a – 1.8), dificultad media (-1.8 a +1.8) y fácil (+1.8 a +3).
- Distribución de las puntuaciones por sujeto**.** Se calculó la media y la desviación estándar de la puntuación de los sujetos para cada modelo de estimación y se comparó la distribución de las puntuaciones en función de la asimetría y la curtosis de los datos.
- Correlación de la habilidad de los sujetos**.** Se obtuvo la correlación de Pearson (r) entre la habilidad estimada en TCT y en Rasch, con el objetivo de observar si los puntajes calibrados tienen relación con la puntuación natural de los sujetos.
- Prueba de diferencia de medias. Se aplicó la prueba *t* de Student para muestras relacionadas, a fin de identificar si existe diferencia en la media de la puntuación de los sujetos estimada con ambos modelos. Como criterio de interpretación se tomaron niveles de significancia de $p < .05$, lo que indica que existe diferencia estadística entre las dos muestras (Pagano, 2008).
- Medidas de posición**.** Finalmente, se realizó un análisis descriptivo de las posiciones que ocupan los aspirantes en función de modelo con que se estima su

puntuación. Se dividieron los puntajes en cuartiles y se comparó la cantidad de sujetos dentro de cada cuartil para cada modelo de estimación, con la finalidad de observar si se penaliza o beneficia a algunos aspirantes al calibrar los ítems.

Resultados

En este apartado se presentan los resultados de las dos fases propuestas para el análisis de datos. En un inicio se describen los resultados del análisis de la calidad métrica de los ítems para cada versión de la prueba. Posteriormente se describen los resultados de la calibración del parámetro de dificultad de las cuatro versiones de la prueba y, finalmente, la correlación del parámetro de dificultad estimado con TCT y con el modelo de Rasch como evidencia de la equiparación de puntuaciones por sujeto.

Análisis de la calidad métrica de los ítems

En esta fase se describen los resultados de las siguientes propiedades psicométricas de los ítems: consistencia interna, coeficiente de discriminación, índice de dificultad, resultados del análisis de unidimensionalidad y resultados del análisis del funcionamiento diferencial de los ítems (DIF). Dichos análisis fueron aplicados a cada versión la prueba.

Las propiedades psicométricas analizadas presentan valores aceptables de acuerdo con los criterios de interpretación sugeridos por la literatura especializada, se registran valores similares en las cuatro versiones. En la tabla 6 se presenta un concentrado con el cálculo de la confiabilidad, dificultad y discriminación para cada una de las versiones de la prueba, considerando el total de ítems, después se describe cada una de las propiedades analizadas.

Tabla 6

Propiedades psicométricas de la prueba de Matemáticas

	Versión			
	A	B	C	D
N	759	751	733	655
No ítems	50	50	50	50
Confiabilidad	.88	.78	.85	.87
Dificultad promedio	.38	.36	.39	.37
Discriminación promedio	.36	.27	.32	.34

Fuente: Elaboración propia. Nota: *N* = Total de participantes por versión

Consistencia interna de los ítems. Se calculó el coeficiente alfa ordinal como estimación de la consistencia interna de los ítems. Las cuatro versiones presentaron valores similares, observándose un valor alfa ordinal >.85 en las versiones A, C y D y un valor de .78 en la versión D. Para la interpretación de este índice, González y Pasmino (2015) consideran que valores entre .70 y .90 indican un alto nivel de consistencia interna.

Discriminación. El coeficiente de discriminación fue calculado mediante la correlación punto biserial (rpbis). La discriminación promedio del conjunto de ítems que conforman las versiones A, C y D fue >30. Por su parte, los ítems de la a versión B presentaron un índice de discriminación promedio de .27, lo que se considera un poder de discriminación regular, sugiriendo la revisión de los ítems para mejorar su comportamiento.

Con relación al total de ítems, cada versión presenta un porcentaje importante de ítems que, según el criterio mencionado, se recomienda revisar a profundidad por obtener niveles de correlación >.20. La versión A presenta 18% de los ítems con niveles de discriminación inferiores a .20, la versión B el 25%, la versión C 28% y la versión D 16% (Apéndice A).

En el caso de los 10 ítems de anclaje (identificados con la letra "X") y con base en la mediana de la puntuación calculada, se observó que el ítem 8C presentó un nivel de discriminación regular (*rbis* = .29) y el ítem 10 un buen poder discriminativo (*rbis* = .39), valores que sugieren la posibilidad de mejorar. El resto de los ítems registraron alto poder discriminativo (*rbis* > .39). Al comparar el comportamiento de las cuatro versiones se observa, de manera general, que el coeficiente de discriminación de los ítems es similar (ver tabla 7).

Tabla 7
Coeficiente de discriminación de los ítems de anclaje

	Versión				
Ítem	A	B	C	D	*Mdn*
1X	.52	.54	.55	.56	.54
2X	.53	.56	.51	.48	.52
3X	.49	.43	.40	.38	.41
4X	.39	.35	.44	.38	.38
5X	.41	.40	.51	.38	.40
6X	.49	.46	.55	.55	.52
7X	.49	.45	.52	.60	.50
8X	.31	.27	.25	.34	.29
9X	.49	.49	.36	.32	.42
10X	.37	.42	.43	.31	.39

Fuente: Elaboración propia. *Nota:* Mdn= Mediana del coeficiente de discriminación por ítem

Índice de dificultad (Dj). La media del índice de dificultad en las cuatro versiones fue .40, lo que se interpreta como un nivel medio de dificultad. El criterio de interpretación para este índice sugiere que entre más cercano a 0 el ítem es más difícil, entre más cercano a 1 más fácil y si se acerca a .5 presenta una dificultad media (Abad, Olea y Ponsoda, 2007). Por otro lado, se observó una distribución adecuada de los ítems por nivel de dificultad en cada versión al presentar ítems en los tres niveles de dificultad (Apéndice B).

En el análisis de los ítems de anclaje, los ítems 3 y 4 presentaron los índices de dificultad más altos observando una dificultad media, de .21 y .22, en las 4 versiones; mientras que el ítem que presentó menor dificultad promedio fue el ítem 6 (*Dj*= .50) con valores similares para las cuatro versiones (Ver tabla 8).

Tabla 8
Índice de dificultad de los ítems de anclaje

	Versión				
Ítem	A	B	C	D	*Mdn*
1X	.40	.38	.46	.46	0.43
2X	.46	.48	.50	.49	0.48
3X	.21	.21	.24	.20	0.21
4X	.20	.21	.24	.22	0.21
5X	.33	.33	.37	.36	0.34
6X	.50	.54	.50	.50	0.50
7X	.34	.43	.41	.38	0.39
8X	.34	.36	.30	.29	0.32
9X	.32	.28	.27	.27	0.27
10X	.42	.43	.41	.45	0.42

Fuente: Elaboración propia. *Nota: Mdn*= Mediana del índice de dificultad por ítem

Análisis de unidimensionalidad: Los resultados del análisis paralelo indican que las cuatro versiones presentan la propiedad de unidimensionalidad ($p < .00$), al presentar el segundo autovalor de los datos observados sustancialmente mayor que el segundo autovalor de los datos bajo el supuesto del modelo de TRI (Ver figura 2).

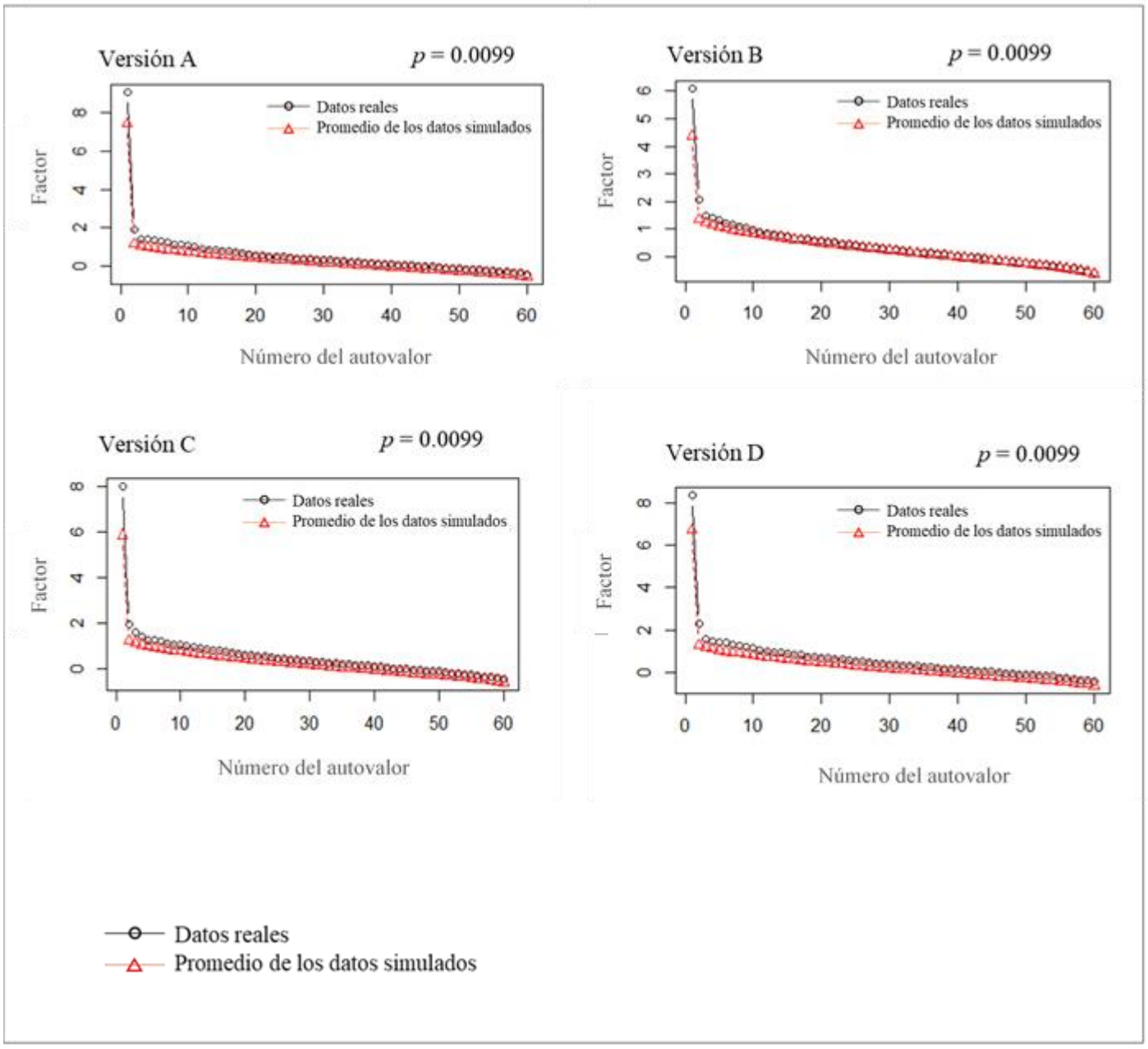

Figura 2. Evidencia de unidimensionalidad por versión, mediante el análisis paralelo

Funcionamiento diferencial del ítem (DIF). La propiedad de invarianza poblacional fue analizada mediante el DIF por sexo, bajo el método de Mantel-Haenszel (1959 en Guilera, et al., 2007), se observó que tres de los ítems únicos de la versión A presentan DIF; uno en la versión B; cuatro en la versión C y tres en la versión D. Respecto a los ítems comunes, el ítem 2X presenta DIF en las cuatro versiones, el ítem 3X en la versión B y el ítem 9X en la versión A (Ver tabla 9). El resto de los ítems, tanto únicos como comunes, no registran indicios de invarianza en ninguna de las versiones (Apéndice C).

Tabla 9
Ítems que presentan DIF en función del sexo

Versión											
A			B			C			D		
Ítem	DIF	Efecto	Ítem	DIF	Efecto	Ítem	DIF	Efecto	Ítem	DIF	Efecto
1	-1.22	B	17	-1.76	C	11	-1.01	B	33	1.02	B
19	1.56	C	2X	-2.22	C	16	1.13	B	40	1.10	B
27	1.18	B	3X	1.49	B	21	1.36	B	44	1.30	B
2X	-1.09	B				33	1.00	B	2X	-1.04	B
9X	1.61	C				2X	-1.52	C			

Fuente: Elaboración propia. Nota: Criterio de interpretación del efecto del DIF, propuesto por Dorans y Holland (1993), B = Moderado (>1, <1.5), C = Severo (>1.5).

Análisis de la equiparación de puntuaciones

Una vez analizadas las propiedades psicométricas de cada versión, se integraron los resultados de las cuatro versiones en una sola cadena de datos para aplicar la técnica de calibración concurrente y obtener el parámetro de dificultad calibrado bajo el modelo de Rasch. El método de calibración concurrente consistió en estimar simultáneamente el parámetro de dificultad de los ítems y considerar los índices INFIT y OUTFIT como

indicadores de ajuste interno. Se observó una dificultad promedio de 0.052 *lógitos*, valor que se encuentra dentro de los criterios propuestos en el modelo de Rasch por Prieto y Delgado en 2003 (– ∞ a +∞ *lógitos*), aunque proponen situar la dificultad media en ±.5 *lógitos* con una media de 0.052 *log*, lo que sugiere que la calibración controló el efecto de la dificultad en los ítems (ver Apéndice D). Los índices de ajuste interno del conjunto de ítems ancla se encontraron dentro del margen razonable de -2 a + 2 *log* (Badanes, 2009), con una media de .95 en INFIT y .94 en OUFIT, valores que sugieren un buen ajuste interno (Ver tabla 10).

Tabla 10
Parámetros de dificultad e índices de ajuste de ítems ancla bajo el modelo de Rasch

Ítem	*Dificultad en Rasch*	N	INFIT (-2 a + 2)	OUTFIT (-2 a + 2)
1X	-0.25637	2898	0.90	0.89
2X	-0.51831	2898	0.91	0.90
3X	0.81373	2898	0.97	0.97
4X	0.78533	2898	0.98	0.99
5X	0.09759	2898	0.96	0.96
6X	-0.63214	2898	0.91	0.90
7X	-0.10242	2898	0.92	0.90
8X	0.21195	2898	1.02	1.03
9X	0.40832	2898	0.97	0.95
10X	-0.27937	2898	0.98	0.97

Fuente: Elaboración propia. *Nota:* N= Total de personas que respondieron el ítem.

Respecto a la proporción de ítems por nivel de dificultad y con base en la clasificación de fácil, dificultad media y difícil, se identificó que una vez calibrados se obtiene un mejor ajuste del índice de dificultad (Apéndice E), observando una importante diferencia en la cantidad de ítems dentro de cada nivel. En el caso de los ítems de anclaje, en la estimación del parámetro con TCT, se presentan 3 ítems difíciles y 7 de dificultad media, mismos que al

calibrarse registraron un nivel medio de dificultad. En la tabla 11 se observa la cantidad de ítems en cada nivel, según el modelo de estimación.

Tabla 11
Proporción de ítems ancla e ítems únicos por nivel de dificultad y modelo de estimación

Ítem	Dificultad	TCT	TRI
Ancla	Fácil	0	0
	Media	7	10
	Difícil	3	0
Únicos	Fácil	9	1
	Media	114	192
	Difícil	73	7

Fuente: Elaboración propia. Criterio de interpretación para los niveles en TCT, propuestos por Abad et. al (2007): Fácil (0- .30), media (.30 -.70), difícil (.70-1). Criterio de interpretación para TRI según Prieto y Delgado (2003): difícil (-3 a – 1.8), dificultad media (-1.8 a +1.8) y fácil (+1.8 a +3).

En el caso de los criterios de interpretación al comparar el índice de dificultad de los ítems, es importante mencionar que en el modelo de Rasch los valores de la dificultad se encuentran entre $-\infty$ a $+\infty$ lógitos, considerando que valores más altos de lógitos implican mayor dificultad. Por su parte, en la TCT el criterio del índice de dificultad (Dj) registra valores de 0 a 1, entendiendo que un valor cercano a 1 significa menor dificultad.

Distribución de las puntuaciones por sujeto. La *media* de las puntuaciones de los sujetos calibradas con TRI fue de -.59 con una S= .627. En la figura 3, se muestra la distribución de las puntuaciones calibradas en comparación con los puntajes naturales obtenidos con TCT. En esta es posible observar que la distribución de los puntajes en TRI presenta una gran concentración de los valores en torno a la media, obteniendo una curva leptocúrtica, mientras que en los puntajes sin calibrar la distribución tiende a generar una

curva mayormente mesocúrtica. Si bien se observa una concentración normal de los valores con relación a la media, estos se distribuyen con mayor normalidad en los puntajes obtenidos con TRI.

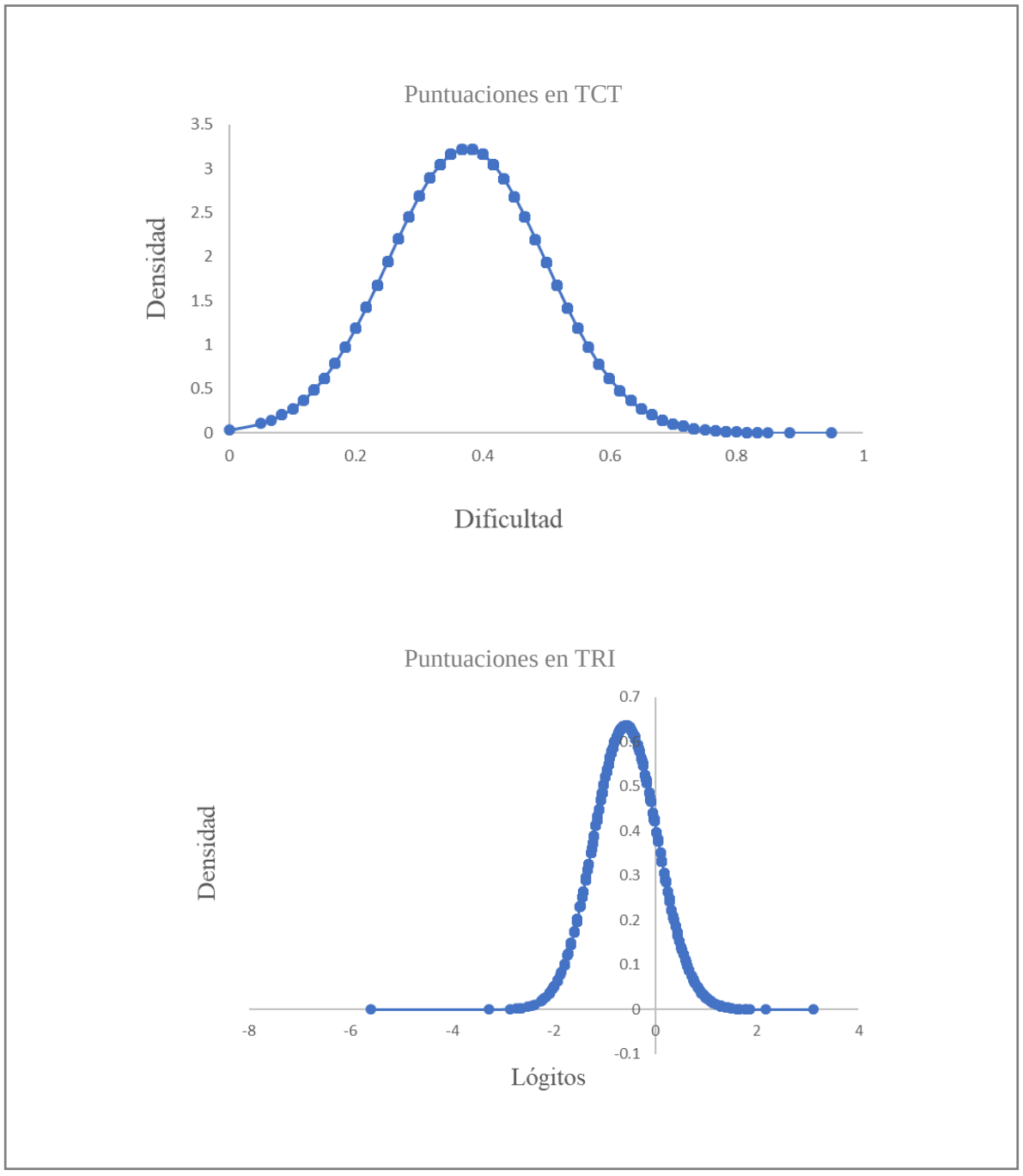

Figura 3. Distribución de las puntuaciones bajo el modelo de la TCT y la TRI por sujeto.

Distribución de los sujetos por modelo de estimación. Se observó una correlación positiva alta entre la puntuación de los sujetos estimadas con TCT y la estimada con el modelo de Rasch ($r > .95$, $p < .05$), lo que sugiere una relación lineal entre ambas estimaciones. Este resultado señala que a medida que aumenta el valor de los lógitos en el modelo de Rasch se acerca el valor a 1 calculado con TCT, es decir conforme aumenta el número de aciertos calculados con TCT, aumenta la habilidad de los sujetos observada con TRI (Ver figura 4).

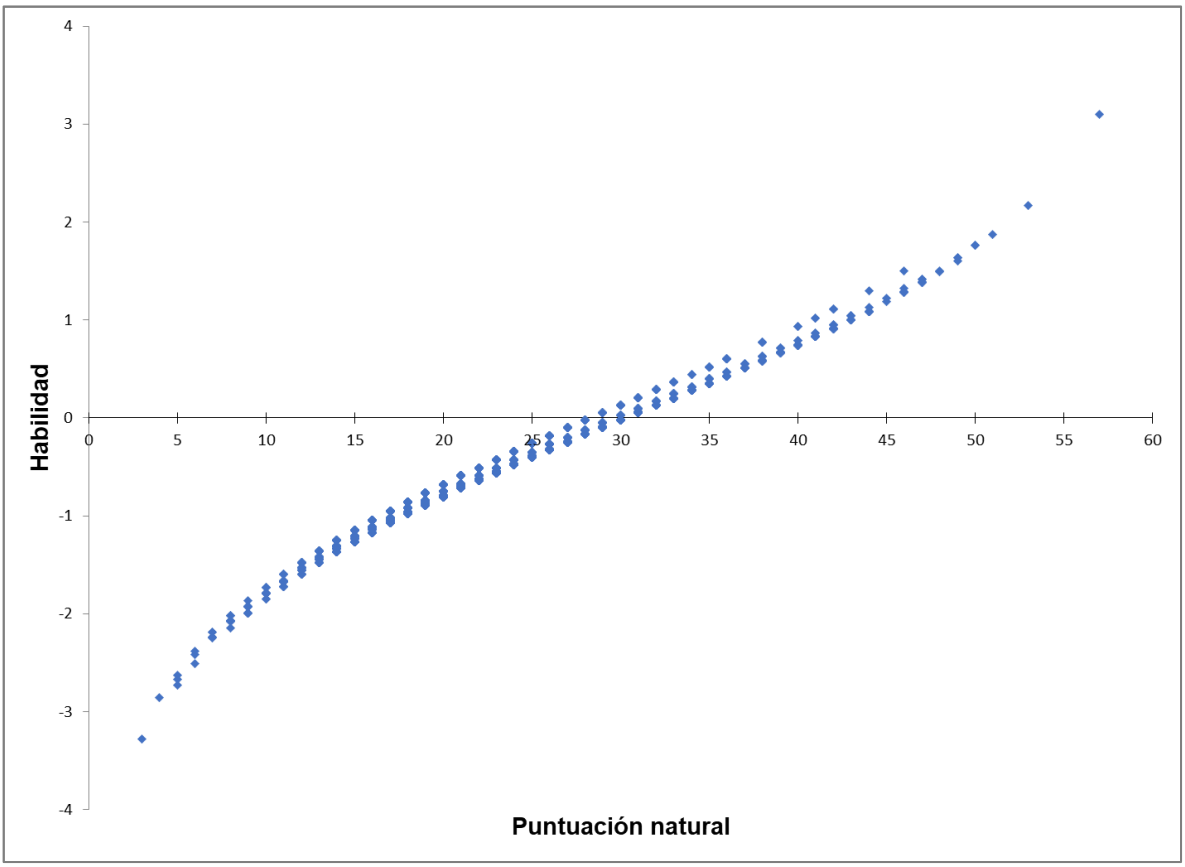

Figura 4. Distribución de la puntuación de los sujetos por modelo de estimación.

Prueba t para muestras relacionadas**.** Se analizó la diferencia de la media de las puntuaciones de la habilidad de los sujetos, estimada con TCT y con TRI, identificando una diferencia estadísticamente significativa entre ambas estimaciones ($p < .05$). En la tabla 12 se presenta la media estimada en cada modelo y la prueba t para diferencia de medias. El nivel de significancia de la prueba *t*, sugiere que hay diferencia entre las medias de las puntuaciones, lo cual es posible dada la calibración de los puntajes.

Tabla 12.
Estadísticos y prueba t de muestras relacionadas

N	Modelo	Media	Desv. Est.	*t*	*gl*	*p*
2898	TCT	.374	.123	-102.297	2987	.000
	TRI	-.585	.627			

Fuente: Elaboración propia. *Nota: gl* = Grados de libertad; *p* = Nivel de significancia

Finalmente, con el fin de documentar la importancia de la calibración de los puntajes, se realizó un análisis descriptivo de las posiciones que ocupan los aspirantes según el modelo de estimación. Se calcularon las frecuencias por cuartiles Q, en ambas estimaciones el Q_1 representa la posición con las puntuaciones más bajas y el Q_4, las más altas. Al calibrar las puntuaciones se obtuvieron los siguientes resultados: El 6.3 % de los aspirantes baja del Q_4 al Q_3, lo que indica que se estaba beneficiando a 48 estudiantes, que sin calibrar sus puntuaciones quedaban en el punto de corte con las puntuaciones más altas.

En el caso del Q_3, el 5.1% pasa al Q_4 y el 7.7% al Q_2, resultado que señala que se estaba penalizando a 37 aspirantes y beneficiando a 54. Respecto al Q_2, se identificó que el 5.8 % baja al Q_3 y el 3.7% sube un cuartil, lo que significa que se penalizaba a 44 aspirantes

y beneficiando a 26, estos últimos ocupaban la posición del segundo punto de corte y al calibrar su puntuación se quedan en el primero. Finalmente, en el Q_1 el 4.1% pasa al segundo cuartil, es decir, se beneficia a 29 estudiantes que estaban el último lugar. En la tabla 13 se aprecia este cambio en la proporción de los aspirantes por cuartil según el modelo de estimación.

Tabla 13
Clasificación de aspirantes por cuartiles en función del modelo de estimación

		TCT				
		Q_4	Q_3	Q_2	Q_1	Total
TRI	Q_4	694	37	-	-	731
	Q_3	48	665	44	-	757
	Q_2	-	54	616	29	699
	Q_1	-	-	26	685	711
Total		742	756	686	714	2898

Fuente: Elaboración propia. *Nota:* Q = Número de cuartil

Discusión

La presente investigación tuvo como objetivo equiparar las puntuaciones de las versiones alternativas de la prueba de matemáticas del Examen de Ingreso a la Educación Superior (ExIES). Para lograr dicho objetivo se trabajó en apego a los lineamientos y estándares de calidad de test de medición propuestos por la *American Psychological Association*, la *American Educational Research Association* y el *National Council on Measurement in Education* (APA, AERA y NCME, 2014) y por la *International Test Comission* (ITC, 2013) y en la metodología para equiparación de puntuaciones propuesta por Kolen y Brennan (2014).

Con base en la metodología empleada y los resultados obtenidos, el objetivo general de este estudio se cumplió satisfactoriamente. Al equiparar las puntuaciones de la prueba se aportó evidencia de validez de la estructura interna del ExIES, así como de la calidad psicométrica de sus ítems. También, se documentó el proceso metodológico empleado para estudios en esta línea de investigación, misma que reitera la necesidad de equiparar puntuaciones cuando estas provienen de diferentes versiones o formatos pero que pretenden evaluar el mismo rasgo o atributo. En el caso particular de las pruebas de acceso a algún programa educativo, se identifica la ventaja que tiene realizar análisis de equiparación para garantizar equidad y justicia a los examinados, observándose una importante diferencia en la estimación de su puntaje después de calibrar los ítems en la misma escala.

Para lograr el objetivo general se utilizó un diseño de equiparación de grupos no equivalentes con ítems comunes o de anclaje, lo que significó que las cuatro versiones compartieran un conjunto de 10 ítems que se tomaron como referencia para equiparar el resto de los ítems que las conformaban, además estos ítems sirvieron como base principal para analizar la calidad métrica de la prueba y el efecto de la calibración en las puntuaciones.

El procedimiento metodológico se realizó en dos fases. En la primera fase, y como requisito previo a la aplicación de la técnica estadística indicada en la equiparación de puntuaciones, se analizaron las propiedades psicométricas de las distintas versiones de la prueba de matemáticas, con la finalidad de asegurar que las cuatro versiones cumplían con criterios de calidad psicométrica. La segunda fase consistió en equiparar las puntuaciones de las cuatro versiones de la prueba con el método de calibración concurrente, método que calibra el parámetro de dificultad de los ítems.

En lo general, los análisis realizados en la primera fase registraron resultados satisfactorios en cuanto a las propiedades psicométricas de la prueba y sus ítems. Respecto a la evidencia de confiabilidad de las puntuaciones, los ítems presentaron índices de consistencia interna superiores al .70 en las cuatro versiones, que de acuerdo con los criterios de valoración sugeridos por González y Pasmiño (2015), se interpretan como índices de confiabilidad altos.

Por su parte, los ítems de las versiones A, C y D presentaron coeficientes de discriminación promedio >.30, que según Ebel y Frisbie (1986), sugieren un buen poder de discriminación, mientras que la versión B presentó un índice promedio <.27, lo que, de acuerdo con estos autores, indica una discriminación regular. Cabe mencionar que los autores del ExIES han optado por conservar estos ítems, utilizando como criterio que los ítems

registren índices de discriminación ≥ .20 (Caso, Díaz, Castro y Martínez, 2017), lo cual se cumple en la totalidad de los ítems de las cuatro versiones, las cuales presentan índices de discriminación promedio similares, tanto en el conjunto de ítems únicos como en los ítems de anclaje. Al respecto, Abad, Olea y Ponsoda (2007), sostienen que el índice de discriminación informa sobre el grado que en que un ítem en particular mide lo mismo que evalúa la prueba globalmente, por lo que los resultados obtenidos sugieren que los ítems de las cuatro versiones realizan un aporte significativo a la consistencia interna de la prueba.

En relación con el índice de dificultad de los ítems, los resultados indican una adecuada distribución en los tres niveles de dificultad (alto, medio y bajo) propuestos por Abad et. al, (2007). Del conjunto de ítems únicos, el 8% (4 ítems) presentaron valores atípicos en las diferentes versiones (en dos versiones mostraron índices bajos y en dos versiones valores altos), lo que sugirió la necesidad de equiparar el puntaje de las cuatro versiones para ajustar el sesgo que puedan concentrar. De maneral general se observó que todos los ítems, tanto únicos como de anclaje, presentaron un índice de dificultad >.20, criterio propuesto por los autores del examen para conservarlos.

Por otro lado, se examinó la unidimensionalidad de las cuatro versiones mediante el análisis paralelo modificado (MPA). De acuerdo con los resultados de este análisis, las cuatro versiones presentan la propiedad de unidimensionalidad, al agrupar el conjunto de datos en el primer factor ($p < .05$). La literatura especializada en esta materia señala que resulta indispensable que en todo proceso de equiparación de puntuaciones las versiones sean equivalentes en contenido y especificaciones y que se asegure que están midiendo el mismo constructo (Holland y Dorans, 2006; Kolen y Brennan, 2014; Von Davier, 2011), al ser

equivalentes las cuatro versiones en la propiedad de unidimensionalidad se cumple con este criterio.

Posteriormente, se estimó la invarianza poblacional de los ítems en cada prueba mediante el funcionamiento diferencial del ítem (DIF) en función del sexo, encontrando que solo un ítem común presentó DIF en las cuatro versiones (ítem 2X), lo que exigirá la revisión de su contenido y el de sus opciones de respuesta a fin de realizar los ajustes que resulten pertinentes para su uso en futuras aplicaciones. En general, las cuatro versiones de la prueba no presentan un número cuantioso de ítems únicos con DIF de nivel severo que representarían fuentes significativas de sesgo para el examen.

Los resultados de los análisis descritos con anterioridad cumplieron satisfactoriamente con los criterios y valores exigidos como prerrequisito a la aplicación de la técnica de equiparación. En lo particular, se obtuvo evidencia en las cuatro versiones de la prueba, de la equivalencia de constructo, de los índices de confiabilidad y dificultad y de su invarianza poblacional, propiedades exigidas en los procesos de equiparación de acuerdo con Kolen y Brennan (2014).

Con relación a la segunda fase, al calibrar los ítems en la misma escala, se obtuvo la estimación del parámetro de dificultad bajo el modelo de Rasch, con índices de ajuste interno aceptables (con una media de .95 en INFIT y .94 en OUFIT). En cuanto al índice de dificultad, el modelo de Rasch propone un criterio de $-\infty$ a $+\infty$ *lógitos*, aunque sugieren situar la dificultad media en ±.5 para conservar los ítems (Prieto y Delgado, 2003), por lo que pudo asegurarse que la calibración de los ítems controló el efecto de la dificultad en los mismos, al presentar una media de 0.052 *lógitos.*

Al comparar la cantidad de ítems en cada nivel de dificultad en función del modelo de estimación, se observó que la cantidad ítems considerados difíciles en TCT, disminuye al calibrarlos; al calibrar los ítems considerados fáciles en TCT, su nivel de dificultad se ajusta a dificultad media. Resultados que señalan la importancia de calibrar los ítems para ajustar sus parámetros.

Se comparó la distribución de las puntuaciones de los sujetos estimadas en ambos modelos, encontrado diferencia en la curtosis de la curva. En el caso de las puntuaciones calibradas se observa una distribución con menor variabilidad, de manera que los datos se distribuyen con mayor normalidad. Asimismo, se aplicó la prueba *r* para estimar la correlación entre los puntajes de los sujetos calculados con TCT y con Rasch, observándose una correlación positiva alta en las cuatro versiones ($r > .95$, $p < .05$), es decir, se identifica una relación lineal fuerte entre ambas estimaciones.

Respecto a la comparación de la media de las puntuaciones, se observó que existe diferencia estadísticamente significativa ($t = -102.297$, $p > .05$, $gl = 2897$). Estos resultados sugieren que, al existir diferencia significativa, el parámetro calibrado modifica la puntuación de los sujetos y con ello la posibilidad de pertenecer a diferentes puntos de corte. De tal modo que, al analizar la posición de los sujetos por punto de corte, se observan importantes diferencias al comparar las posiciones a partir del modelo de estimación. Este hallazgo proporciona evidencia significativa de la importancia de calibrar las puntuaciones, a fin de no penalizar injustamente la posición de los sujetos con respecto a su resultado.

Los resultados de este estudio se suman a los referidos en otras investigaciones en las que se ha documentado la equiparación de pruebas que son aplicadas en diferentes formatos; estudios en los que se ha comprobado que al aplicar procedimientos de equiparación se

obtienen pruebas con puntajes similares en cuanto a su media, distribución, parámetros, índices de ajuste y error (Jiménez, 2016; Lancheros, 2013; Ricker y Davier, 2007 y Rodríguez, 2007).

Por su parte, el desarrollo del presente estudio permitió demostrar la importancia de asegurar las condiciones previas al proceso de equiparación. En este sentido, el cuidado en el diseño, desarrollo y administración de las diferentes versiones dará como resultados índices de calidad psicométrica que facilitaran los procedimientos de equiparación. Por otro lado, fue posible ilustrar la utilidad de este tipo de estudios en aspectos relacionados con la seguridad de la prueba, de modo que al aplicar diferentes versiones se controla la posibilidad de copia entre los examinados y se propicia la construcción de un banco de ítems para futuras aplicaciones al contar con mayor cantidad de ítems para construir versiones.

La presente investigación registra algunas limitaciones. El hecho de disponer solo de 724 examinados en promedio por versión, impidió la aplicación de técnicas de equiparación más robustas, lo que hubiera permitido proponer un modelo logístico de al menos dos parámetros, los cuales exigen muestras más grandes para lograr un mejor ajuste (Abad et. al, 2011). Lo anterior sugiere que para réplicas futuras de este estudio se considere muestras de mayor tamaño.

Por otro lado, al no contar con la libertad para incidir en el diseño de la prueba y para determinar el tamaño de la muestra, se dificultó el análisis de equiparación con otros procedimientos basados en TRI, lo que permitiría comparar diferentes métodos de equiparación para contrastar sus parámetros, índices de ajuste y de error. Lo anterior proporcionaría mayor precisión en los puntajes a fin de seleccionar el método de equiparación más apropiado.

Estudios futuros podrían orientarse hacia la comparación de los parámetros en tres momentos: (1) analizar los parámetros con el modelo de TCT; (2) calibrar los parámetros de los ítems con TRI; (3). Depurar los ítems en función de su comportamiento en el análisis de unidimensionalidad y del DIF. De esta manera, sería posible identificar la aportación de la calibración al ajuste de las puntuaciones, asimismo, se esperaría que a medida que se depuran las versiones, eliminando los ítems con poco aporte o que presenten evidencia de invarianza poblacional las puntuaciones serán mayormente equivalentes.

Una sugerencia respecto a los análisis por versión es analizar la propiedad de invarianza poblacional en función de otras variables además del sexo, por ejemplo, en estudios similares sería importante analizar las diferencias por carrera o área de conocimiento, bachillerato de origen y sede de aplicación, entre otras, análisis que aportaría mayor evidencia de equidad y justicia a los aspirantes. Otra recomendación importante es replicar el estudio de equiparación en el resto de las pruebas que conforman en ExIES, así como al examen en su totalidad, ejercicio que permitirá obtener mayor evidencia de validez del examen.

A manera de conclusión, es posible afirmar que el objetivo general de este estudio se cumplió de manera satisfactoria. Los resultados de este proporcionan evidencia en cuanto a la equivalencia de las cuatro versiones en función de la similitud de los índices de dificultad, de la discriminación de los ítems, su consistencia interna y unidimensionalidad. Aunado a ello, los índices de ajuste interno en la calibración son similares para todos los ítems, la distribución de las puntuaciones calibradas presenta menor variabilidad y se corrigen posibles errores al posicionar a los examinados en el punto de corte correspondiente a su nivel de habilidad para responder la prueba.

La documentación de las distintas fases de este estudio permitió, además de documentar la metodología asociada con los procesos equiparación -línea de investigación de incipiente desarrollo en nuestro país-, llamar la atención en torno a la importancia que tienen este tipo de prácticas en el fomento de la equidad en las evaluaciones y de la justicia para los examinados, valores que deben permear los procesos relacionados con la aplicación de exámenes de selección y en cualquier medición que sea realizada con versiones alternativas o en formatos diferentes.

Referencias

Abad, F., Olea, J. y Ponsoda, V. (2007). *Introducción a la psicometría. Teoría Clásica de los Test y Teoría de Respuesta al Ítem.* Universidad Autónoma de Madrid, Madrid, España.

Abad, F., Olea, J., Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud.* España: Síntesis

American College Testing. (2016). What ACT Does. EE.UU. Recuperado de http://www.act.org/content/act/en/about-act.html

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

Antillón, L., Larrazolo, N. y Backhoff, E. (2008). Igualación lineal de tres versiones del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Iberoamericana de Evaluación Educativa*. 1 (2), 193-203.

Atkinson, A., Ariza, F. y García-Balboa, L. (2007). Estimadores robustos: una solución en la utilización de valores atípicos para el control de la calidad posicional. *GeoFocus. 7*, 171-187.

Backhoff, E. (2001). *Desarrollo, validación e implementación de un sistema para la selección de estudiantes a la Universidad Autónoma de Baja California.* (Tesis doctoral). Universidad Autónoma de Aguascalientes, Aguascalientes, México.

Backhoff, E., Larrazolo, N., y Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). REDIE. *Revista Electrónica de Investigación Educativa*, 2 (1), 11-28

Caso, J., Díaz, C., Castro, M. y Martínez, R. (2017). *Manual Técnico del Examen de Ingreso a la Educación Superior*. México: Universidad Autónoma de Baja California.

Centro Nacional de Evaluación. (2017). Marco de referencia. México. Recuperado de http://www.ceneval.edu.mx/ceneval-web/content.do?page=5757

College Board. (2016). About the SAT Suite of Assessments. EE.UU. Recuperado de https://collegereadiness.collegeboard.org/about

Dorans, N., Moses, T. & Eignor, D. (2010). *Principles and Practices of Test Score Equating.* Research Report. Educational Testing Service.

Dorans, N., Pommerich, M. & Holland, P. (2007). Linking and Aligning Scores and Scales. USA: Springer

Drasgow, F. & Lissak, R. (1983). Modified Parallel Analysis: A Procedure for Examining the Latent Dimensionality of Dichotomously Scored Item Responses. *Journal of Applied Psychology*. 68 (3), 363-373

Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of Education Measurement*. Englewood Cliffs, NJ: Prentice Hall.

Edwards, D., Coates, Hamish & Friedman, T. (2012). A Survey of International Practice in University Admissions Testing. Higher Education Management and Policy. 24 (1), 2-18

Freiberg, A., Stover, J., De la Iglesia, G. y Fernández, M. (2013). Correlaciones policóricas y tetracóricas en estudios factoriales exploratorios y confirmatorios. *Ciencias psicológicas,* 7(2), 151-164

Gempp, R. (2010). Equiparación, alineamiento y predicción de puntuaciones en medición educativa. *Revista Iberoamericana de Evaluación Educativa.* 3(2), 103-126

González, F. (2007). *Instrumentos de evaluación psicológica.* La Habana, Cuba: Ecmed

González, F. (2011). *Detección del Funcionamiento Diferencial del Ítem en Test Adaptativos Informatizados*. (Tesis doctoral). Universidad Autónoma de Madrid, Madrid, España.

González, J. & Wiberg, M. (2017). Applying Test Equating Methods. Using R. doi: 10.1007/978-3-319-51824-4

González, J. y Pazmino, M. (2015). Cálculo e interpretación del Alfa de Cronbach para el caso de validación de la consistencia interna de un cuestionario, con dos posibles escalas tipo Likert. *Revista Publicando*, 2(1), 62-77

Guilera, G., Gómez-Benito, Hidalgo, M. y Sánchez-Meca, J. (2007). Un meta-análisis del procedimiento Mantel-Haenszel en la detección del DIF en ítems dicotómicos. *Anuario de psicología.* 38(3), 431-442

Holland, P. & Dorans, N. (2006). A handbook on Measurement assessment, and evaluation in higher education. *Linking and equating.* Educational Testing Service

International Test Commission (2013). International Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores. En www.intestcom.org

Jiménez, E. (2016). *Propuesta metodológica para el estudio de la Equivalencia entre dos versiones de una prueba: Funcionamiento diferencial de versiones Utilizando propensity score* (tesis doctoral). Universidad Complutense de Madrid. Madrid, España

Kim, J. (2007). A comparison of calibration methods and proficiency estimators for creating IRT vertical scales. *Iowa Research Online.* University of Iowa

Kolen, M. & Brennan, R. (2014). Test Equating, Scaling and Linking. Methods and Practices. doi: 0.1007/978-1-4939-0317-7

Lancheros, L. (2013). *Métodos de equiparación de puntuaciones: los exámenes de estado en población con y sin limitación visual* (tesis doctoral). Universidad Nacional de Colombia, Colombia

Livingston. S. (2014). Equating Test Scores (without IRT) 2ed. Educational Testing Service.

López, José (1995). Estimación de parámetros en la TRI: Una evaluación de Bilog en muestras pequeñas. *Psicothema.* 7 (1), 173-185

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

Magis, D., Beland, S., Tuerlinckx, F. & De Boeck, P (2010). A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods, 42, 847-862.

Martínez, R., Lloreda, J. y Lloreda V. (2006). *Psicometría.* España: Alianza Editorial

Métrica Educativa A.C. (2017). Examen de Habilidades y Conocimientos Básicos (EXHCOBA). México. Recuperado de http://metrica.edu.mx/examenes/exhcoba/

Mohana, P., McHorney C., Stump, T. & Pekins, A. (2007). Odds Ratio, Delta, ETS Classification, and Standardization Measures of DIF Magnitude for Binary Logistic Regression. *Journal of Educational and Behavioral Statistics.* 32(1), 92-109

Navas, M. (2000). Equiparación de puntuaciones: Exigencias actuales y retos de cara al futuro. *Metodología de las ciencias del comportamiento.* 2(2), 151-165

Noble, J. & Wayne, J. (2003). Issues in College Admissions Testing. Department Of Education.

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

Oliden, P., & Zumbo, B. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema*, 20 (4), 896-901

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

Pagano, R. (2008). *Estadística para las ciencias del comportamiento y de la salud*. México. Cengage Learning Editores

Prieto, G y Delgado, A (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*. 15 (1), 94-100

Raju, N. & Areson, E. (2002, abril). *Developing a Common Metric in Item Response Theory: An Area-Minimization Approach.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, EEUU.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.10.

Ricker, K. & Von Davier, A. (2007). *The Impact of Anchor Test Length on Equating Results in a Nonequivalent Groups Desing.* Research Report. Educational Testing Service.

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, Journal of Statistical Software, 17 (5), 1-25. URL http://www.jstatsoft.org/v17/i05/

Rodríguez, D. (2016). *Evaluación empírica de dos métodos de equiparación: buscando equidad en la evaluación de personas con limitación visual* (tesis de maestría). Universidad Nacional de Colombia, Colombia

Rodríguez, O. (Septiembre, 2007). Equiparación de puntuaciones con tri y tct en una prueba de ingeniería. En 2ª Reunión Regional Norte, Centro América y Caribe de Evaluación Educativa.

Sánchez-Mendiola, M. y Delgado-Maldonado, L. (2016). Exámenes de alto impacto: implicaciones educativas. *Investigación en educación médica.* 6(21), 52-62

Silva, F. y Santelices, V. (2016). Funcionamiento Diferencial del Ítem en una Evaluación Estandarizada según Necesidades Educativas Especiales Transitorias. *Iberoamericana de Evaluación Educativa*. 9 (1), 145-160

Von Davier, A. (2011). Statistical Models for Test Equating, Scaling, and Linking. doi: 10.1007/978-0-387-98138-3

Zwick, R. (2007). A handbook on Measurement assessment, and evaluation in higher education. *Higher Educational Admissions Testing.* Educational Testing Service

Apéndices

Apéndice A. Coeficiente de discriminación de los ítems

Ítem	A	B	C	D
1	.45	.43	-.03	.56
2	.33	.39	.31	.46
3	.41	.49	.04	.54
4	.20	.04	.49	.24
5	.32	.34	.56	.24
6	.42	.31	.43	.24
7	.58	.34	.39	.55
8	.49	-.08	.53	.39
9	.11	.23	.42	.35
10	.21	.40	.28	.43
11	.18	.41	.30	.27
12	.57	.18	.54	.32
13	.49	.28	.37	.42
14	.48	.36	.35	.31
15	.38	.25	.63	.42
16	.58	.34	.18	.50
17	.43	.43	.15	.36
18	.45	.43	-.03	.35
19	.46	.29	.04	-.01
20	.13	-.09	.32	.07
21	.24	-.15	.06	.55
22	.39	.29	.18	.47
23	.38	.36	.39	.42
24	.23	.32	.12	.08

25	.36	.48	.17	.34
26	-.04	.07	.21	.48
27	.07	.03	.34	.28
28	.28	.25	.38	.09
29	.26	.35	.56	.52
30	.50	-.01	.25	.30
31	.52	-.03	.09	.24
32	.50	.26	.46	.34
33	.48	.35	.41	.37
34	.46	.41	.47	.36
35	.39	-.07	.11	.20
36	.52	.13	.59	.42
37	.54	.38	.50	.36
38	.44	.12	.31	.42
39	.07	.27	.05	.30
40	.19	.22	.16	.26
41	.51	.22	.17	.50
42	.37	.23	.20	.01
43	.48	.36	.59	.49
44	.26	.13	.17	.06
45	.19	.20	.10	.13
46	.24	.32	.20	.46
47	.15	.02	.25	.13
48	.42	.35	.38	.51
49	.00	.09	.23	.06
50	.11	.30	.23	.09
1X	.52	.54	.55	.56
2X	.53	.56	.51	.48
3X	.49	.43	.40	.38
4X	.39	.35	.44	.38
5X	.41	.40	.51	.38

6X	.49	.46	.55	.55
7X	.49	.45	.52	.60
8X	.31	.27	.25	.34
9X	.49	.49	.36	.32
10X	.37	.42	.43	.31
Ẋ	0.36	0.27	0.32	0.34

Apéndice B. Índice de dificultad de los ítems

Ítem	A	B	C	D
1	.59	.56	.29	.21
2	.58	.68	.46	.33
3	.21	.54	.70	.60
4	.26	.19	.52	.28
5	.60	.50	.45	.47
6	.44	.54	.35	.19
7	.41	.58	.33	.37
8	.67	.39	.39	.50
9	.30	.15	.81	.30
10	.38	.17	.25	.15
11	.43	.51	.62	.47
12	.52	.37	.37	.34
13	.72	.23	.40	.90
14	.47	.60	.36	.35
15	.49	.32	.50	.60
16	.43	.63	.33	.62
17	.39	.35	.29	.39
18	.78	.12	.20	.28
19	.20	.18	.46	.48
20	.24	.30	.18	.30
21	.31	.20	.21	.45
22	.67	.61	.38	.61
23	.48	.85	.89	.56
24	.24	.44	.21	.28
25	.39	.76	.81	.84
26	.31	.28	.26	.29

27	.18	.18	.58	.25
28	.27	.23	.47	.27
29	.37	.57	.64	.53
30	.51	.21	.38	.45
31	.23	.12	.16	.36
32	.55	.28	.39	.37
33	.38	.33	.47	.36
34	.31	.34	.53	.29
35	.43	.07	.25	.27
36	.52	.27	.34	.28
37	.38	.45	.37	.31
38	.30	.16	.31	.30
39	.36	.25	.33	.33
40	.21	.23	.30	.31
41	.42	.15	.24	.38
42	.35	.33	.35	.23
43	.32	.31	.25	.25
44	.24	.24	.39	.19
45	.26	.23	.25	.33
46	.28	.28	.27	.26
47	.17	.27	.56	.23
48	.28	.53	.47	.57
49	.24	.53	.31	.19
50	.24	.15	.15	.19
1X	.40	.38	.46	.46
2X	.46	.48	.50	.49
3X	.21	.21	.24	.20
4X	.20	.21	.24	.22
5X	.33	.33	.37	.36
6X	.50	.54	.50	.50
7X	.34	.43	.41	.38

8X	.34	.36	.30	.29
9X	.32	.28	.27	.27
10X	.42	.43	.41	.45
$\dot{X}$	0.38	0.36	0.39	0.37

Apéndice C. Efecto del funcionamiento diferencial del ítem por versión

Ítem	A		B		C		D	
	DIF	*Efecto*	*DIF*	*Efecto*	*DIF*	*Efecto*	*DIF*	*Efecto*
1	-1.2273	B	0.9399	A	0.145	A	0.4385	A
2	0.0728	A	0.0898	A	-0.5387	A	-0.4539	A
3	-0.1232	A	-0.188	A	-0.0869	A	-0.6396	A
4	0.1956	A	0.6382	A	-0.3813	A	-0.7895	A
5	-0.0351	A	-0.278	A	0.5357	A	0.006	A
6	-0.7797	A	0.243	A	-0.0926	A	0.8928	A
7	-0.1197	A	-0.3416	A	0.1912	A	0.4407	A
8	-0.5964	A	0.9366	A	0.2668	A	-0.179	A
9	0.1182	A	-0.4009	A	-0.7678	A	0.1884	A
10	-0.589	A	-0.1288	A	-0.3828	A	-0.2482	A
11	-0.4501	A	-0.5397	A	-1.018	B	-0.5351	A
12	-0.7928	A	-0.4287	A	-0.6756	A	-0.4297	A
13	-0.3741	A	-0.2259	A	-0.6207	A	-0.8803	A
14	-0.0502	A	-0.4789	A	-0.3854	A	0.0975	A
15	0.3618	A	-0.7128	A	-0.3598	A	-0.8184	A
16	-0.9038	A	-0.7148	A	-1.1318	B	-0.0174	A
17	0.0691	A	-1.7613	C	0.0762	A	0.0881	A
18	0.1325	A	0.0376	A	-0.0823	A	0.8431	A
19	1.5648	C	0.0327	A	0.6222	A	-0.0487	A
20	0.3429	A	0.5579	A	-0.9277	A	-0.416	A
21	0.5052	A	0.856	A	1.3672	B	-0.5038	A
22	-0.3336	A	-0.1847	A	0.4029	A	-0.4126	A
23	0.4968	A	0.4933	A	0.6676	A	0.0003	A
24	-0.5295	A	-0.2599	A	0.4597	A	-0.5303	A
25	-0.2048	A	-0.4956	A	-0.284	A	0.4546	A
26	0.8816	A	0.4142	A	0.2007	A	0.595	A

27	1.1803	B	0.0217	A	0.733	A	-0.0988	A
28	0.4579	A	-0.4885	A	0.1534	A	-0.0003	A
29	-0.0183	A	0.3778	A	-0.2979	A	0.463	A
30	0.5885	A	0.2022	A	-0.2365	A	-0.0067	A
31	-0.7492	A	-0.2046	A	-0.022	A	0.5327	A
32	0.0674	A	0.0809	A	0.7054	A	0.3757	A
33	0.4352	A	0.0246	A	1.0026	B	1.023	B
34	0.045	A	0.1154	A	-0.2417	A	-0.5574	A
35	-0.0334	A	-0.0587	A	-0.0804	A	-0.0338	A
36	0.2449	A	0.9033	A	0.7112	A	-0.6928	A
37	-0.4596	A	0.0669	A	0.8546	A	0.9277	A
38	-0.5582	A	0.2839	A	0.0319	A	-0.1207	A
39	0.5937	A	0.0433	A	-0.0413	A	0.2311	A
40	-0.2028	A	0.6786	A	0.0358	A	1.1014	B
41	-0.5358	A	-0.3421	A	-0.3581	A	0.4408	A
42	0.6633	A	0.0522	A	-0.0778	A	-0.0992	A
43	0.8715	A	0.1672	A	-0.1954	A	-0.5544	A
44	0.5385	A	-0.1554	A	0.1349	A	1.3096	B
45	0.9276	A	0.4883	A	0.0623	A	-0.1458	A
46	-0.1272	A	-0.4048	A	-0.0966	A	-0.0145	A
47	-0.4356	A	0.4549	A	-0.6225	A	0.5742	A
48	-0.299	A	-0.3366	A	0.5391	A	0.0126	A
49	0.0153	A	0.9624	A	0.3186	A	-0.1481	A
50	0.5039	A	0.7088	A	-0.4264	A	-0.2794	A
1X	0.059	A	0.0566	A	0.4487	A	0.4087	A
2X	-1.0998	B	-2.2248	C	-1.524	C	-1.044	B
3X	-0.3941	A	1.4915	B	0.1871	A	0.2246	A
4X	-0.369	A	0.2292	A	-0.1234	A	0.108	A
5X	0.2893	A	-0.055	A	0.2632	A	0.4371	A
6X	0.2062	A	-0.69	A	0.1522	A	-0.4744	A
7X	-0.6408	A	-0.3811	A	0.1271	A	-0.8372	A

8X	-0.4704	A	-0.1096	A	-0.0856	A	-0.0106	A
9X	1.6122	C	0.9171	A	0.9059	A	0.5913	A
10X	0.0523	A	-0.159	A	-0.0802	A	-0.4076	A

Apéndice D. Parámetro de dificultad en el modelo de Rasch

Ítem	MEASURE	COUNT	INFIT.MSQ	OUTFIT.MSQ	MODLSE	SCORE
1	-1.00978	759	0.9485	0.9277	0.07713	449
2	-0.96825	759	1.0076	1.0145	0.07691	442
3	0.88368	759	0.9806	1.0174	0.0939	156
4	0.53633	759	1.0918	1.1316	0.08656	199
5	-1.03363	759	1.0054	1.0117	0.07726	453
6	-0.34436	759	0.9749	0.9707	0.07679	335
7	-0.18356	759	0.8863	0.8747	0.07769	308
8	-1.36046	759	0.9234	0.8802	0.08003	506
9	0.32138	759	1.1368	1.1875	0.08306	229
10	-0.06774	759	1.0898	1.1005	0.07858	289
11	-0.29116	759	1.0997	1.1238	0.07705	326
12	-0.71105	759	0.881	0.8683	0.07616	398
13	-1.65605	759	0.9099	0.8666	0.08408	550
14	-0.49065	759	0.9349	0.9221	0.0763	360
15	-0.56615	759	0.9903	0.9764	0.07617	373
16	-0.29709	759	0.8888	0.875	0.07701	327
17	-0.09848	759	0.9714	0.9891	0.07832	294
18	-1.9904	759	0.927	0.8823	0.09056	594
19	0.91025	759	0.9584	1.0091	0.09456	153
20	0.67498	759	1.1298	1.1518	0.08923	181
21	0.27357	759	1.0711	1.0869	0.08238	236
22	-1.35404	759	0.9602	0.9639	0.07996	505
23	-0.53134	759	0.9866	0.9862	0.07622	367
24	0.66706	759	1.0775	1.1126	0.08907	182
25	-0.12293	759	1.0036	1.0049	0.07813	298
26	0.28716	759	1.2129	1.2734	0.08257	234

27	1.09783	759	1.1328	1.2583	0.09956	133
28	0.51401	759	1.0488	1.0692	0.08616	202
29	-0.00564	759	1.0569	1.0933	0.07914	279
30	-0.6357	759	0.9235	0.9139	0.07612	385
31	0.75595	759	0.9425	0.9068	0.09095	171
32	-0.80401	759	0.9234	0.8991	0.07631	414
33	-0.0739	759	0.9431	0.9286	0.07853	290
34	0.28036	759	0.958	0.9868	0.08247	235
35	-0.28523	759	0.9869	0.9906	0.07708	325
36	-0.67627	759	0.918	0.8952	0.07613	392
37	-0.0492	759	0.9129	0.9094	0.07874	286
38	0.34212	759	0.9699	0.978	0.08336	226
39	0.03838	759	1.1567	1.1786	0.07957	272
40	0.84881	759	1.0921	1.151	0.09307	160
41	-0.26145	759	0.9237	0.9046	0.07721	321
42	0.05739	759	1.0049	0.9972	0.07977	269
43	0.21985	759	0.95	0.963	0.08166	244
44	0.65128	759	1.0578	1.0971	0.08875	184
45	0.52887	759	1.0952	1.1263	0.08643	200
46	0.45543	759	1.0716	1.0756	0.08515	210
47	1.14798	759	1.0957	1.2265	0.10102	128
48	0.43381	759	0.9761	0.9958	0.08479	213
49	0.67498	759	1.1864	1.2815	0.08923	181
50	0.68294	759	1.1258	1.2059	0.0894	180
1	-0.83854	751	0.9485	0.93	0.07581	418
2	-1.38024	751	0.9605	0.9279	0.0803	508
3	-0.78123	751	0.9177	0.9036	0.07562	408
4	0.97659	751	1.0751	1.157	0.09616	139
5	-0.58777	751	0.9727	0.9738	0.07536	374
6	-0.76409	751	0.9876	0.9889	0.07558	405
7	-0.96003	751	0.9721	0.9693	0.0764	439

8	-0.13783	751	1.1434	1.174	0.07707	296
9	1.26984	751	1.0163	1.0232	0.10528	110
10	1.11154	751	0.9596	0.9674	0.10011	125
11	-0.65592	751	0.9409	0.9426	0.07539	386
12	-0.0418	751	1.0483	1.0621	0.07786	280
13	0.70207	751	1.0034	1.0607	0.0893	171
14	-1.00686	751	0.9641	0.9616	0.07668	447
15	0.22852	751	1.0148	1.0064	0.08089	237
16	-1.16799	751	0.9787	0.9777	0.07795	474
17	0.07511	751	0.9422	0.9308	0.07902	261
18	1.48464	751	0.9575	0.9481	0.11327	92
19	0.99522	751	0.9992	1.0312	0.09668	137
20	0.2814	751	1.14	1.1738	0.08163	229
21	0.87753	751	1.136	1.271	0.0935	150
22	-1.06594	751	0.9952	0.993	0.0771	457
23	-2.41318	751	0.9648	0.9243	0.10397	637
24	-0.31925	751	0.9942	0.9953	0.07599	327
25	-1.79736	751	0.9181	0.8556	0.08726	568
26	0.40421	751	1.0835	1.104	0.08353	211
27	0.99522	751	1.0775	1.158	0.09668	137
28	0.67825	751	1.0118	1.017	0.08878	174
29	-0.91353	751	0.9621	0.9776	0.07614	431
30	0.82572	751	1.0999	1.15	0.0922	156
31	1.55046	751	1.076	1.1982	0.11596	87
32	0.41121	751	1.0067	1.0225	0.08365	210
33	0.16372	751	0.9765	0.9686	0.08005	247
34	0.12546	751	0.9587	0.957	0.07959	253
35	2.1993	751	1.0534	1.3336	0.14931	49
36	0.45361	751	1.06	1.0758	0.08438	204
37	-0.39404	751	0.9588	0.9535	0.0757	340
38	1.15223	751	1.0399	1.1448	0.10139	121

39	0.54817	751	1.0076	1.0035	0.08611	191
40	0.71811	751	1.0231	1.0516	0.08966	169
41	1.28099	751	1.0127	1.0455	0.10567	109
42	0.13816	751	1.0172	1.014	0.07974	251
43	0.27473	751	0.9771	0.9841	0.08153	230
44	0.64692	751	1.0562	1.0945	0.08811	178
45	0.70207	751	1.0308	1.0613	0.0893	171
46	0.38334	751	0.9887	0.987	0.08319	214
47	0.45361	751	1.0983	1.1458	0.08438	204
48	-0.72418	751	0.9633	0.9633	0.07548	398
49	-0.72418	751	1.0716	1.0901	0.07548	398
50	1.23687	751	0.9984	0.9889	0.10416	113
1	0.42022	733	1.1584	1.2133	0.0847	211
2	-0.39747	733	1.0093	1.0095	0.07713	339
3	-1.4578	733	1.0969	1.1876	0.08306	511
4	-0.62834	733	0.9242	0.915	0.0769	378
5	-0.3498	733	0.8936	0.8817	0.07728	331
6	0.11398	733	0.9528	0.9575	0.0806	256
7	0.20611	733	0.9725	0.9701	0.08167	242
8	-0.09499	733	0.9082	0.8994	0.07868	289
9	-2.11303	733	0.966	0.864	0.09657	594
10	0.62188	733	1.0216	1.0367	0.08829	184
11	-1.0799	733	1.0097	1.002	0.07886	453
12	0.02408	733	0.9076	0.8998	0.07969	270
13	-0.10735	733	0.9767	0.9833	0.07859	291
14	0.06236	733	0.9951	0.9961	0.08006	264
15	-0.5456	733	0.8559	0.8429	0.07689	364
16	0.1928	733	1.0684	1.0914	0.08151	244
17	0.42022	733	1.0791	1.1062	0.0847	211
18	0.95965	733	1.1408	1.2108	0.09601	144
19	-0.36769	733	1.1333	1.1654	0.07722	334

20	1.1036	733	0.9922	1.0376	0.1	129
21	0.86089	733	1.0941	1.2487	0.09352	155
22	-0.02642	733	1.0695	1.082	0.07924	278
23	-2.74437	733	0.9667	0.8755	0.11857	650
24	0.85216	733	1.0883	1.15	0.09331	156
25	-2.12238	733	1.0275	1.1377	0.09683	595
26	0.56795	733	1.0458	1.1114	0.08726	191
27	-0.92057	733	0.9817	1.0226	0.07779	427
28	-0.40936	733	0.9706	0.9663	0.0771	341
29	-1.17409	733	0.8929	0.8475	0.07968	468
30	-0.05146	733	1.0379	1.036	0.07903	282
31	1.22815	733	1.0784	1.1717	0.10382	117
32	-0.09499	733	0.9427	0.932	0.07868	289
33	-0.45091	733	0.9615	0.9534	0.077	348
34	-0.68751	733	0.9323	0.9145	0.07697	388
35	0.63752	733	1.0938	1.1325	0.0886	182
36	0.15317	733	0.8826	0.8597	0.08104	250
37	-0.00124	733	0.9212	0.9047	0.07946	274
38	0.28716	733	1.0127	1.0082	0.08273	230
39	0.1928	733	1.1239	1.1493	0.08151	244
40	0.34941	733	1.0821	1.081	0.08361	221
41	0.68515	733	1.0709	1.1009	0.08957	176
42	0.10101	733	1.0602	1.0683	0.08046	258
43	0.65328	733	0.8903	0.8628	0.08892	180
44	-0.06394	733	1.073	1.0761	0.07892	284
45	0.61409	733	1.0955	1.1259	0.08814	185
46	0.53018	733	1.0569	1.089	0.08657	196
47	-0.82434	733	1.0286	1.0363	0.07735	411
48	-0.42125	733	0.9747	0.9636	0.07707	343
49	0.28716	733	1.0478	1.0365	0.08273	230
50	1.3398	733	1.032	1.0764	0.10755	107

1	0.84388	655	0.9117	0.876	0.10003	136
2	0.16744	655	0.9489	0.9459	0.08688	215
3	-1.03472	655	0.9024	0.8614	0.08291	390
4	0.42767	655	1.0468	1.1028	0.09093	182
5	-0.49773	655	1.0488	1.0538	0.08176	310
6	0.96805	655	1.0436	1.1578	0.10342	124
7	-0.03794	655	0.8999	0.9096	0.08452	243
8	-0.61111	655	0.9767	0.9761	0.08161	327
9	0.31434	655	1.001	1.0361	0.08902	196
10	1.25884	655	0.961	0.9994	0.1127	99
11	-0.47098	655	1.0342	1.0339	0.08183	306
12	0.10016	655	1.0163	0.9959	0.08603	224
13	-2.98121	655	0.9492	0.8468	0.13396	591
14	0.04871	655	1.0187	1.0318	0.08543	231
15	-1.05537	655	0.9582	0.938	0.08305	393
16	-1.12474	655	0.9141	0.8895	0.08357	403
17	-0.12988	655	0.9942	0.9902	0.08369	256
18	0.40297	655	1.0016	1.0003	0.09049	185
19	-0.5378	655	1.1664	1.2146	0.08168	316
20	0.33024	655	1.1332	1.1842	0.08927	194
21	-0.3769	655	0.8996	0.8788	0.08215	292
22	-1.09689	655	0.9266	0.9037	0.08335	399
23	-0.86491	655	0.9646	0.9448	0.08203	365
24	0.41941	655	1.1278	1.1478	0.09078	183
25	-2.42637	655	0.9778	0.9445	0.11041	553
26	0.3704	655	0.9398	0.9244	0.08994	189
27	0.56374	655	1.0307	1.0663	0.09354	166
28	0.45261	655	1.1214	1.1856	0.09138	179
29	-0.75103	655	0.9094	0.8953	0.08171	348
30	-0.38364	655	1.0197	1.0383	0.08212	293
31	0.00515	655	1.057	1.1007	0.08495	237

32	-0.03078	655	1.0058	1.0129	0.08459	242
33	-0.00206	655	0.9966	0.9842	0.08488	238
34	0.36232	655	1.0009	0.9891	0.0898	190
35	0.44426	655	1.0649	1.1086	0.09123	180
36	0.41941	655	0.9683	0.9678	0.09078	183
37	0.2828	655	1.0017	0.9994	0.08853	200
38	0.29065	655	0.9693	0.9672	0.08865	199
39	0.18258	655	1.0272	1.0421	0.08708	213
40	0.25161	655	1.046	1.0377	0.08806	204
41	-0.05932	655	0.9283	0.9123	0.08431	246
42	0.69911	655	1.1542	1.2287	0.09648	151
43	0.56374	655	0.935	0.9434	0.09354	166
44	0.94677	655	1.1206	1.184	0.10282	126
45	0.175	655	1.1009	1.1225	0.08698	214
46	0.52038	655	0.9536	0.9482	0.09267	171
47	0.70845	655	1.0978	1.1492	0.0967	150
48	-0.93918	655	0.9093	0.8784	0.08236	376
49	0.97879	655	1.1256	1.1831	0.10373	123
50	0.96805	655	1.0931	1.2425	0.10342	124
1X	-0.25637	2898	0.9032	0.8889	0.03919	1226
2X	-0.51831	2898	0.9059	0.897	0.03873	1399
3X	0.81373	2898	0.9651	0.9666	0.0469	621
4X	0.78533	2898	0.9761	0.9886	0.04657	634
5X	0.09759	2898	0.9613	0.9597	0.04067	1003
6X	-0.63214	2898	0.9117	0.8969	0.0387	1475
7X	-0.10242	2898	0.9151	0.8992	0.03971	1127
8X	0.21195	2898	1.0226	1.0313	0.04137	935
9X	0.40832	2898	0.9681	0.9541	0.04282	824
10X	-0.27937	2898	0.9759	0.9681	0.03913	1241

Apéndice E. Comparación del parámetro de dificultad de los ítems, según el modelo de estimación

	TCT		TRI	
ITEM	Puntuación	Nivel	Puntuación	Nivel
1	0.59	Media	-1.01	Media
2	0.58	Media	-0.97	Media
3	0.21	Difícil	0.88	Media
4	0.26	Difícil	0.54	Media
5	0.60	Media	-1.03	Media
6	0.44	Media	-0.34	Media
7	0.41	Media	-0.18	Media
8	0.67	Media	-1.36	Media
9	0.30	Media	0.32	Media
10	0.38	Media	-0.07	Media
11	0.43	Media	-0.29	Media
12	0.52	Media	-0.71	Media
13	0.72	Fácil	-1.66	Media
14	0.47	Media	-0.49	Media
15	0.49	Media	-0.57	Media
16	0.43	Media	-0.30	Media
17	0.39	Media	-0.10	Media
18	0.78	Fácil	-1.99	Difícil
19	0.20	Difícil	0.91	Media
20	0.24	Difícil	0.67	Media
21	0.31	Media	0.27	Media
22	0.67	Media	-1.35	Media
23	0.48	Media	-0.53	Media
24	0.24	Difícil	0.67	Media
25	0.39	Media	-0.12	Media

26	0.31	Media	0.29	Media
27	0.18	Difícil	1.10	Media
28	0.27	Difícil	0.51	Media
29	0.37	Media	-0.01	Media
30	0.51	Media	-0.64	Media
31	0.23	Difícil	0.76	Media
32	0.55	Media	-0.80	Media
33	0.38	Media	-0.07	Media
34	0.31	Media	0.28	Media
35	0.43	Media	-0.29	Media
36	0.52	Media	-0.68	Media
37	0.38	Media	-0.05	Media
38	0.30	Media	0.34	Media
39	0.36	Media	0.04	Media
40	0.21	Difícil	0.85	Media
41	0.42	Media	-0.26	Media
42	0.35	Media	0.06	Media
43	0.32	Media	0.22	Media
44	0.24	Difícil	0.65	Media
45	0.26	Difícil	0.53	Media
46	0.28	Difícil	0.46	Media
47	0.17	Difícil	1.15	Media
48	0.28	Difícil	0.43	Media
49	0.24	Difícil	0.67	Media
50	0.24	Difícil	0.68	Media
51	0.56	Media	-0.84	Media
52	0.68	Media	-1.38	Media
53	0.54	Media	-0.78	Media
54	0.19	Difícil	0.98	Media
55	0.50	Media	-0.59	Media
56	0.54	Media	-0.76	Media

57	0.58	Media	-0.96	Media
58	0.39	Media	-0.14	Media
59	0.15	Difícil	1.27	Media
60	0.17	Difícil	1.11	Media
61	0.51	Media	-0.66	Media
62	0.37	Media	-0.04	Media
63	0.23	Difícil	0.70	Media
64	0.60	Media	-1.01	Media
65	0.32	Media	0.23	Media
66	0.63	Media	-1.17	Media
67	0.35	Media	0.08	Media
68	0.12	Difícil	1.48	Media
69	0.18	Difícil	1.00	Media
70	0.30	Media	0.28	Media
71	0.20	Difícil	0.88	Media
72	0.61	Media	-1.07	Media
73	0.85	Fácil	-2.41	Difícil
74	0.44	Media	-0.32	Media
75	0.76	Fácil	-1.80	Media
76	0.28	Difícil	0.40	Media
77	0.18	Difícil	1.00	Media
78	0.23	Difícil	0.68	Media
79	0.57	Media	-0.91	Media
80	0.21	Difícil	0.83	Media
81	0.12	Difícil	1.55	Media
82	0.28	Difícil	0.41	Media
83	0.33	Media	0.16	Media
84	0.34	Media	0.13	Media
85	0.07	Difícil	2.20	Fácil
86	0.27	Difícil	0.45	Media
87	0.45	Media	-0.39	Media

88	0.16	Difícil	1.15	Media
89	0.25	Difícil	0.55	Media
90	0.23	Difícil	0.72	Media
91	0.15	Difícil	1.28	Media
92	0.33	Media	0.14	Media
93	0.31	Media	0.27	Media
94	0.24	Difícil	0.65	Media
95	0.23	Difícil	0.70	Media
96	0.28	Difícil	0.38	Media
97	0.27	Difícil	0.45	Media
98	0.53	Media	-0.72	Media
99	0.53	Media	-0.72	Media
100	0.15	Difícil	1.24	Media
101	0.29	Difícil	0.42	Media
102	0.46	Media	-0.40	Media
103	0.70	Media	-1.46	Media
104	0.52	Media	-0.63	Media
105	0.45	Media	-0.35	Media
106	0.35	Media	0.11	Media
107	0.33	Media	0.21	Media
108	0.39	Media	-0.09	Media
109	0.81	Fácil	-2.11	Difícil
110	0.25	Difícil	0.62	Media
111	0.62	Media	-1.08	Media
112	0.37	Media	0.02	Media
113	0.40	Media	-0.11	Media
114	0.36	Media	0.06	Media
115	0.50	Media	-0.55	Media
116	0.33	Media	0.19	Media
117	0.29	Difícil	0.42	Media
118	0.20	Difícil	0.96	Media

119	0.46	Media	-0.37	Media
120	0.18	Difícil	1.10	Media
121	0.21	Difícil	0.86	Media
122	0.38	Media	-0.03	Media
123	0.89	Fácil	-2.74	Difícil
124	0.21	Difícil	0.85	Media
125	0.81	Fácil	-2.12	Difícil
126	0.26	Difícil	0.57	Media
127	0.58	Media	-0.92	Media
128	0.47	Media	-0.41	Media
129	0.64	Media	-1.17	Media
130	0.38	Media	-0.05	Media
131	0.16	Difícil	1.23	Media
132	0.39	Media	-0.09	Media
133	0.47	Media	-0.45	Media
134	0.53	Media	-0.69	Media
135	0.25	Difícil	0.64	Media
136	0.34	Media	0.15	Media
137	0.37	Media	0.00	Media
138	0.31	Media	0.29	Media
139	0.33	Media	0.19	Media
140	0.30	Media	0.35	Media
141	0.24	Difícil	0.69	Media
142	0.35	Media	0.10	Media
143	0.25	Difícil	0.65	Media
144	0.39	Media	-0.06	Media
145	0.25	Difícil	0.61	Media
146	0.27	Difícil	0.53	Media
147	0.56	Media	-0.82	Media
148	0.47	Media	-0.42	Media
149	0.31	Media	0.29	Media

150	0.15	Difícil	1.34	Media
151	0.21	Difícil	0.84	Media
152	0.33	Media	0.17	Media
153	0.60	Media	-1.03	Media
154	0.28	Difícil	0.43	Media
155	0.47	Media	-0.50	Media
156	0.19	Difícil	0.97	Media
157	0.37	Media	-0.04	Media
158	0.50	Media	-0.61	Media
159	0.30	Media	0.31	Media
160	0.15	Difícil	1.26	Media
161	0.47	Media	-0.47	Media
162	0.34	Media	0.10	Media
163	0.90	Fácil	-2.98	Difícil
164	0.35	Media	0.05	Media
165	0.60	Media	-1.06	Media
166	0.62	Media	-1.12	Media
167	0.39	Media	-0.13	Media
168	0.28	Difícil	0.40	Media
169	0.48	Media	-0.54	Media
170	0.30	Media	0.33	Media
171	0.45	Media	-0.38	Media
172	0.61	Media	-1.10	Media
173	0.56	Media	-0.86	Media
174	0.28	Difícil	0.42	Media
175	0.84	Fácil	-2.43	Difícil
176	0.29	Difícil	0.37	Media
177	0.25	Difícil	0.56	Media
178	0.27	Difícil	0.45	Media
179	0.53	Media	-0.75	Media
180	0.45	Media	-0.38	Media

181	0.36	Media	0.01	Media
182	0.37	Media	-0.03	Media
183	0.36	Media	0.00	Media
184	0.29	Difícil	0.36	Media
185	0.27	Difícil	0.44	Media
186	0.28	Difícil	0.42	Media
187	0.31	Media	0.28	Media
188	0.30	Media	0.29	Media
189	0.33	Media	0.18	Media
190	0.31	Media	0.25	Media
191	0.38	Media	-0.06	Media
192	0.23	Difícil	0.70	Media
193	0.25	Difícil	0.56	Media
194	0.19	Difícil	0.95	Media
195	0.33	Media	0.18	Media
196	0.26	Difícil	0.52	Media
197	0.23	Difícil	0.71	Media
198	0.57	Media	-0.94	Media
199	0.19	Difícil	0.98	Media
200	0.19	Difícil	0.97	Media
201	0.42	Media	-0.26	Media
202	0.48	Media	-0.52	Media
203	0.21	Difícil	0.81	Media
204	0.22	Difícil	0.79	Media
205	0.35	Media	0.10	Media
206	0.51	Media	-0.63	Media
207	0.39	Media	-0.10	Media
208	0.32	Media	0.21	Media
209	0.28	Difícil	0.41	Media
210	0.43	Media	-0.28	Media