UNIVERSIDAD DE
COSTA RICA

FACULTAD DE EDUCACIÓN
ESCUELA DE FORMACIÓN DOCENTE

ANÁLISIS DEL FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM BAJO EL MARCO DE LA TEORÍA DE RESPUESTA AL ÍTEM EN LA PRUEBA DE APTITUD ACADÉMICA

Trabajo final de graduación sometido a la consideración de la
Escuela de Formación Docente

Sustentantes
Kenner Ordóñez Lacayo (996664)
Diego Solís Worsfold (B06199)

Ciudad Universitaria Rodrigo Facio, Costa Rica
2018

"Este proyecto de graduación fue aceptado por la Comisión de Trabajos Finales de Graduación de la Escuela de Formación Docente de la Universidad de Costa Rica, como parte de los requisitos para aspirar al título y grado de Licenciatura en Enseñanza de la Matemática".

Dra. Annia Espeleta Sibaja

Representante de la Escuela de Formación Docente

Dr. Guaner Rojas Rojas

Director

Dra. Susan Francis Salazar

Lectora externa

M. Sc. Karol Jiménez Alfaro

Asesora

M. Sc. Luis Rojas Torres

Asesor

Kenner Ordóñez Lacayo

Candidato

Diego Solís Worsfold

Candidato

Dedicatoria de Kenner

A mis padres, colegas del PPPAA, docentes, compañeros y compañeras de carrera.

Dedicatoria de Diego

A mi familia, amigos y compañeros de generación.

Reconocimientos

Se le agradece al Programa Permanente de la Prueba de Aptitud Académica por permitir realizar este TFG con la Prueba de Aptitud Académica de la Universidad de Costa Rica, considerando que este TFG está adscrito en la carrera de Enseñanza de la Matemática. Especialmente, por autorizar el uso de las instalaciones del edificio y parte del equipo informático.

Como parte fundamental del TFG, se le agradece al Equipo Técnico de Investigación por autorizar el uso de los datos de la Fórmula 1 de la PAA 2016, en especial al M. Ed. Danny Cerdas, por su colaboración en la elaboración de la base con las variables requeridas.

Especial reconocimiento al Dr. Guaner Rojas (director del TFG), al M. Sc. Luis Rojas y a la M. Sc. Karol Jiménez, por sus aportes académicos tanto antes del TFG como en el proceso de elaboración del mismo.

Índice general

Índice de figuras

Índice de tablas

Índice de abreviaturas

Abrev.	Significado	Notas
2PL	Dos parámetros libres	TRI
AERA	Asociación Americana de Investigación Educativa	Siglas del inglés
AFE	Análisis Factorial Exploratorio	
APA	Asociación Americana de Psicología	Siglas del inglés
DGSC	Dirección General de Servicio Civil	República de Costa Rica
ETI	Equipo Técnico de Investigación	Del PPPAA, UCR
ETS	*Educational Testing Service*	Estados Unidos de América
FDI	Funcionamiento Diferencial del Ítem	
ITCR	Instituto Tecnológico de Costa Rica	República de Costa Rica
MEP	Ministerio de Educación Pública	República de Costa Rica
MH	Modelo Mantel-Haenszel	
MH	Estadístico Mantel-Haenszel	En cursiva
NCME	Consejo Nacional de Medición en Educación	Siglas del inglés
OECD	Organización para la cooperación económica y el desarrollo	Siglas del inglés
PAA	Prueba de Aptitud Académica	Universidad de Costa Rica

Continúa en la próxima página

Continúa de la página anterior

Abrev.	Significado	Notas
PISA	Programa para la Evaluación Internacional de Estudiantes	Siglas del inglés
PPPAA	Programa Permanente de la PAA	Instituto de Investigaciones Psicológicas, UCR
SAT	*Scholastic Aptitude Test*	Estados Unidos de América
TCT	Teoría Clásica de los Tests	
Timss	Estudio Internacional de Tendencias en Matemáticas y Ciencias	Siglas del inglés
TRI	Teoría de Respuestas al Ítem	
UCR	Universidad de Costa Rica	República de Costa Rica
UNA	Universidad Nacional Autónoma de Costa Rica	República de Costa Rica

Resumen

El presente TFG tiene el objetivo de describir y analizar 5 métodos de detección del funcionamiento diferencial del ítem en pruebas estandarizadas. En particular, el estudio se centra en la PAA de la UCR aplicada en el año 2016. El problema de investigación nace de la necesidad de identificar el método de detección más cercano al contexto de aplicación de la PAA. En ese sentido, la estructura del estudio está compuesta en tres fases: 1) la descripción teórica de 5 métodos de detección, 2) estudio de simulación de los 5 métodos y 3) el estudio del funcionamiento diferencial del item en la PAA 2016. Por lo tanto, la investigación se considera un estudio explicativo que se fundamenta en el enfoque cuantitativo.

Los resultados mostraron un desempeño favorable de los métodos de detección que tienen la capacidad de identificar funcionamiento diferencial uniforme y no uniforme. En particular, la Prueba Chi-Cuadrado de Lord y Regresión Logística presentaron la menor proporción de error tipo I. Por lo tanto, se recomienda ampliar el estudio sobre el funcionamiento diferencial del ítem en investigaciones que aborden los métodos de detección de forma individual y realizar ese abordaje como la etapa inicial frente a la posible existencia de sesgo en la prueba.

Adicionalmente, se sugiere complementar la formación de docentes de matemáticas en la descripción del FDI y la creación de ítems para pruebas estandarizadas.

Palabras clave: Formación de docentes, FDI, TRI, TCT, Equidad, Validez.

Capítulo 1

Introducción

1.1. Justificación

La complejidad del mundo globalizado y la diversidad de contextos educativos imponen grandes dificultades para determinar el nivel adecuado que debe alcanzar un estudiante en su formación. Para que sea posible tomar una decisión es necesario que distintos entes participen en el diseño de un currículo mínimo que se debe cumplir en cada uno de los niveles académicos.

En particular, el MEP implementó una nueva propuesta curricular en matemáticas a partir del año 2012. Su propósito era ofrecer a los estudiantes costarricenses mejores instrumentos formativos para incrementar sus condiciones de vida. En ese sentido, se buscaba dar a todos los sectores sociales y culturales un programa de matemáticas moderno y sólido que promoviera la equidad.

Este proceso de mejora curricular conlleva, paralelamente, la actualización de los planes de formación docente. El impacto de dicha transformación se traduciría

en mejores prácticas evaluativas, nuevas líneas de investigación académica y la sistematización de nuevos estándares de educación.

Ante ese nuevo escenario de formación, debe existir algún medio que permita verificar si los estudiantes lograron superar los objetivos planteados en la nueva propuesta currícular. Una herramienta que permite realizar esa verificación son las pruebas estandarizadas. Aunque algunas personas las critican (Barrenechea, 2010; Bautista Sánchez, 2015; Martínez Rizo, 2009), estas se han utilizado por la rigurosidad con que se pueden administrar, así como la posibilidad de hacer estudios que permitan analizar su confiabilidad y validez. Al respecto, Alfaro-Rojas y Rojas Torres (2016, p. 217) afirman que la estandarización es muy importante para poder realizar inferencias para cada examinado respecto al constructo.

En esta investigación se entenderá constructo o variable latente en el sentido de la AERA, APA y NCME, es decir, una fuente de varianza que por ser oculta, generalmente, se puede medir según las puntuaciones obtenidas en distintos tests diseñados para tal fin (AERA, APA y NCME, 2014, p. 11). Esta definición de constructo brinda la importancia que tienen los tests y, eventualmente, la necesidad de que estos sean estandarizados. Al respecto, Borsboom, Mellenbergh y Heerden afirman que uno de los indicadores de una variable latente (constructo) es la variabilidad de los puntajes obtenidos por los examinados (Borsboom, Mellenbergh y van Heerden, 2004, p. 1069).

Uno de los constructos medidos en pruebas estandarizadas es el nivel que han alcanzado los estudiantes en los contenidos y procedimientos de las asignaturas o del currículo general, lo cual permite realizar comparaciones a nivel nacional e internacional. Ejemplos de estas son las conocidas pruebas Timss (Martin, Mullis y Hooper, 2016) y PISA (OECD, 2016). A saber, el Estado de la Educación del Programa Estado de la Nación afirma que las PISA permiten comparar el estado

de la educación nacional respecto a los de otras naciones, en particular, para competencias de lectura, matemática y ciencias (Estado de la Educación, 2015, p. 254). Las pruebas estandarizadas mencionadas anteriormente son diseñadas y validadas por entes internacionales y sus resultados permiten tener un indicador de la calidad del sistema educativo en el que están inmersos los estudiantes de cada uno de los países participantes.

En el escenario costarricense, además de las pruebas mencionadas anteriormente, existen otras que se diseñan y administran con los mismos estándares que exige cualquier prueba estandarizada. Algunas de ellas son administradas por el Ministerio de Educación Pública (MEP), la Dirección General de Servicio Civil (DGSC), la Universidad de Costa Rica (UCR), el Instituto Tecnológico de Costa Rica (ITCR), la Universidad Nacional Autónoma de Costa Rica (UNA) y en algunas investigaciones de la Fundación Omar Dengo.

Dadas las consecuencias asociadas a cada una de ellas, las pruebas deben ser sometidas a procesos de validación para garantizar previamente su confiabilidad. En ese contexto, existen diferentes teorías y modelos matemáticos que verifican si los indicadores miden satisfactoriamente lo pronosticado y si el modelo utilizado tiene una teoría que la respalde. Las principales teorías para analizar la confiabilidad de pruebas estandarizadas compuestas por ítems de elección única son la Teoría Clásica de los Test (TCT) y la Teoría de Respuesta al Ítem (TRI) con uno, dos y tres parámetros libres.

Aún con toda la rigurosidad con la que se diseñan cada una de las pruebas estandarizadas del MEP, el ITCR y la UCR, actualmente existe en el país una gran desigualdad en el acceso a la educación superior. Al respecto, Montero-Rojas et al. señalan que en Costa Rica algunos indicadores de equidad han venido descendiendo (2013, p. 105). Este es un punto que atañe a cualquier docente de

1.1. JUSTIFICACIÓN

Enseñanza de la Matemática, quien debe conocer el contexto actual para adecuar las clases.

En situaciones como las que vive Costa Rica respecto a la inequidad en el acceso a la educación de calidad, es muy probable que los resultados de las pruebas estandarizadas, centradas en medir logro o éxito académico, están influidos por desigualdad en el acceso a la educación superior de calidad. En particular, preocupan las oportunidades que tienen las poblaciones indígenas del país y los estudiantes con necesidades educativas especiales. Al respecto, Montero-Rojas et al. indican que el poco acceso a una educación de calidad que tienen estas poblaciones impiden que, tanto ellas como sus comunidades, tengan un amplio desarrollo económico y social (2013, p. 105).

Ante esta situación cada docente en matemática debería conocer la implementación de métodos o estrategias que permitan reducir los efectos que conlleven a aumentar la desigualdad en el acceso a la educación superior pública, efecto que no es exclusivo de Costa Rica.

Desde 1972 se ha implementado el estudio del funcionamiento diferencial del ítem (FDI) para contrastar dos grupos de una población, y determinar si la prueba o alguno de los ítems beneficia a un grupo en particular. Dichos estudios tenían la intención de identificar evidencias de sesgo en pruebas estandarizadas. Desde el punto de vista técnico, Hortensius (2012) indica que es indispensable definir las diferentes hipótesis con que se realizaría el análisis del FDI, ya sea TRI o no, y su uniformidad en el nivel de habilidad. Aun así, dicho estudio asume que los grupos de referencia y focales tienen una misma distribución de habilidades, situación que puede ser difícil de encontrar en la realidad.

1.2. Delimitación del problema

La Prueba de Aptitud Académica (PAA) de la Universidad de Costa Rica surge en 1960 por la necesidad de limitar el ingreso a nuevos estudiantes. En particular, la infraestructura universitaria no permitía atender la cantidad demandada de cupos. Eso se sumaba a las deficientes habilidades iniciales que los estudiantes revelaban para afrontar sus estudios universitarios (Mainieri Hidalgo, 2010). Ante este panorama, en 1960 se implementa una prueba semejante a la *Scholastic Aptitude Test* (SAT) de los Estados Unidos, construido por el *Educational Testing Service* (ETS) (Mainieri Hidalgo, 2010) como modelo de admisión en la universidades norteamericanas. Sin embargo, las semejanzas entre la PAA y SAT no se mantuvieron a lo largo de los años. Actualmente estas pruebas miden constructos diferentes y responden a marcos de referencia científicos distintos (Smith-Castro, 2014, p. 287).

La Universidad de Costa Rica es una de las instituciones nacionales más prestigiosas y ello se debe a la excelencia académica de esta, por lo que la demanda es significativa y esto a su vez implica que es necesario utilizar un sistema de ingreso que permita seleccionar a las personas con mayor probabilidad de éxito, para lo cual se optó por el uso de un instrumento con alta calidad técnica (Jiménez-Alfaro y Morales-Fernández, 2010, p. 52).

De lo anterior se confirma que la población meta de la PAA son todas aquellas personas que aspiran a realizar estudios de pregrado o de grado en la Universidad de Costa Rica. Además, es necesario que estas personas aprueben o que hayan aprobado las pruebas nacionales de bachillerato para el año previo al que desean ingresar a la UCR.

1.2. DELIMITACIÓN DEL PROBLEMA

Actualmente la PAA mide razonamiento general en contextos verbales y matemáticos mediante ítems de elección única. Los folletos de examen utilizados en el año 2016 estaban compuestos por 85 ítems de los cuales 70 ya habían sido utilizados en la población meta y los restantes, estaban en calidad de ítems piloto. De los ítems que fueron aplicados anteriormente a la población meta, en el momento de ser seleccionados para el ensamblaje se contaba con propiedades de calidad técnica establecidas por el Equipo Técnico de Investigación (ETI). Los ítems en condición de pilotaje, para efectos del cálculo de la nota de la PAA, se asignan como ítems respondidos correctamente, sin embargo, para efectos de sus indicadores de calidad técnica se utilizan los patrones de respuestas de toda la población que los resolvieron.

Montero-Rojas et al. (2013, p. 109) realizan las siguientes aclaraciones respecto al uso de la PAA:

- El promedio de admisión es uno de los elementos que se utilizan para el ingreso a carreras de pregrado y grado de la Universidad de Costa Rica.
- El promedio de admisión pretende ordenar a todos los examinados según la probabilidad de éxito en la Universidad en comparación con los demás examinados, es decir, un modelo referido a normas.
- Los examinados adquieren el derecho para optar por los distintos cupos que existen para cada una de las carreras.

Además, Jiménez-Alfaro y Morales-Fernández (2010, p. 51) como parte de sus investigaciones concluyen que los mejores predictores de éxito en la UCR son tanto el promedio de educación diversificada como el puntaje obtenido en la PAA.

Sin embargo, aun teniendo evidencias de la calidad técnica de la PAA tanto en la confiabilidad como en la validez, se debe considerar que, según Jiménez-Alfaro y Morales-Fernández (2010, p. 33), en las ciencias sociales se hace complejo estudiar distintas variables pues estas están afectadas por una gran diversidad de factores. En particular, se debe considerar que la sociedad costarricense sufre de una desigualdad considerable cuando se trata del acceso a educación de calidad (Estado de la Educación, 2015, pp. 280-286), lo cual puede afectar la validez de las inferencias que se hacen con diversas pruebas estandarizadas aplicadas en el país.

Por lo anterior, es necesario revisar la pertinencia y la precisión de los estudios de equidad que se le realizan actualmente a la PAA para así disminuir los efectos negativos que se pueden presentar. Es por ello que se hace necesario verificar cuál es el modelo teórico que mejor se ajusta a la PAA para que, en función de este modelo, se pueda determinar cuáles son los mejores métodos de detección del FDI, según el diseño y condiciones de aplicación de la PAA.

1.3. Planteamiento del problema

Por lo anterior, el análisis del FDI para cada uno de los ítems podrían disminuir la inequidad en la aplicación de pruebas estandarizadas. Como lo indica Moreira Mora (2008), estos análisis son otro indicador de qué tan bien miden el constructo cada uno de los ítems.

Por lo tanto, se considera necesario realizar un análisis del FDI que estén acordes en el contexto nacional y los modelos de medición más frecuentes. Considerando la trayectoria de la Prueba de Aptitud Académica (PAA) de la UCR se decide centrar el siguiente Trabajo Final de Graduación en:

1.3.1. Objetivo General

Examinar las técnicas para capturar el funcionamiento diferencial del ítem (FDI), según sexo, en la Prueba de Aptitud Académica (PAA) 2016.

1.3.2. Objetivos Específicos

Para cumplir con el objetivo general se pretenden realizar las actividades específicas pertinentes para lograr:

1. Proporcionar una revisión de métodos estadísticos para evaluar empíricamente FDI en ítems de pruebas estandarizadas.
2. Proponer una guía de análisis del FDI para pruebas estandarizadas con ítems de razonamiento general.
3. Identificar la proporción de ítems de la PAA 2016 con funcionamiento diferencial según sexo.

1.4. Antecedentes

Los primeros estudios sobre sesgo en pruebas estandarizadas se realizaron en 1960 cuando se asumió que las diferencias de éxito en pruebas de habilidad cognitiva entre grupos etarios (negros/latinos y blancos) se debía a reactivos con contenido fuera del contexto de las minorías. Dichos estudios sociales se centraban en la necesidad de identificar los ítems parcializados y eliminarlos de las pruebas estandarizadas. Angoff (1993) señala que los investigadores cometieron errores

metodológicos significativos en sus resultados por utilizar los resultados de las pruebas completas para generar comparaciones de habilidad entre los grupos de *estudio*, *y no* separar el estudio por ítem.

La preocupación por realizar estudios de justicia selectiva hizo que durante los años sesentas existiera una confusión entre la definición social y estadística del sesgo. Señala Angoff (1993) que el término "Funcionamiento Diferencial del Ítem (FDI)" se llegaría a consolidar como referencia conceptual en el caso en que un ítem evidencie distintas propiedades estadísticas para grupos distintos con la misma habilidad.

El primer estudio formal sobre sesgo del ítem (aún no definido FDI) fue realizado por Cardall y Coffman en 1964, cuando aplicaron análisis de varianza en una prueba SAT entre grupos de negros y blancos. Ese acercamiento inicial fue superado por Angoff en 1972 cuando estudió diferencias culturales mediante el método delta-plot. Dicho procedimiento se basaba en el cálculo de valores *p* por ítem en cada grupo comparativo y luego convertido a una normal distribución normal.

Otro gran avance para el FDI, es la publicación de la TRI por parte de Lord en 1952. Dicha publicación es especialmente importante porque su foco de atención son los ítems, no la puntuación total de la prueba. No mucho tiempo después, se utilizaría dicho modelo para caracterizar el funcionamiento diferencial del ítem. Esta propuesta teórica permite realizar comparaciones entre grupos mediante la curva característica del ítem. En particular, se pueden realizar pruebas estadísticas para verificar si la diferencia entre las curvas es significativa.

En términos generales, los métodos de detección del FDI asociados a la TRI fueron ignorados por la cantidad de datos necesarios para obtener resultados estables y la complejidad del modelo. En esas condiciones, Angoff (1993) señala que

otros métodos como el delta-plot tuvieron mejor aceptación por su simple aplicación, aunque eso conllevara a obtener resultados erróneos. En particular, el error podría justificarse por no considerar el rol de la discriminación del ítem y centrarse en la dificultad del mismo. Esta característica sería compartida por el modelo de Rasch propuesto en 1960.

En el año 1988, Holland y Thayer describirían un procedimiento para investigar el FDI que usaba una técnica desarrollada por Mantel y Haenszel en 1959. Esta técnica se basa en una prueba Chi-cuadrado sobre tablas de contingencia, por intervalo de habilidad, que describen las frecuencias de respuestas correctas e incorrectas por grupo de interés. Unos años antes, en 1986, Dorans y Kulick propondrían un método de detección del FDI muy similar a MH. Se le llamó proceso de estandarización, dado que realiza una comparación entre las diferencias de los valores *p* esperados por grupo de interés.

Capítulo 2

Marco teórico

2.1. Validación de pruebas

Según Montero Rojas (2013), los temas de mayor interés en educación y psicología incluyen la definición y medición de constructos tan complejos como: habilidad intelectual, aptitudes académicas y rasgos de personalidad. En dichos estudios un extracto de una población debe ser categorizado mediante una prueba estandarizada. Para Nunnally y Bernstein (1995) el medio utilizado es una herramienta de medición que estima calificaciones en una escala numérica para revelar el constructo. Ante la función selectiva de la prueba, la principal preocupación es demostrar que las decisiones se toman con justicia frente a los examinados. Montero Rojas (2013) insiste en la necesidad de abordar estos temas desde la psicometría y de esa forma consolidar indicadores del grado de validez.

Según Zumbo (1999, p. 10), cuando se utiliza la palabra "validez" no se infiere directamente la validez de las puntuaciones obtenidas por el examinado, sino a las

inferencias que se realizan a partir de esas puntuaciones. Dicha interpretación es una nueva postura académica por asumir las consecuencias que conlleva la aplicación de una prueba. En ese nuevo sentido académico, la validez incluye:

1. Evidencia de constructo (multidimensionalidad).
2. La precisión de las mediciones (confiabilidad).
3. La validez como un continuo.
4. La validez como teoría, no como conjunto de herramientas.
5. La necesidad de estudiar estadísticamente el sesgo.

Desde este nuevo enfoque, la validez tiene que formar parte de todo el proceso de creación de una prueba: la definición del constructo, la creación de los ítems y de la interpretación de los resultados (Messick, 1989, p. 5). En ese sentido, Montero Rojas (2013) caracteriza la calidad técnica de una prueba como un proceso riguroso y sistemático de acumulación de evidencia empírica para consolidar las inferencias generadas a partir de las mediciones de la prueba.

En el contexto de validación es importante dejar claro el alcance de ciertos conceptos que se utilizan con regularidad. Su definición permite un acercamiento teórico claro y preciso. Para dicha aproximación se proponen las definiciones proporcionadas por Mellenbergh (1982) y Zumbo (1999, p. 12) (propuesta original Zumbo y Hubley (1998), Camilli y Shepard (1994) y Clauser y Mazor (1998)):

Análisis de ítems: Grupo de técnicas estadísticas para estudiar el rendimiento de ítems de forma individual. Es un estudio relevante cuando se busca consolidar un examen o para adoptar una medida específica.

Impacto de un ítem: Es evidente cuando examinados de diferentes grupos tienen probabilidades distintas de responder correctamente un ítem debido a que existen verdaderas diferencias entre los grupos a través de la habilidad que se pretende medir con el ítem.

FDI: FDI ocurre cuando examinados de diferentes grupos revelan diferentes probabilidades de éxito frente al ítem a pesar de compartir la misma habilidad que el ítem pretende medir.

FDI Uniforme: La magnitud de dependencia condicional por grupo de estudio es relativamente invariante en el continuo del rasgo latente.

FDI No Uniforme: La magnitud de dependencia condicional por grupo de estudio varia y cambia de dirección en el continuo del rasgo latente.

Sesgo del ítem: Ocurre cuando examinados de un grupo tienen menor probabilidad de éxito frente al ítem debido a una característica propia del ítem o una situación evaluativa que no forma parte del propósito de la prueba. El FDI es una condición necesaria pero no suficiente para identificar sesgo en un ítem.

Impacto Adverso: Es el término legal que describe la situación en donde diferencias grupales en el rendimiento de una prueba resulta en una selección desproporcionada de examinados o decisiones parecidas. Este factor no es evidencia de sesgo de la prueba.

Frente al escenario de un estudio sobre sesgo, se debe considerar el papel que juega un estudio del FDI. Lo más relevante de dicho estudio es que la no confirmación del FDI certifica la imparcialidad de un ítem, pero la verificación del FDI es tan solo la primera condición necesaria, para la existencia de sesgo. En ese escenario, Zumbo (1999) recomienda hacer un seguimiento estricto utilizando técnicas de

reconocimiento de sesgo, por ejemplo: análisis de contenido o evaluaciones estadísticas. La primera se puede realizar mediante un panel de expertos que emiten un criterio de verificación, pero se recomienda utilizar las segundas por su rigurosidad técnica.

A continuación, se exponen algunas condiciones que podrían generar sesgo durante el proceso de diseño de ítems de pruebas educativas:

- Contextos narrativos históricamente asociados a algún sexo. Por ejemplo, la pesca y la casería.
- Vocabulario variado según la lengua nativa. Por ejemplo, el término "asistencia" se puede interpretar como que se preste ayuda o como se presentarse en algún lugar.
- Vocabulario variado según la zona geográfica. Por ejemplo, el término "maquinilla" se puede interpretar como el diminutivo de "máquina" o como "tajador"
- La selección de los distractores. Esto se refiere al hecho de que el contenido sea más atractivo según el sexo.

Si se elige el FDI como herramienta primaria de estudio, se deben considerar las siguientes recomendaciones realizadas por Zumbo (1999, p. 14):

1. Existen gran variedad de subgrupos que se pueden contrastar. Se debe clarificar cuál es el foco de atención. Las comparaciones estándar se basan en género, raza, subcultura o lengua.
2. Se debe discutir el grado del valor del estadístico que se pretende tolerar, antes de clasificar un ítem como FDI.

3. El análisis de datos debe realizarse tomando en cuenta los casos en que se favorece al grupo de referencia y al grupo focal.

4. Se debe considerar la etapa constructiva en que se aplica el estudio del FDI: prueba existente/prueba nueva y prueba piloto/prueba oficial. Los análisis del FDI se deben realizar antes de reportar los resultados.

5. La identificación del FDI no es la etapa final de un estudio. Se deben analizar sus causas a partir de estudios de contenido con expertos y pruebas de validación.

Toda investigación que pretenda realizar una descripción del FDI, tiene que reconocer que una parte fundamental del estudio se centra en la teoría moderna de los test, donde el centro de atención es la variación del continuo de la variable latente. Dichas teorías sobre los test permiten definir e interpretar los resultados de un estudio diferencial del ítem. Los principales representantes de dichas teorías son TCT y TRI.

2.2. Modelos de medición

Todas las pruebas estandarizadas deben tener una alta calidad técnica que asegure un buen diseño de los ítems y una correcta interpretación de los resultados. Esta calidad debe estar garantizada por procedimientos y teorías científicas que permitan a los tomadores y administradores de pruebas estar conformes con cada una de sus obligaciones. Para esto existen dos teorías principales que permiten estudiar la calidad interna de cada instrumento, las cuales, al combinarlas con otros enfoques y metodologías, proporcionan indicadores de qué tan adecuado es el uso

que se le da a cada instrumento de medición. Estas teorías se conocen como Teoría clásica de los Tests (TCT) y Teoría de Respuesta a los Ítems (TRI).

Los primeros trabajos de la TCT corresponden a Spearman publicados en los primeros años del siglo XX. Más tarde, con los diversos avances teóricos y tecnológicos, Lord y Novick proporcionan los primeros elementos de la TRI (Muñiz, 2010, p. 59).

Muñiz (2010, pp. 60-61) indica que, según Spearman, para aplicar la TCT se deben cumplir los siguientes supuestos:

1. La "Verdadera Puntuación" de un examinado es la "Esperanza matemática de la Puntuación Empírica"
2. La "Verdadera Puntuación" no debe estar relacionada con el error de medición.
3. Los errores de medida de una medición no deben estar relacionados con otras mediciones.

Un modelo que puede describir la TCT es

$$X = V + \varepsilon \tag{2.1}$$

donde

- X es la puntuación esperada del examinado.
- V es la puntuación verdadera del examinado.
- ε es el error de medición.

Según Muñiz (2010, p. 62), algunas de las críticas de la TCT es que los puntajes obtenidos en instrumentos que miden el mismo constructo, no siempre se podrán comparar pues estos no tienen la misma métrica. Otra de las críticas es que el análisis de los instrumentos depende en gran medida de la población que realiza cada uno de los tests, lo cual produce problemas al realizar o utilizar baterías de ítems o fórmulas. Y una de las últimas críticas es que el uso de coeficientes como el Alfa de Cronbach es el mismo para todos los examinados, sin importar su nivel en el constructo.

La TRI es en términos matemáticos y estadísticos más compleja que la TCT, pero permite abordar o corregir las críticas antes mencionadas que se le atribuían a la TCT. En el caso de la TRI se debe satisfacer que exista una relación entre lo que mide el ítem y la probabilidad de que los examinados acierten estos (Muñiz, 2010, p. 64). Para esta teoría se pueden considerar distintos modelos que están en función de tres parámetros: la discriminación, la dificultad y el índice de acertar por azar.

La ecuación matemática de la TRI con dos parámetros es

$$P(\theta) = \frac{1}{1 + e^{Da(\theta - b)}} \tag{2.2}$$

donde:

- θ es el nivel de habilidad del examinado.
- $P(\theta)$ es la probabilidad de que un examinado de habilidad θ responda correctamente el ítem.
- a se interpreta como la discriminación del ítem.
- b se interpreta como la dificultad del ítem.

- D es un valor de normalización, generalmente, 1.17.

Según Muñiz (2010), en los distintos modelos TRI se debe asumir que los ítems solamente miden un constructo, es decir, que son unidimensionales. El tercer supuesto es la independencia local, es decir, que las respuestas de los ítems deben ser independientes de las de los demás. Esto significa que para los tests que se fundamenten o utilicen la TRI, se deben verificar si se satisfacen los tres supuestos antes mencionados.

Como aspectos generales de ambas teorías se puede mencionar que en la TCT se le da énfasis al test como un todo y que se recomienda su uso en muestras entre los 200 y 500 examinados, por su lado, en la TRI se le da énfasis a cada uno de los ítems que componen el test y su uso se recomienda en muestras de más de 500 examinados, aunque depende del modelo seleccionado.

2.3. Caracterización teórica de los método de FDI

En la tabla 2.1, Magis, Béland, Tuerlinckx y Boeck (2010, p. 849) presenta algunos de los métodos de detección de FDI más tradicionales hasta ese momento. En esta tabla, se presenta cada uno de los métodos distribuidos por modelo de medición y si están en capacidad de detectar FDI uniforme o no. De estos diez métodos, siete se implementan en el paquete difR, Magis, Beland y Raiche (2015).

En función de lo anterior, se realiza un estudio detallado de las cinco técnicas sobre el FDI elegidas para la investigación: MH, Estandarización, Regresión Logística, Lord y Raju. Dicho estudio considera los aspectos estructurales de cada modelo y las condiciones en que mejor se desempeñan. La caracterización teórica

Tabla 2.1

Métodos tradicionales para detectar Funcionamiento Diferencial del Ítem (FDI)

		Número de grupos	
Modelo	Efecto FDI	2	> 2
No-TRI	Uniforme	Mantel–Haenszel* Estandarización* SIBTEST Regresión Logística*	Comparaciones por pares Mantel–Haenszel Generalizado*
No-TRI	No uniforme	Regresión Logística* Breslow–Day* NU.MH NU.SIBTEST	Comparaciones por pares
TRI	Uniforme	LRT* Lord* Raju*	Breslow–Day Lord Generalizado*
TRI	No uniforme	LRT* Lord* Raju*	Breslow–Day Lord Generalizado*

Notas:
NU.MH: modificación de Mantel–Haenszel para FDI no uniforme.
NU.SIBTEST: modificación de SIBTEST para FDI no uniforme.
LRT: Prueba de Razón de Verosimilitud.
* Actualmente implementado en el paquete difR (versión 2.2).
Tomado de Magis et al. (2010, p. 849).

identifica posibles limitaciones de cada modelo y señala la naturaleza descriptiva de los datos.

2.3.1. Métodos en la TCT

Los instrumentos estadísticos utilizados para identificar el FDI son definidos por Shepard (1982, p. 23) como métodos internos diseñados para asegurar que las

interpretaciones, que los ítems individuales atribuyen al puntaje total de la prueba, es el mismo para todos los subgrupos. Dicha definición está contextualizada dentro de la Teoría Clásica de los Test. En ese sentido, los tres modelos estadísticos más representativos son Mantel-Haenszel (Holland y Thayer, 1988), Estandarización (Dorans y Kulick, 1986) y la Regresión Logística.

Según Lord (1977), un estudio que se basa en la proporción de respuestas correctas puede arrastrar anomalías. Su principal inquietud es la dependencia de los resultados al grupo respectivo.

Para Dorans y Holland (1993) el puntaje total de una prueba es una medida mucho más confiable que la medida propuesta por ítems individuales, siempre que se demuestre la validez de la prueba y se pueda asegurar que el puntaje de la prueba se obtiene bajo las mismas condiciones para el grupo focal y el grupo de referencia.

Mantel-Haenszel

En el año 1959, Mantel y Haenszel introducen un nuevo procedimiento para el estudio de grupos emparejados. Según Dorans y Holland (1993), serían Holland y Thayer, en 1988, quienes adaptarían el procedimiento general para el estudio diferencial del ítem.

En su etapa inicial, los datos básicos para el estudio MH deben colocarse en M tablas de contingencia 2×2 o una tabla tridimensional $2 \times 2 \times M$ como se muestra en la tabla 2.3.

El método MH asume como hipótesis nula que la probabilidad de obtener un ítem correcto en un nivel de habilidad específico según la variable latente es la misma

Tabla 2.3

Organización de datos en tabla de contingencia 2 × 2

	Puntuación del Ítem		
Grupo	Correcto	Incorrecto	Total
Focal (*f*)	C_{fm}	I_{fm}	N_{fm}
Referencia (*r*)	C_{rm}	I_{rm}	N_{rm}
Total Grupos	C_{tm}	I_{tm}	N_{tm}

Nota: Tomado de Dorans y Holland (1993, p. 39).

tanto para el grupo focal como para el grupo de referencia en todos los *M* niveles estudiados.

$$H_0 : \frac{C_{rm}}{I_{rm}} = \frac{C_{fm}}{I_{fm}}, \; m \in 1, \ldots, M \tag{2.3}$$

Dorans y Holland (1993) advierten que en la propuesta original, Mantel y Haenszel desarrollaron una prueba de chi cuadrado de la hipótesis nula contra una alternativa particular conocida como hipótesis de razón de probabilidad constante, definida como:

$$H_a : \frac{C_{rm}}{I_{rm}} = \alpha_m \cdot \frac{C_{fm}}{I_{fm}}; \; \alpha_m = 1, \; m \in 1, \ldots, M \tag{2.4}$$

El estadístico asociado al método MH que verifica la validez de la hipótesis nula señalada en la ecuación (2.1) se define como *MH*-χ^2. Dicho estadístico se distribuye aproximadamente como chi cuadrado con un grado de libertad.

$$MH - \chi^2 = \frac{\left[\left|\sum_{m=1}^{M} C_{rm} - \sum_{m=1}^{M} \mathrm{E}(C_{rm})\right| - 0.5\right]^2}{\sum_{m=1}^{M} \mathrm{Var}(C_{rm})} \tag{2.5}$$

donde

$$E(C_{rm}) = E(C_{rm} \mid \alpha_m = 1) = \frac{N_{rm}C_{tm}}{N_{tm}}$$

$$Var(C_{rm}) = E(Var_{rm} \mid \alpha_m = 1) = \frac{N_{rm}C_{tm}N_{fm}I_{tm}}{N_{tm}^2(N_{tm}-1)}$$

De forma adicional, el método MH estima la razón de probabilidades constantes a través de todas las categorías de habilidad asociadas. El índice se define como

$$\alpha_{MH} = \frac{\sum_{m=1}^{M} C_{rm}I_{fm}/N_{tm}}{\sum_{m=1}^{M} C_{fm}I_{rm}/N_{tm}}, \tag{2.6}$$

el cual representa, según Angoff (1993), el factor promedio en el cual se compara la probabilidad de que un individuo del grupo de referencia obtenga la respuesta correcta frente a la probabilidad de un individuo del grupo focal (medida del efecto del tamaño).

Dicho factor promedio señalado en la ecuación (2.6) se puede transformar según la propuesta de Holland y Thayer (1985). Esa transformación facilita la interpretación del factor dado que es simétrico alrededor de cero,

$$MHD - DIF = -2.35 \cdot \log(\alpha_{MH}) \tag{2.7}$$

donde valores positivos de *MHD – DIF* favorecen al grupo focal y valores negativos al grupo de referencia.

La clasificación formal de ítems según *MHD – DIF* para identificar los distintos grados del FDI fue creado por la ETS (Dorans y Holland, 1993, p. 10). La designación toma en consideración un valor de significancia del 5%. La clasificación es FDI despreciable, FDI intermedio y FDI amplio.

Tabla 2.4

Designación del FDI para MH según a ETS

Clasificación			*MHD – DIF*		
Despreciable (A)			$\lvert MHD - DIF \rvert$	$\leq$	1
Intermedio (B)	1	$<$	$\lvert MHD - DIF \rvert$	$<$	1,5
Amplio (C)	1,5	$\leq$	$\lvert MHD - DIF \rvert$		

Angoff (1993) hace una aclaración final sobre el uso de pruebas significativas ($MH - \chi^2$) y la medida del efecto del tamaño (*MHD – DIF*). Señala que las pruebas estadísticas serán significativas si la muestra es lo suficientemente grande y la medida del efecto de tamaño permitirá conclusiones asociadas a la significancia cuando la medida del efecto sea pequeña.

Estandarización

En 1982, Dorans concluye que un nuevo método para estudiar el FDI debe utilizarse dado que la metodología delta-plot y la log-linear tenían grandes problemas para ajustarse al modelo TRI de un parámetro.

Según Dorans y Holland (1993), serían Dorans y Kulick, en 1983, quienes utilizarían un nuevo método que compararía las curvas de respuesta del ítem empíricas, donde el puntaje total sería el estimado de habilidad.

Dorans y Holland (1993) señalan que el proceso de estandarización (Est) permite identificar el FDI cuando el funcionamiento esperado de un ítem difiere entre estudiantes con la misma habilidad pero pertenecientes a distintos grupos. Dicho funcionamiento esperado se puede operacionalizar por medio de regresiones item-test no paramétricas. En particular, diferencias en la regresión item-test empírica es indicativa de FDI.

2.3. CARACTERIZACIÓN TEÓRICA DE LOS MÉTODO DE FDI

Uno de los principios fundamentales del proceso de estandarización es utilizar todos los datos accesibles de la muestra para calcular el comportamiento del ítem en cada estrato de habilidad de la variable latente por grupo. De hecho, Dorans y Holland (1993) advierten que menospreciar datos solo conlleva a estimaciones pobres sobre el efecto del tamaño de la muestra, que tiene mayores errores estandarizados asociados que los estimados según la muestra completa.

La definición del FDI según el proceso de estandarización asume como hipótesis nula

$$H_0 : E_f(I \mid M) = E_r(I \mid M) \tag{2.8}$$

o de forma análoga por diferencias en el nivel m de habilidad

$$D_m = E_{fm} - E_{rm} \tag{2.9}$$

donde $E_f(I \mid M)$ es la regresión item-test empírica para el grupo focal según la puntuación I del ítem para el nivel de habilidad M y $E_r(I \mid M)$ es la regresión item-test empírica para el grupo de referencia según la puntuación I del ítem y para el nivel de habilidad M.

Las diferencias D_m son las medidas fundamentales según el proceso de estandarización debido a que esas cantidades son diferencias p en el rendimiento del ítem por grupos distintos cuyas habilidades son comparadas en la variable latente medida por el test. Dorans y Holland (1993) aclaran que cualquier diferencia significativa que exista después de dicha comparación no puede ser explicada por diferencias en la habilidad. Las posibles diferencias encontradas son inesperadas frente a posibles diferencias existentes en la habilidad por grupo. Gráficas de las diferencias, así como de las regresiones $E_f(I \mid M)$ y $E_r(I \mid M)$, proveen evidencias visuales del comportamiento diferencial del ítem.

La diferencia *STD – PDIF* se pesa en términos de un grupo estandarizado (i.e.: grupo focal),

$$STD - PDIF = \frac{\sum_{m=1}^{M} w_m(E_{fm} - E_{rm})}{\sum_{m=1}^{M} w_m} \tag{2.10}$$

donde $\frac{w_m}{\sum_{m=1}^{M} w_m}$ es el factor de peso en el nivel *m* revelado por el grupo de estandarización para comparar los pesos del funcionamiento de cada ítem entre el grupo focal (E_{fm}) y el grupo de referencia (E_{rm}).

El índice *STD – PDIF* revela valores entre –1 y 1. Los valores positivos indican que el ítem favorece al grupo focal, mientras que valores negativos no favorecen al grupo focal.

Tabla 2.5

Designación del FDI para Estandarización

Clasificación			*STD – PDIF*		
Despreciable (A)			$\lvert STD - PDIF \rvert$	$\leq$	0.05
Intermedio (B)	0.05	$<$	$\lvert STD - PDIF \rvert$	$<$	0.10
Amplio (C)	0.10	$\leq$	$\lvert STD - PDIF \rvert$		

Regresión Logística

Según indican Magis et al. (2010, p. 851), Hortensius (2012, p. 3), Zumbo (1999, p. 26) y Magis et al. (2015, p. 40), Swaminathan y Rogers, en 1990, propusieron el método de regresión logística para detectar FDI (Log), la cual fue otra alternativa a Mantel-Haenszel. En esta regresión, teóricamente la variable dependiente es la probabilidad que posee cada examinado para acertar la respuesta de cada ítem, con datos observados serían las respuestas que dan los examinados a cada uno de los

ítems. Las variables explicativas son las puntuaciones obtenidas por los examinados en el test, si el examinado pertenece al grupo focal o no y la interacción entre las respuestas dadas por los examinados según su pertenencia o no al grupo focal.

De lo anterior se desprende que la versión general del modelos es la siguiente:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 S_i + \beta_2 G_i + \beta_3 (SG)_i \tag{2.11}$$

donde:

- π_i es la probabilidad que tiene el examinado *i* de acertar el ítem.
- S_i es el puntaje total obtenido por el examinado.
- G_i es la indicatriz de pertenencia al grupo focal *G*.
- (SG_i) son los puntajes obtenidos por el examinado *i*, según pertenezcan o no al grupo focal *G*.
- β_0, β_1, β_2 y β_3 son los parámetros que se deben determinar para la regresión.

En los parámetros β_0 y β_1 indicarían si los datos ajustan adecuadamente en esta regresión, la cual se sabe que corresponde a la TCT por considerar los puntajes totales de los examinados en el test. En el caso de que β_2 sea significativo sería un indicador de presencia de FDI. Hasta este punto, según Hortensius (2012, p. 5) este método es semejante respecto a la eficacia al de Mantel-Haenszel, salvo el hecho de que el método de regresión logística proporciona mas falsos positivos.

Sin embargo, una vez que se sabe que un ítem posee FDI, este método proporciona mayores ventajas respecto al de Mantel-Haenszel. En el caso del método de regresión logística permite detectar FDI no uniforme, el cual se presenta

si el parámetro β_3 es significativo. En el caso de que β_3 no sea significativo, entonces el FDI es uniforme.

Para la estimación de los parámetros en este tipo de regresiones se suele utilizar el método de máxima verosimilitud y para el análisis del ajuste, según Magis et al. (2010, p. 851) y Magis et al. (2015, p. 41) se pueden utilizar las ΔR^2 de Cox y Snell o la de Nagelkerke. Algunos de los cortes que se pueden utilizar son los de Zumbo y Thomas, principalmente para test de diagnóstico, y el de Jodoin y Gierl, para test de altas consecuencias. A continuación se proporcionan cada no de estos cortes:

- Cortes según Zumbo y Thomas (laxo):
  - Si $\Delta R^2 \leq 0.13$, entonces es Insignificante.
  - Si $0.13 < \Delta R^2 \leq 0.26$, entonces es Moderado.
  - Si $0.26 < \Delta R^2$, entonces es Alto.
- Cortes según Jodoin y Gierl (estricto):
  - Si $\Delta R^2 \leq 0.035$, entonces es Insignificante.
  - Si $0.035 < \Delta R^2 \leq 0.07$, entonces es Moderado.
  - Si $0.07 < \Delta R^2$, entonces es Alto.

2.3.2. Métodos en la TRI

Los dos métodos estadísticos que se ajustan al modelo de medición TRI, escogidos para el estudio son: la prueba de Chi-cuadrado de Lord y la prueba de Áreas de Raju.

Prueba Chi Cuadrado de Lord

La prueba chi-cuadrado de Lord fue propuesta por Federik Lord en 1980. Magis et al. (2010) señalan que el método asume que los grupos de referencia y focal tienen los mismos parámetros según la TRI, con un estadístico distribuido según chi-cuadrado bajo la hipótesis nula. Según Magis et al. (2010), este procedimiento tiene la ventaja de ajustarse a cualquier modelo TRI, siempre que los parámetros se ajusten a una métrica común antes del estudio.

El estadístico asociado al ítem *i* tiene la siguiente forma:

$$Q_i = (V_{iR} - V_{iF})' \left(\sum_i R - \sum_i F \right)^{-1} (V_{iR} - V_{iF}) \qquad (2.12)$$

donde $V_{iR} = (a_{iR}. b_{iR}. c_{iR})$ y $V_{iF} = (a_{iF}. b_{iF}. c_{iF})$ son los vectores de discriminación, dificultad y parámetro de azar para el ítem *i* según el grupo de referencia y grupo focal, respectivamente, y $\sum_i R$ y $\sum_i F$ son las matrices de varianza-covarianza.

El estadístico Q_i tiene una distribución chi-cuadrada asintótica, que depende de la distribución normal asintótica de los estimados máximos (verosimilitud) de los parámetros. Magis et al. (2010) indican que los grados de libertad dependen de los parámetros estimados por modelo.

En particular, Mclaughlin y Drasgow (1987) señalan que los estimados de los parámetros son particularmente problemáticos cuando su cálculo se realiza simultáneamente con los parámetos de las personas, debido a que no obtendrían las propiedades usuales de los estimados máximos. Por lo tando, la prueba estadística no seguiría una distribución chi-cuadrada que permita validar la prueba del FDI.

Sin embargo, Mclaughlin y Drasgow (1987) advierten que los *cálculos* simultáneos de parámetros (ítem y personas) sigue siendo una práctica común, particularmente para pruebas con una cantidad de ítems moderada o grande.

Áreas de Raju

Este método fue propuesto en Raju (1988) y Raju (1990) y consiste en comparar el área bajo entre las curvas características de los ítems correspondientes al grupo focal y el grupo de referencia. Se debe recordar que para utilizar modelos TRI se debe contar con muestras de al menos 500 examinados (Raju, 1990, p. 202). Este método no es tan apropiado utilizarlo en modelo 3PL, es más útil en los modelos 1PL y 2PL (Raju, 1990, p. 206).

Para aplicar el método de Áreas de Raju, primeramente las CCI, del grupo focal y del referencia, se deben equiparar para que estén en la misma métrica (Kim y Cohen, 1995, p. 292). Una vez que están en la misma métrica, se espera que en este método, la diferencia entre las CCI, sea cero. Para esto se utiliza el estadístico Z [1] (Magis et al., 2010, p. 852).

En modelos 1PL se da directamente con las diferencias de las habilidades estimadas entre ambos grupos, es decir,

$$Z = \frac{b_{jR} - b_{jF}}{\sqrt{\sigma^2_{jR} + \sigma^2_{jF}}}, \tag{2.13}$$

donde b_{ji} y σ^2_{ji} son la habilidad estimada y el error estándar, respectivamente, para el ítem j y grupo i, donde i puede ser R o F, (Magis et al., 2010, p. 852). En el paquete difR, Magis et al. (2015, p. 67), cuando se aplica el modelo 1PL en realidad se ejecuta

[1] El estadístico Z se basa en la distribución normal.

el modelo de Rasch, pues por omisión se asume que el parámetro de discriminación es 1.

El método Raju para la detección de FDI en los modelos TRI-3PL indican que si los parámetros de "azar" son distintos, entonces necesariamente el ítem posee FDI (Raju, 1988, p. 499). Esto significa que los parámetros de "azar" en ambas poblaciones debe ser el mismo.

Al tratarse de modelos de la TRI se debe considerar que los parámetros de discriminación deben ser positivos, pues, de no ser así, el ítem está midiendo lo opuesto al constructo deseado. Entonces, si para las dos poblaciones, los parámetros de discriminación son distintos, es claro que el ítem presenta FDI pues para una población está midiendo el constructo y para la otra está midiéndolo de forma opuesta. Por lo tanto, para ambas poblaciones, los parámetros de discriminación deben ser positivos.

Considerando a *SA* como las diferencias de las CCI de dos grupos (referencia y focal), si las *SA* se distribuyen normalmente, centrada en 0 y normalizada, entonces se utiliza

$$Z = \frac{SA - 0}{\sigma(SA)} \tag{2.14}$$

con $SA = (1 - c)\left(\hat{b}_2 - \hat{b}_1\right)$ y $\hat{a}_1 = \hat{a}_2$. En un caso como este, de existir FDI, este sería uniforme.

Si las *SA* no se distribuyen normalmente, entonces se utiliza

$$Z = \frac{H - 0}{\sigma(H)} \tag{2.15}$$

con $H = (1 - c)\frac{2(\hat{a}_2 - \hat{a}_1)}{D\hat{a}_1\hat{a}_2}\ln\left(1 + \exp\left(\frac{D\hat{a}_1\hat{a}_2\left(\hat{b}_2 - \hat{b}_1\right)}{\hat{a}_2 - \hat{a}_1}\right)\right) - \left(\hat{b}_2 - \hat{b}_1\right)$ y $\hat{a}_1 \neq \hat{a}_2$. En un caso como este, de existir FDI, este sería no uniforme.

En estos resultados los subíndices 1 y 2 se pueden sustituir por *R* (referencia) o *F* (focal), dependiendo a cuál grupo se le quiere dar prioridad en los cálculos.

Para los cortes, el paquete difR, Magis et al. (2015, p. 68), utiliza como referencia a los cuantiles de la distribución normal estándar con colas inferiores $1 - \frac{\alpha}{2}$

2.4. Potencia Estadística

Según Ellis (2010) la potencia de cualquier prueba estadísticamente significativa se define como la probabilidad de observar en la muestra una determinada diferencia que sí existe en la población. Es decir, la probabilidad de no cometer un error tipo II. En otras palabras, potencia estadística se define como la probabilidad de rechazar la hipótesis nula cuando la hipótesis alternativa es verdadera.

De forma adicional, Ellis (2010) señala que la potencia estadística afecta directamente la calidad de las inferencias resultantes de las muestras escogidas. En particular, si la potencia promedio es baja, entonces la proporción de conclusiones incorrectas sería alta.

En este TFG, la hipótesis nula de la "Potencia estadística" es: El método de detección de FDI indica si el ítem fue modificado para que muestre FDI o si no sufrió modificación. Semejante a este concepto, es necesario recordar que el "Error tipo I" está vinculado con los falsos positivos, es decir, en el contexto del FDI, que el método de detección de FDI indique que el ítem posee FDI sin que realmente sea

así. En este sentido, la hipótesis nula es: El método de detección de FDI indica si un ítem que no fue modificado muestra FDI.

Capítulo 3

Metodología

La propuesta de diseño nace de la necesidad de crear una herramienta que analice y aporte evidencias respecto al alcance de los métodos de detección del FDI. En primera instancia, la caracterización teórica responde a la obligación por describir las condiciones psicométricas apropiadas para un estudio sobre el FDI. En ese sentido, es indispensable valorar, en la fase dos, las condiciones en las cuales las cinco técnicas que identifican el FDI son más adecuadas. Finalmente, la fase tres implementa el análisis del comportamiento diferencial del ítem con los datos obtenidos en la PAA del 2016 en razonamiento verbal y contexto matemático.

3.1. Experimentación de técnicas para identificar el FDI

Los experimentos están diseñados para medir el efecto que tiene la modificación en la dificultad y discriminación de ítems (variables independientes) sobre la potencia

y error tipo I de los métodos de detección del FDI (variables dependientes). Por lo tanto, es indispensable crear las condiciones experimentales más cercanas a la realidad de un estudio sobre el FDI. Eso requiere generar interacciones probables de examinados frente a una lista de ítems.

En el año 2016, la PAA estuvo conformada por cuatro fórmulas (compuestas por 85 ítems) aplicadas a un total de 60 000 estudiantes. El estudio del FDI sobre la PAA se realizará tomando en cuenta la base de datos para la fórmula 1, por lo que el estudio de simulación busca replicar las condiciones en esa primera fórmula. Dicha distribución permite utilizar la función creada (ver detalles en Anexo A) en R Core Team (2016) con réplicas de datos de 15 000 individuos sobre 85 ítems.

Dicha función tiene tres entradas (cantidad de individuos, cantidad de ítems, número de iteración) y permite generar una base de datos. Sobre esta base de datos se realiza el estudio del FDI. Inicialmente, a cada individuo se le asigna una habilidad a partir de la distribución normal. Paralelamente, cada ítem recibe un valor de dificultad y de discriminación a partir de la distribución normal. Luego, se calcula la probabilidad de éxito del individuo, dada la habilidad asignada frente a la dificultad y discriminación de los ítems, según la ecuación 2.2 del modelo de dos parámetros de la TRI.

El objetivo del estudio de simulación es evaluar las técnicas de detección del FDI en escenarios de funcionamiento diferencial fuerte, funcionamiento diferencial moderado y funcionamiento diferencial débil. En este caso dicha clasificación responde al grado de modificación de la dificultad (uniforme) o discriminación (no uniforme) sobre 4 ítems pertenecientes al grupo focal. Si la modificación realizada es exclusiva de la dificultad del ítem, entonces la condición se clasifica como uniforme. Si la modificación realizada es exclusiva de la discriminación del ítem, entonces la condición se clasifica como no uniforme. Si la modificación realizada es compartida

entre dificultad y discriminación (2 ítems cada una), entonces la condición se clasifica como mixta.

Para estos efectos, se espera que la potencia sea lo más cercana a 1, mientras que el "Error tipo I" esté cercano al 5%, esto porque los 4 ítems modificados son aproximadamente el 5% de la cantidad total de ítems, 85.

La tabla 3.1 muestra el grado de las modificaciones según el tipo de FDI:

Tabla 3.1

Variables Independientes

	Grado de Modificación		
Tipo FDI	Fuerte	Moderado	Débil
Uniforme	±1	±0.5	±0.25
No Uniforme	±0,5	±0.25	±0.1

Las cinco técnicas de detección del FDI serán estratificadas según su error tipo I y potencia estadística (variables dependientes). Se realizarán 100 réplicas de datos por tipo de FDI (uniforme, no uniforme y mixto) en tres grados distintos (fuerte, moderado y débil), para un total de 900 réplicas de datos. Los resultados se resumirán en tablas mediante el promedio (incluye desviación estándar) obtenido de las 100 réplicas de datos por tipo de FDI y grado de modificación. Se considera que 100 réplicas son suficientes para garantizar resultados estables, pues son métodos ampliamente estudiados.

3.2. Estudio del FDI en la PAA 2016 según sexo

Para el 2016, la PAA contó con cuatro cuadernillos (denominados Fórmulas) para la población sin adecuaciones en acceso, los cuales estaban compuestos por 85 ítems. Para efectos de este estudio se decidió contemplar únicamente la Fórmula 1, pues el proceso de análisis de los demás es análogo.

Esta fórmula la realizaron 11 592 personas, de las cuales, algunas aspiraban ingresar a la UCR en el 2017 o esperaban trasladarse de carrera en el 2017. Estas representan todas las zonas geográficas de Costa Rica, distintas modalidades de la Educación Diversificada de Costa Rica, distintos niveles económicos y la mayoría de ellas están entre los 17 y 20 años.

En esta muestra de personas que realizaron la PAA, 9 400 son egresadas de sistemas educativos públicos (4 157 hombre y 5 243 mujeres) y los restantes 2 192, de sistemas privados y subvencionados (1 041 hombres y 1 151 mujeres).

Por razones teóricas es indispensable clasificar los grupos comparativos según sexo. Por lo tanto, es necesario enfatizar que la muestra considerada estaba compuesta por 5 198 hombres (44,84 %) y 6 394 mujeres (55,16 %).

Para realizar análisis del FDI, como mínimo, se debe garantizar la unidimensionalidad de la prueba y una alta calidad técnica del test. En la primer etapa se puede utilizar una técnica multivariada como el Análisis Factorial Exploratorio. Y para garantizar una alta calidad técnica de la PAA, se utilizan los modelos de medición TCT y TRI.

Capítulo 4

Resultados

4.1. Estudio de simulación

La siguiente sección muestra los resultados obtenidos en el estudio de simulación realizadas para medir la potencia y error tipo I de las técnicas de detección del FDI. La distribución de los resultados se divide por tipo de FDI (uniforme, no uniforme y mixto) y el grado de modificación (fuerte, débil y moderado).

4.1.1. FDI Uniforme

Las cinco técnicas de detección del FDI tienen una alta dependencia de la dificultad del ítem. Por lo tanto, suelen identificar correctamente el FDI uniforme. La diferencia entre las técnicas de detección es evidente según el grado de modificación (fuerte, débil, moderado).

Modificación Fuerte

En la tabla 4.1, se observa que las cinco técnicas de detección del FDI tienen un excelente rendimiento en modificaciones fuertes. Cuatro de las técnicas presentan un 100% de potencia. Resalta el rendimiento del proceso de estandarización cuyo promedio de error tipo I es 0%. Por el contrario, Raju presenta el porcentaje de error tipo I más alto de la experimentación con un 32.7%.

Tabla 4.1

Resultados Simulación Uniforme Fuerte

	Potencia		Error Tipo 1	
Método	$\bar{X}$	*SD*	$\bar{X}$	*SD*
TRI-Lord	1.000	0.000	0.217	0.136
TRI-Raju	1.000	0.000	0.327	0.144
TCT-Log	1.000	0.000	0.273	0.120
TCT-MH	1.000	0.000	0.321	0.133
TCT-Est	0.995	0.035	0.000	0.000

Modificación Moderada

En la tabla 4.2 se observa que cuatro técnicas de detección del FDI presentan un excelente rendimiento en modificaciones moderadas. Las cuatro técnicas presentan un 100% de potencia. Se debe resaltar que el rendimiento del proceso de estandarización es regular. Logró captar un 66% de los ítems con funcionamiento diferencial. Raju presenta el porcentaje de error tipo I más alto de la experimentación con un 14.5%.

Tabla 4.2

Resultados Simulación Uniforme Moderada

	Potencia		Error Tipo 1	
Método	$\overline{X}$	*SD*	$\overline{X}$	*SD*
TRI-Lord	1.000	0,000	0,073	0,039
TRI-Raju	1.000	0,000	0,145	0,053
TCT-Log	1.000	0,000	0,108	0,043
TCT-MH	1.000	0,000	0,125	0,048
TCT-Est	0,662	0,239	0,000	0,000

Modificación Débil

En la tabla 4.3 se observa que cuatro técnicas de detección del FDI presentan un excelente rendimiento en modificaciones débiles. Las cuatro técnicas presentan un valor cercano al 100 % de potencia. Se debe resaltar que el rendimiento del proceso de estandarización es deficiente. No logró captar ningún ítem con FDI, con una potencia del 0 %. Además, Raju presenta el porcentaje de error tipo I más alto de la experimentación con un 10 %.

Tabla 4.3

Resultados Simulación Uniforme Débil

	Potencia		Error Tipo 1	
Método	$\overline{X}$	*SD*	$\overline{X}$	*SD*
TRI-Lord	1.000	0,000	0,045	0,025
TRI-Raju	0.995	0,035	0,100	0,041
TCT-Log	1.000	0,000	0,066	0,027
TCT-MH	1.000	0,000	0,068	0,027
TCT-Est	0.000	0,000	0,000	0,000

4.1.2. FDI No Uniforme

Tres técnicas de detección del FDI (Lord, Raju y Logística) suelen identificar correctamente el FDI no uniforme. Las técnicas que se basan en la TCT generalmente fallan en la detección de modificaciones en la discriminación del ítem. La diferencia entre las técnicas de detección es evidente según el grado de modificación (fuerte, débil, moderado).

Modificación Fuerte

En la tabla 4.4 se observa que las tres técnicas basadas en TCT obtuvieron resultados muy distintos respecto a la potencia. La técnica logística se ajusta perfectamente a las técnicas TRI con una potencia del 100%. MH presenta un resultado aceptable con una potencia del 83.5%. El proceso de estandarización es deficiente, con una potencia del 9%. Raju se mantiene como la técnica de mayor error tipo I con un 7.8%.

Tabla 4.4

Resultados Simulación No Uniforme Fuerte

	Potencia		Error Tipo 1	
Método	$\bar{X}$	*SD*	$\bar{X}$	*SD*
TRI-Lord	1.000	0.000	0.033	0.026
TRI-Raju	1.000	0.000	0.078	0.039
TCT-Log	1.000	0.000	0.058	0.026
TCT-MH	0.835	0.189	0.056	0.026
TCT-Est	0.095	0.158	0.000	0.000

Modificación Moderada

En la tabla 4.5 se observa que las tres técnicas basadas en TCT obtuvieron resultados muy distintos con respecto a la potencia. La técnica logística se ajusta perfectamente a las técnicas TRI con una potencia cercana al 100%. MH presenta un promedio regular con una potencia del 71%. El proceso de estandarización es deficiente, con una potencia del 0%. Raju se mantiene como la técnica de mayor error tipo I con un 8.2%.

Tabla 4.5

Resultados Simulación No Uniforme Moderada

	Potencia		Error Tipo 1	
Método	$\overline{X}$	*SD*	$\overline{X}$	*SD*
TRI-Lord	0.992	0.043	0.037	0.034
TRI-Raju	0.992	0.043	0.082	0.042
TCT-Log	0.992	0.043	0.051	0.023
TCT-MH	0.710	0.224	0.052	0.023
TCT-Est	0.000	0.000	0.000	0.000

Modificación Débil

En la tabla 4.6 se observa que tres técnicas (Lord, Raju y Logística) sobresalen en el escenario no uniforme con modificación débil. Las tres técnicas tiene una potencia cercana al 80%, frente al pobre desempeño de MH (34%) y el proceso de estandarización (0%). Raju se mantiene como la técnica de mayor error tipo I con un 8.9%.

Tabla 4.6

Resultados Simulación No Uniforme Débil

	Potencia		Error Tipo 1	
Método	$\overline{X}$	*SD*	$\overline{X}$	*SD*
TRI-Lord	0.752	0.206	0.036	0.031
TRI-Raju	0.805	0.200	0.089	0.053
TCT-Log	0.790	0.194	0.049	0.023
TCT-MH	0.340	0.209	0.048	0.025
TCT-Est	0.000	0.000	0.000	0.000

4.1.3. FDI Mixto

La condición de FDI mixto permite estudiar las cinco técnicas de detección desde el tipo de FDI (uniforme o no uniforme). Las técnicas que se basan en la TRI generalmente tienen buen desempeño en ambos tipos de FDI. La diferencia entre las técnicas de detección es evidente según el grado de modificación (fuerte, débil, moderado).

Modificación Fuerte

En la tabla 4.7 se observa que cuatro técnicas (Lord, Raju, Logística y MH) presentan un excelente promedio de potencia. MH tiene el valor más bajo con una potencia del 92.5. El proceso de estandarización se aleja del comportamiento de las demás técnicas con una potencia del 56.7%. Raju se mantiene como la técnica de mayor error tipo I con un 16.6%.

Tabla 4.7

Resultados Simulación Mixto Fuerte

	Potencia		Error Tipo 1	
Método	$\overline{X}$	*SD*	$\overline{X}$	*SD*
TRI-Lord	1,000	0,000	0,083	0,046
TRI-Raju	1,000	0,000	0,166	0,065
TCT-Log	1,000	0,000	0,131	0,060
TCT-MH	0,925	0,126	0,146	0,070
TCT-Est	0,567	0,122	0,000	0,000

Modificación Moderada

En la tabla 4.8 se observa que tres técnicas (Lord, Raju y Logística) sobresalen en el escenario mixto con modificación moderada. Estas tres técnicas tienen una potencia del 100 %. Por otro lado, se evidencia un buen desempeño en la potencia de MH (83,7 %). El proceso de estandarización (35 %) obtuvo la potencia más baja de la simulación. Raju se mantiene como la técnica de mayor error tipo I con un 9,8 %.

Tabla 4.8

Resultados Simulación Mixto Moderado

	Potencia		Error Tipo 1	
Método	$\overline{X}$	*SD*	$\overline{X}$	*SD*
TRI-Lord	1,000	0,000	0,042	0,024
TRI-Raju	1,000	0,000	0,098	0,033
TCT-Log	1,000	0,000	0,069	0,031
TCT-MH	0,837	0,148	0,072	0,031
TCT-Est	0,350	0,159	0,000	0,000

Modificación Débil

En la tabla 4.9 se observa que tres técnicas (Lord, Raju y Logística) sobresalen en el escenario mixto con modificación débil. Las tres técnicas tiene una potencia cercana al 85%, frente al regular desempeño de MH (65%). El proceso de estandarización (0%) no logró captar correctamente el FDI. Raju se mantiene como la técnica de mayor error tipo I con un 8,7%.

Tabla 4.9

Resultados Simulación Mixto Débil

	Potencia		Error Tipo 1	
Método	$\bar{X}$	*SD*	$\bar{X}$	*SD*
TRI-Lord	0,837	0.148	0,039	0.023
TRI-Raju	0,867	0,157	0,087	0,035
TCT-Log	0,847	0,150	0,055	0,027
TCT-MH	0,650	0,151	0,055	0,025
TCT-Est	0,000	0.000	0.000	0.000

4.2. Evidencias de constructo

4.2.1. Análisis de confiabilidad según la TCT

Para analizar la confiabilidad de la Fórmula 1 de la PAA, se establecieron dos grupos de ítems: 1) todos los ítems ensamblados, es decir, los 85 ítems y 2) todos los ítems que cumplen con los parámetros del PPPAA para considerar que un ítem puede formar parte de su batería o banco de ítems. Este segundo grupo de ítems se denominaran como Ajustados.

Como primer resultado de este análisis, en la tabla 4.10 se *observa que* para ambos grupos de ítems, el Alfa de Cronbach es aceptable para pruebas estandarizadas, sin embargo, y como es de esperar, una vez que se eliminan los 8 ítems que no ajustaban, el Alpha de Cronbach de la fórmula aumentó en 3 milésimas, lo cual es deseable.

Tabla 4.10

Alpha según bases

Items	Núm. de ítems	raw_alpha	std.alpha
Todos	85	0.917	0.915
Ajustados	77	0.920	0.919

Lo anterior significa que para ambos grupos de ítems, si se aplican a poblaciones equivalentes, los resultados de las puntuaciones de la Prueba serán semejantes, lo cual es uno de los resultados deseables en una prueba estandarizada.

Profundizando en este estudio se concluye que si se excluyen de los análisis a los ítems i14, i19, i27, i57, i70 e i85, se mejora el Alpha de Cronbach estandarizado, pues, para cada uno, pasa de 0.915 a 0.916, aunque la diferencia no es tan significativa. Esta información, correspondiente a la TCT, sugiere que no es necesario excluir ningún ítem de los análisis, pero de ser así, se debe considerar si alguno de los ítems o combinaciones de los mismos, miden elementos del constructo que no miden los demás. Por tal motivo, es necesario analizar el comportamiento de los ítems en términos del constructo.

Para decidir la exclusión de los 8 ítems se consideró el proceso que se utiliza en el PPPAA para tomar la decisión de si un ítem se mantiene en el banco de ítems o si, es un ítem piloto, definir su ingreso al banco de ítems. A continuación, se detallan los resultados de dicho análisis.

4.2.2. Análisis Factorial Exploratorio: todos los ítems

En la figura 4.1 se observa que los ítems de la Fórmula 1 de la PAA 2016 se configuran en un solo factor, siendo el primer valor propio 10.667 y el segundo, 1.264. Esto es una primera aproximación que sugiere que toda la fórmula mide un único factor, lo cual combinado con el Alfa de Cronbach, se espera que dicha medición sea bastante estable. En principio, esto significa que todos los ítems miden el mismo constructo y que, por tanto, se pueden excluir una cantidad moderada de ítems, en caso de ser necesario para repetir el procedimiento. Para saber cuáles ítems no miden adecuadamente este constructo, lo cual es una razón para desear excluirlos de los análisis, es preferible conocer, para cada uno de los ítems, qué tan bien mide el factor. Para este análisis, se recomienda estudiar las cargas factoriales de cada uno de los ítems.

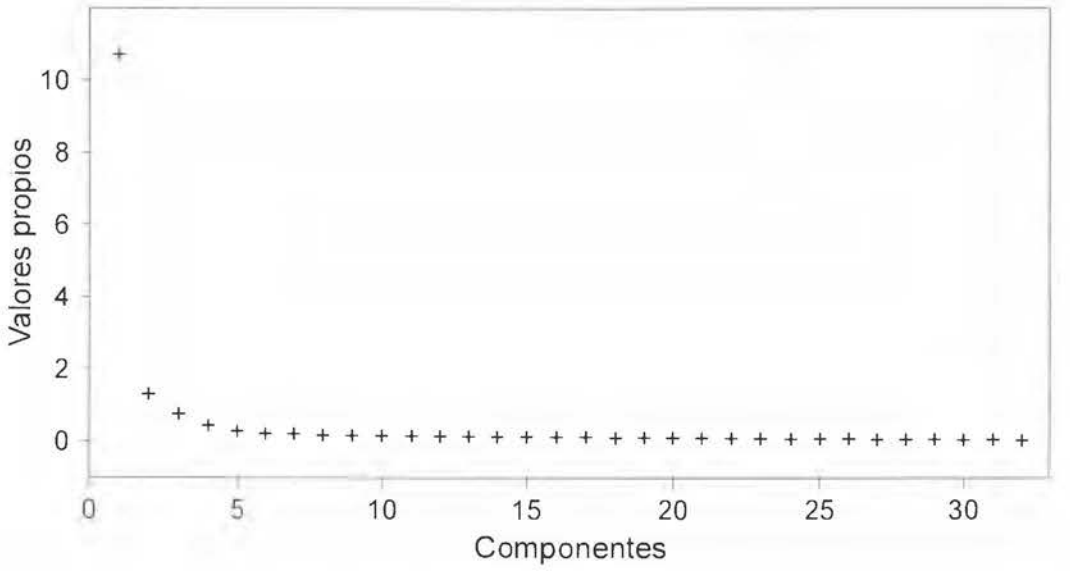

Figura 4.1. Gráfico de sedimentación (todos los ítems)

Para realizar el análisis de las cargas factoriales de los ítems, normalmente, se considera que la carga factorial del modelo de un factor sea superior a 0.20. Como

resultado de este, se obtuvo que los 6 ítems que presentaron deficiencias en el Alfa de Cronbach, también tienen cargas factoriales bajas. Esto aporta más evidencias de que estos ítems deben ser excluidos de los análisis. Adicionalmente, de este análisis, los ítems i80 e i81 tienen cargas factoriales por debajo de 0.17, mientras que los ítems i9, i26, i76 e i84 poseen cargas entre los 0,17 y 0,20. Las cargas factoriales de los ítems estuvieron entre 0,1830 y 0,5280, siendo el promedio 0,3584 y la desviación estándar 0,0922.

Una vez realizado el análisis factorial exploratorio y procurando más evidencias para tomar la decisión para excluir ítems del análisis, se procede al análisis de los ítems con la TCT y la TRI.

4.2.3. TCT y TRI: todos los ítems

Como parte del estudio desde la TCT se observa que los ítems i14, i19, i27, i57, i70, i76, i80, i81 e i85 no alcanzan el nivel mínimo del valor de discriminación, el cual es de 0.2. De estos 9 ítems, 3 son ítems de banco, de donde se rescata que el ítem 85, de nuevo presenta índices de ajuste poco aceptables. Además, se concluye que los 6 ítems que presentaron deficiencias en el Alfa de Cronbach reinciden, al presentar problemas de ajuste, esta vez, con el parámetro de discriminación. Además, ocurre lo mismo para los ítems i80 e i81, pues apenas alcanzan un parámetro de discriminación de 0.196. Los demás ítems muestran índices de discriminación superiores a 0.20 e índices de dificultad aceptables. El parámetro de dificultad de la TCT estuvo entre 0.1290 y 0.7550, siendo el promedio 0,4289 y la desviación estándar 0.1519; mientras que el de discriminación estuvo entre 0,2110 y 0,5280, con promedio 0.3738 y desviación estándar 0.0861.

Para realizar el análisis con el modelo TRI-2PL se considera que la discriminación no sea menos a 0,35, que la dificultad del ítem esté en [−3,5; 3,5] y que la imagen de 1 sobre la curva de información no sea menor a 0,1. Como parte de los resultados obtenidos se observa que los 8 ítems que presentaron problemas en la TCT, también presentan valores no aceptables en los parámetros de discriminación y en la imagen sobre la curva de información, incluso, cuatro de ellos no están dentro de los parámetros de dificultad aceptables. El parámetro de dificultad de la TRI estuvo entre −1,4730 y 3,4820, siendo el promedio 0.5034 y la desviación estándar 1,0430; mientras que el de discriminación estuvo entre 0.2200 y 0.8460, con promedio 0.5157 y desviación estándar 0,1610. También, la imagen del nivel de habilidad 1 de la curva de información de los ítems estuvo entre 0.0330 y 0,5040, con promedio 0.1554 y desviación estándar 0.0918.

Como consecuencia de los resultados obtenidos, se decide considerar todos los demás ítems, es decir, excluir en análisis subsecuentes los ítems i14, i19, i27, i57, i70, i80, i81 e i85.

4.2.4. Análisis Factorial Exploratorio: ítems ajustados

Una vez que se eliminaron los 8 ítems, en la figura 4.2 se observa que los gráficos de sedimentación siguen siendo prácticamente el mismo. La diferencia respecto al análisis AFE con todos los ítems y este, es que para los ítems ajustados, el primer valor propio 10,538 y el segundo, 1,200.

Para el caso de las cargas factoriales, solamente los ítems i9, i26, i76 e i84 tienen pesos ligeramente menores al esperado de 0.20.

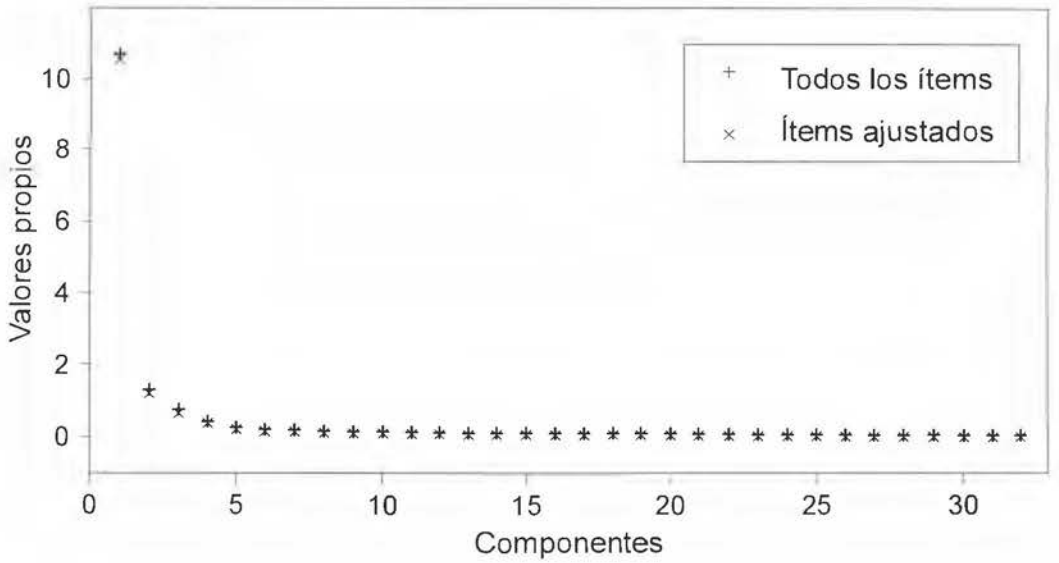

Figura 4.2. Gráfico de sedimentación (todos los ítems)

4.3. Estudio del análisis Funcionamiento Diferencial del Ítem de la PAA 2016

Las siguientes tablas muestran un resumen del análisis FDI comportamiento de la Fórmula 1 de la PAA 2016, sin incorporar los ítems que no ajustaron al modelo de medición y con una población de 11 592 individuos.

Tabla 4.11

Estudio DIF sin no ajustables: 11 592 individuos

item	DIF	Lord	DIF	Raju	DIF	Log	DIF	MH	DIF	Est
i1	0	2,69	0	1,42	0	1,76	0	1,77	0	0,01
i2	1	14.79	1	3.39	1	10.86	1	10.08	0	-0.03
i3	0	4.77	0	1.10	1	5.99	1	5.63	0	0.02
i4	1	24 07	1	4.16	1	29.48	1	29.21	0	0.05
i5	1	35.57	1	5.92	1	48.22	1	45.35	0	0.06
i6	1	28.19	1	4.72	1	19.90	1	4.54	0	-0.02

Continúa en la próxima página

Tabla 4.11 (*Continúa de la página anterior*)

item	DIF	Lord	DIF	Raju	DIF	Log	DIF	MH	DIF	Est
i7	1	58,19	1	6,35	1	68.77	1	64.83	0	0.07
i8	1	21,97	1	4,52	1	11,79	0	0.07	0	-0.01
i9	1	40,37	1	-6,32	1	47.65	1	46.45	0	0.06
i10	1	39,94	1	6,30	1	52.07	1	50.11	0	0.06
i11	1	35,16	1	-4,18	1	50,42	1	39.48	0	0.06
i12	1	9,24	1	2,34	1	8,49	1	8.64	0	0.03
i13	1	42,73	1	5.79	1	44.12	1	41.60	0	0.05
i15	0	1,15	0	0,91	0	0,78	0	0,71	0	0,01
i16	1	19,44	1	4,02	1	11,40	0	0.01	0	-0.00
i17	1	45,08	1	-5,72	1	54,49	1	48.84	0	0.06
i18	1	49,57	1	5,32	1	52.35	1	47.67	0	0.06
i20	1	7,65	1	2,71	0	1,64	0	0.10	0	0.00
i21	0	5,75	0	1,74	0	4,23	0	3,38	0	0,01
i22	0	5,14	1	-2,19	1	9,92	0	0.17	0	-0.00
i23	1	66,26	1	6,85	1	43,58	0	0.07	0	-0.00
i24	1	20,21	1	3,34	1	16.16	1	10.62	0	-0.02
i25	1	49,61	1	6,65	1	44,83	1	29.88	0	0 03
i26	1	64,75	1	4,54	1	50,32	0	1.23	0	-0.02
i28	0	0,60	0	-0,77	0	3,06	0	0,00	0	0.00
i29	1	12,02	1	3,42	0	2.62	0	2.23	0	0.01
i30	0	4,13	1	-2,01	0	5.17	1	4.50	0	-0.02
i31	1	19,37	1	3,85	1	20,60	1	20.50	0	0.04
i32	1	107,09	1	10.06	1	134.85	1	119.14	0	0.09
i33	0	3,30	0	1,62	0	3,45	0	3,05	0	-0 01
i34	1	92,37	1	7,91	1	106.23	1	101.55	0	0.09
i35	1	29,75	1	5,34	1	40.17	1	39.14	0	-0.05
i36	1	15,33	1	3,36	1	16,07	1	15.15	0	-0.03
i37	0	0,38	0	0,60	0	0,62	0	0,04	0	-0.00
i38	0	4,59	1	2.08	0	1.67	0	0.52	0	0.00
i39	1	126,66	1	10.10	1	143.43	1	139.12	0	0.09
i40	1	16,46	1	3,39	1	12.73	1	10.23	0	0.02
i41	1	9,02	1	2.93	0	2.69	0	0.22	0	-0.01
i42	1	17,90	1	3.96	1	9.18	1	4.57	0	-0.02
i43	1	10,93	1	3,13	0	5.34	1	4.71	0	0.01
i44	1	21,57	1	4,04	1	23.36	1	22.76	0	0.04
i45	1	16,97	1	3,53	1	18.69	1	18.43	0	-0.04
i46	1	33,52	1	4,80	1	36.09	1	36.80	0	-0.05

Continúa en la próxima página

Tabla 4.11 (*Continúa de la página anterior*)

item	DIF	Lord	DIF	Raju	DIF	Log	DIF	MH	DIF	Est
i47	1	56,36	1	-4,49	1	70,13	1	38,76	0	0,05
i48	0	1,34	0	1,15	0	3,39	0	2,77	0	-0,01
i49	1	32,54	1	5,05	1	21,43	0	1,64	0	-0,02
i50	1	10,16	1	2,83	1	8,25	1	6,82	0	0,01
i51	1	12,46	1	3,39	1	6,92	0	0,00	0	-0,00
i52	1	47,98	1	6,71	1	57,47	1	57,24	0	-0,07
i53	1	9,27	1	2,72	1	7,98	1	5,85	0	0,02
i54	1	51,24	1	6,61	1	57,73	1	53,14	0	-0,06
i55	1	22,85	1	-4,33	1	22,53	1	19,63	0	-0,04
i56	1	29,29	1	5,40	1	39,34	1	37,52	0	-0,05
i58	1	40,42	1	5,19	1	43,68	1	43,09	0	-0,05
i59	0	0,13	0	-0,30	0	2,35	0	1,03	0	0,01
i60	1	18,97	1	3,99	1	9,66	1	3,89	0	-0,02
i61	1	77,99	1	6,67	1	98,41	1	82,83	0	-0,08
i62	1	19,08	1	4,05	1	19,67	1	19,33	0	-0,04
i63	1	24,91	1	3,71	1	19,08	1	12,79	0	-0,03
i64	1	7,91	1	2,74	1	10,74	1	9,93	0	-0,02
i65	1	79,08	1	-8,82	1	111,55	1	106,13	0	-0,08
i66	1	10,46	1	2,73	1	16,60	1	15,51	0	-0,03
i67	0	4,95	1	-1,96	1	8,55	1	6,35	0	-0,02
i68	1	32,59	1	-4,84	1	46,91	1	15,78	0	-0,03
i69	1	17,85	1	-3,90	1	33,15	1	8,49	0	-0,02
i71	1	15,28	1	3,68	1	14,77	1	14,16	0	-0,03
i72	1	21,15	1	-3,40	1	34,78	1	30,93	0	-0,04
i73	1	47,31	1	-4,09	1	60,54	1	38,73	0	0,06
i74	1	11,57	1	-2,72	1	18,65	1	9,58	0	-0,02
i75	1	11,26	1	-3,14	1	24,54	1	10,25	0	-0,02
i76	1	8,35	1	-2,28	1	14,02	1	7,81	0	0,02
i77	1	15,87	1	-3,60	1	29,94	1	13,98	0	-0,02
i78	1	6,45	1	-2,01	1	17,44	1	11,36	0	-0,02
i79	0	4,81	0	-1,37	1	9,27	1	5,68	0	-0,01
i82	0	2,66	0	-0,83	0	5,76	0	1,84	0	0,01
i83	0	2,59	0	1,42	0	1,84	0	1,21	0	0,01
i84	1	6,97	0	-1,51	1	11,74	1	8,49	0	-0,02

El estudio de FDI, sobre la PAA 2016 con la población completa de la fórmula 1, revela comportamientos dispares entre los métodos de detección. Los resultados se agruparon en los siguientes bloques:

4.3.1. FDI nulo

El estudio reveló concordancia entre los cinco métodos de detección del FDI en 10 ítems: i1, i15, i21, i28, i33, i37, i48, i59, i82 e i83. Dicho resultado debe resaltarse debido a las diferencias sustanciales entre los métodos de detección y su desempeño variado en el estudio de simulación preliminares. Las salidas generadas por el proceso de estandarización carecen de importancia, dado que todas las salidas fueron 'FDI nulo". Dicho resultado concuerda con las salidas registradas en algunas condiciones del estudio de simulación (0% potencia, 0% error tipo I).

4.3.2. Ajuste en FDI No Uniforme

El estudio también reveló concordancia entre los tres métodos de detección que tienen la posibilidad de identificar funcionamiento diferencial no uniforme. En este bloque, los resultados se contrastan con MH y Estandarización. Los resultados destacados son los siguientes 6 ítems: i8, i16, i23, i26, i49 e i51. Dicho resultado debe resaltarse debido a que MH es uno de los métodos que se utiliza regularmente en el estudio de la PAA.

4.3.3. Mejor desempeño según error tipo I

Las tres técnicas (que teóricamente identifican FDI no uniforme) que tuvieron mejor desempeño en el estudio de simulación, según error tipo I fueron: Logística, Lord y Raju (de mejor a menor desempeño). Por lo tanto, es indispensable señalar los únicos 6 ítems donde alguno de esos 3 métodos descarta FDI frente al resto. Los 6 ítems identificados son: i3, i30, i43, i67, i79 e i84. El proceso de estandarización no forma parte de la interpretación debido a que su comportamiento fue "FDI nulo"en todos los ítems.

4.3. ESTUDIO DEL ANÁLISIS FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM DE LA PAA 2016

Capítulo 5

Conclusiones y Recomendaciones

5.1. Conclusiones

5.1.1. Resumen de los métodos

Como parte de la revisión de literatura y la disponibilidad en el paquete difR (Magis et al., 2015) se decidió efectuar los estudios de detección de FDI con 5 métodos, 3 de los cuales son de la TCT (MH, Estandarización y Regresión logística) y los otros 2 (Lord y Raju), de la TRI. De estos cinco métodos, 3 tienen mayor potencia para detectar FDI no uniforme (las dos de TRI y Regresión logística), mientras que, según la teoría, las otras dos, no.

El método de MH se centra en la probabilidad que tiene cada grupo de acertar el ítem para lo cual se utiliza un estadístico que se comporta como la χ^2. Este método se considera de la TCT por centrarse en los puntos que obtuvieron los examinados en el test. El método se basa en la habilidad definida por la puntuación *total en el test*,

es decir, es altamente dependiente de la dificultad del ítems y no en la discriminación, por lo que, en principio, no estaría en capacidad de detectar FDI no uniforme.

Por su parte, Estandarización, se enfoca en comparar la regresión entre las respuestas de los ítems y el puntaje obtenido en el test, esto para cada uno de los dos grupos. Debido a que este método se centra en los puntajes obtenidos en el test, este método se considera uno de la Teoría Clásica de los Tests. Además, en este método se define la habilidad del examinado según su puntaje total en el test; es altamente dependiente de la dificultad del ítem por lo que no estaría en la capacidad de detectar FDI no uniforme.

El método Regresión logística, tal y como indica el nombre, realiza una regresión logística donde las variables independientes son las respuestas a los ítems, el grupo al que se pertenece y la combinación de las dos anteriores, es decir, las respuestas de cada ítem dependiendo del grupo al cual se pertenece. Precisamente, esta última variable es la que permite a este método determinar si existe FDI uniforme o no uniforme. Como la variable dependiente es el puntaje obtenido en el test, se sugiere que este método se basa en la TCT.

El método Chi cuadrado de Lord se basa en que los parámetros de TRI deben ser esencialmente los mismos para que no haya FDI. Esto indica que en efecto se refiere a un modelo concebido en la TRI y que según los parámetros se puede asegurar la existencia de FDI uniforme o no uniforme. El estadístico de este método sigue una distribución chi cuadrada asintótica.

El método de Raju consiste en comparar las áreas entre las curvas características de los ítems del grupo focal y del grupo de referencia, considerando los parámetros de cada uno de los grupos. Por esta razón, este método se clasifica dentro de los

modelos de la TRI. Además, al considerar distintos niveles de habilidad, también permite detectar FDI no uniforme.

5.1.2. MH y Estandarización con deficiente rendimiento

El rendimiento de las técnicas de detección MH y Estandarización son deficientes debido a su composición estructural. Ambas técnicas dependen exclusivamente de la dificultad del ítem y del puntaje total de la prueba. Eso imposibilita que las técnicas puedan identificar el funcionamiento diferencial no uniforme (dependiente de la discriminación del ítem). En ese escenario teórico, el estudio de simulación confirmó lo descrito anteriormente.

Tomando como referencia las condiciones del estudio de simulación, en los casos donde las modificaciones se realizaron en la discriminación del ítem (FDI no uniforme), en escenarios moderados/débiles, MH y Estandarización no superan el 85% de potencia. En casos extremos, como la modificación débil, Estandarización no logró captar ningún ítem con funcionamiento diferencial. Dicha potencia del 0% es deficiente y el 34% de potencia para MH, en el mismo escenario, tampoco es un resultado positivo.

La principal preocupación del pobre desempeño de MH y el proceso de estandarización es que los estudios que dependen de dichos métodos no tendrían la capacidad de identificar ítems cuya discriminación actúa diferencialmente. En el caso del estudio sobre la PAA 2016 con 11 592 individuos, se pudo observar que 6 ítems cumplen con esa condición (i8, i16,i23, i26, i49 y i51), donde las técnicas capaces de identificar funcionamiento diferencial no uniforme tuvieron un desempeño distinto frente a las que no.

5.1.3. Efecto del tamaño de la muestra

Los métodos de detección de FDI implementados utilizan pruebas estadísticas que son sensibles a los tamaños de la muestra, ya sea por ser muy pequeñas o demasiado grandes. Esto significa que al utilizar los métodos de detección de FDI seleccionados se debe proceder con cautela al decidir si el ítem realmente posee funcionamiento diferencial según el grupo al cual pertenecen. Wasserstein y Lazar realizaron dicha aclaración en el año 2016 con el manual sobre uso e interpretación de valores p. En particular, la publicación recomienda que las conclusiones científicas no deben basarse únicamente en si un valor p o significancia estadística sobrepasa una tolerancia específica. De esa forma se pueden evitar resultados inestables o interpretaciones que carecen de evidencias sólidas.

5.1.4. PAA 2016 con múltiples ítems con FDI

En el caso de la muestra utilizada en la Fórmula 1 de la PAA 2016, al ser más de 11 mil examinados, todas las pruebas estadísticas tienden a rechazarse, pues los métodos de detección de Funcionamiento Diferencial del Ítem estudiados se vuelven muy sensibles a pequeñas variaciones que se van acumulando hasta rechazarse. Como consecuencia de la cantidad de examinados y al no considerar una equiparación o diseño para compensar este hecho, los resultados de la aplicación de los métodos de detección de FDI arrojan que prácticamente todos los ítems poseen FDI, aunque el tamaño de la diferencia es muy pequeña. Por tal motivo, es necesario realizar más estudios para verificar si realmente todos estos ítems poseen FDI o si es una limitación de las pruebas estadísticas utilizadas en cada uno de los métodos.

5.1.5. Implicaciones en la Enseñanza de la Matemática

Como parte de la búsqueda de equidad en el acceso a educación de calidad planteada en el Programa de Estudios de Matemáticas (MEP, 2012) y considerando el sistema de medición de las pruebas administradas por el Ministerio de Educación, el estudio del FDI debe ser parte fundamental de la calidad técnica de las pruebas estandarizadas. Por ende, es de vital importancia complementar la formación de los docentes de matemáticas en este tipo de pruebas. Esto significa que los futuros docentes de matemática deben demostrar su capacidad en:

1. Caracterizar el FDI.

2. Crear ítems para pruebas estandarizadas que minimicen la presencia de un funcionamiento diferente para los distintos grupos de población.

3. Identificar la influencia de las mediciones en actividades de aula.

Los futuros docentes de matemática pueden tener un acercamiento a las distintas formas cuantitativas de detección de ítems que generaran diferencias a los distintos grupos de población. Esto permitirá realizar análisis cualitativos adicionales que eviten el sesgo que podría generar una herramienta de evaluación que define el futuro de una persona, tal es el caso de las evaluaciones de aula y las pruebas estandarizadas.

5.2. Recomendaciones

5.2.1. Estudios futuros del FDI de la PAA

La PAA es un instrumento que ajusta al modelo de medición TRI en dos parámetros. En ese sentido, cualquier estudio sobre comportamiento diferencial del ítem debería considerar métodos de detección construidos bajo hipótesis TRI.

El cambio en la utilización de métodos de detección es importante porque el análisis de la prueba sería consistente con su modelo de medición. La justificación del cambio es evidente: los métodos de detección que tienen la capacidad de identificar funcionamiento diferencial no uniforme son más potentes en escenarios mixtos, bajo modificaciones fuertes y moderadas.

Se recomienda que el PPPAA no utilice en el futuro los métodos MH y Estandarización, sino que haga una migración hacia métodos que tienen la posibilidad de detectar el comportamiento diferencial uniforme y no uniforme.

5.2.2. Nuevas líneas de investigación

Existen muchos métodos de detección del FDI que tienen la capacidad de identificar comportamiento diferencial no uniforme, algunos de ellos: Chi cuadrado Lord, Raju, Regresión logística, Método de máxima verosimilitud, entre otros.

Se recomienda al PPPAA fomentar nuevas investigaciones que permitan estudiar dichas técnicas de forma individual. La Prueba de Aptitud Académica debe ser analizada considerando el contexto en que se aplica, por lo que el método de detección debe ajustarse a esas condiciones.

Este trabajo final de graduación solo responde a las preguntas básicas sobre el FDI en la PAA; investigaciones más amplias podrían aclarar el panorama completo sobre sesgo en el instrumento utilizado para definir la admisión a la UCR.

Es de vital importancia agregar que la correcta identificación del FDI es el primer paso en la búsqueda de condiciones equitativas de evaluación (condición necesaria no suficiente de sesgo). En un segundo paso, el ítem señalado debe analizarse, no desecharse, para identificar las razones detrás de su comportamiento diferencial. Ya sea el vocabulario utilizado, ubicación de distractores, contexto de aplicación, entre otros; es decir, una línea de investigación complementaria a la identificación del FDI.

5.2.3. Efectos del tamaño de la muestra

Cada fórmula de la PAA la realizan más de 10 mil examinados, lo cual significa que los métodos de detección de FDI, en la forma que fue implementada en este estudio, consistentemente arrojaran que muchos ítems poseen FDI según el sexo, debido a los estadísticos utilizados. Para disminuir los efectos del tamaño de la muestra se recomiendan determinar otros estadísticos o métodos de detección de FDI que permitan compensar los efectos del tamaño de la muestra.

5.2.4. FDI en la formación de docentes de matemáticas

La forma adecuada de fomentar condiciones equitativas en el diseño de pruebas estandarizadas es incorporar al perfil de salida de los docentes de matemáticas la formación en FDI. Se recomienda complementar los cursos de evaluación e investigación educativa con los siguientes apartados:

5.2. RECOMENDACIONES

1. Introducción a FDI y Sesgo.
2. *Incorporar* análisis de Sesgo al diseño de ítems.
3. Incorporar análisis de FDI al estudio de la confiabilidad de pruebas e ítems.
4. Analizar los efectos de una prueba con ítems que poseen FDI.

Apéndice A

Descripción experimento en R

A.1. Código de programación

Se utiliza el lenguaje R Core Team (2016) como plataforma de programación. Se crea una función con tres entradas (cantidad de individuos, cantidad de ítems, número de iteración) que permite generar una base de datos y realizar el estudio del FDI correspondiente. Las siguientes líneas de código permiten identificar las etapas de la programación.

A.1.1. Ajuste a Modelo TRI

A cada individuo se le asigna una habilidad aleatoria (hab). De forma equivalente, cada ítem recibe un valor aleatorio de dificultad y discriminación (disc y dif).

Luego, se calcula la probabilidad de éxito del individuo dada la habilidad asignada frente a la dificultad y discriminación de los ítems. Dicho cálculo se realiza mediante el ajuste a la TRI.

```
analisisdif <- function(x,y,z){
    hab <- rnorm(x);
    disc <- rlnorm(y,0,0.3);
    dif <- rnorm(y);
    hab2 <- cbind(hab,1);
    junto <- cbind(disc, dif);
    prob <- matrix(1, x, y);

    for (i in 1:x){
        for (j in 1:(y)){
            prob[i, j] <- 0;
            prob[i, j] <- prob[i, j]+
                (1/(1+(exp(1))^(-1.7*junto[j,1]*(hab2[i,1]-junto[j,2]))))
        }
    };
```

Código A.1: Ajuste a Modelo TRI

A.1.2. Asignación de Respuestas Correctas

Dada la asignación de éxito (prob) para cada individuo por ítem, se hace una comparación frente a un valor aleatorio con distribución uniforme (comp). Si la probabilidad de acierto es mayor que el valor aleatorio generado entonces el individuo obtendría la solución correcta. En caso contrario, se asume que seleccionó la respuesta incorrecta. El resumen de datos se guarda como una matriz de ceros y unos (binom).

```
    comp <- matrix(1, x, y);

    for (i in 1:x){
        for (j in 1:(y)){
            comp[i, j] <- 0
            comp[i, j] <- comp[i, j]+runif(1,0,1)
        }
    };
```

```
binom<-matrix(1,x,y);

for (i in 1:x){
    for (j in 1:(y)){
        binom[i,j]=2;

        if (prob[i,j]>comp[i,j]){
            binom[i,j]=binom[i,j]-1
        else
            binom[i,j]<-binom[i,j]-2
        }
    }
};
```

Código A.2: Asignación de Respuestas Correctas

A.1.3. Asignación de grupo

El estudio del funcionamiento diferencial de ítem depende de la separación de la población en dos grupos. Dada la naturaleza de la base generada, simplemente se separa la base completa en dos grupos (group). El primer grupo se señala con un (0) y el segundo grupo se clasifica con un (1).

```
group<-matrix(1,x,1)

for (i in 1:(x)){
    group[i,1]=2;

    if (i<(x/2)+1)
        group[i,1]=group[i,1]-2
    else
        group[i,1]=group[i,1]-1;
};
```

Código A.3: Asinación de Grupo

A.1.4. Modificación de Ítem

Cada base aleatoria tiene 4 ítems experimentales con una diferencia significativa en su parámetro de dificultad o discriminación (FDI uniforme o no uniforme).

La modificación del parámetro depende de la asignación aleatoria original. En algunos casos, el grupo focal se verá beneficiado y en otros casos perjudicado.

```
60    juntodif <- cbind(disc, dif);
61
62    if(juntodif[1,2]<1.3)
63        juntodif[1,2] <- juntodif[1,2]+0.25
64    else
65        juntodif[1,2] <- juntodif[1,2]-0.25;
66
67    if(juntodif[2,2]<1.3)
68        juntodif[2,2] <- juntodif[2,2]+0.25
69    else
70        juntodif[2,2] <- juntodif[2,2]-0.25;
71
72    if(juntodif[y-1,2]<1.3)
73        juntodif[y-1,2] <- juntodif[y-1,2]+0.25
74    else
75        juntodif[y-1,2]<-juntodif[y-1,2]-0.25;
76
77    if(juntodif[y,2]<1.3)
78        juntodif[y,2] <- juntodif[y,2]+0.25
79    else
80        juntodif[y,2] <- juntodif[y,2]-0.25;
```

Código A.4: Ítems Experimentales

A.1.5. Asignación por grupo

Se asignan los 4 ítems experimentales al grupo focal (binomdif). La intención es que los dos grupos se diferencien únicamente en la asignación de respuestas para el ítem modificado (ítem con FDI). Se crea una matriz resumen con los datos divididos por grupo que incluye el estudio de los ítems experimentales (finaldif).

```
probdif <- matrix(1,x,y);

for (i in 1:x){
    for (j in 1:(y)){
        probdif[i,j] <- 0;
        probdif[i,j] <- probdif[i,j]+
            (1/(1+(exp(1))^(-1.7*juntodif[j,1]*(hab2[i,1]-juntodif[j,2]))))
    }
};

binomdif<-matrix(1,x,y);

for (i in 1:x){
    for (j in 1:(y)){
        binomdif[i,j]=2;

        if (probdif[i,j]>comp[i,j])
            binomdif[i,j]=binomdif[i,j]-1
        else
            binomdif[i,j] <- binomdif[i,j]-2;
    }
};

binomdif2 <- cbind(group,binomdif);

final <- rbind(binom2,1),
finaldif <- final[1:x,];

for (i in 1:(x)){
    if (i<(x/2)+1)
        finaldif[i,2]=binomdif2[i,2]
    else
        finaldif[i,2]=binom2[i,2]
}

for (i in 1:(x)){
    if (i<(x/2)+1)
        finaldif[i,3]=binomdif2[i,3]
    else
      finaldif[i,3]=binom2[i,3]
  };

for (i in 1:(x)){
    if (i<(x/2)+1)
     finaldif[i,y]=binomdif2[i,y]
    else
     finaldif[i,y]=binom2[i,y]
  };
```

```
for (i in 1:(x)){
    if (i<(x/2)+1)
     finaldif[i,y+1]=binomdif2[i,y+1]
    else
     finaldif[i,y+1]=binom2[i,y+1]
  }
```

Código A.5: Grupo de Referencia

A.1.6. Estudio del FDI

La función tiene la instrucción de realizar el estudio del FDI considerando cinco técnicas distintas: Lord, Raju, Estandarizada, MH y Logística.

```
    a <- difLord(finaldif, group=1, focal.name=1, model="2PL")
    b <- difRaju(finaldif, group=1, focal.name=1, model="2PL")
    c <- difLogistic(finaldif, group=1, focal.name=1);
    d <- difMH(finaldif, group=1, focal.name=1);
    f <- difStd(finaldif, group=1, focal.name=1);

    tabla <- cbind(0,a$LordChi,0,b$RajuZ,0,c$Logistik,0,d$MH,0 f$PDIF);

    ga <- as.data.frame(a$DIFitems);
    gb <- as.data.frame(b$DIFitems);
    gc <- as.data.frame(c$DIFitems);
    gd <- as.data.frame(d$DIFitems);
    gf <- as.data.frame(f$DIFitems);
```

Código A.6: Estudio FDI

A.1.7. Salida de Datos

La salida del programa es una matriz con los estadísticos por técnica y la confirmación del FDI. Se imprime, de forma adicional, la matriz con los datos generados aleatoriamente.

```
    write.csv(potfinal,paste(c("BaseUNIFLIT",toString(z)),collapse=""))
```

```
202|     write.csv(finaldif, paste(c("Replica", toString(z)), collapse=""))
```

Código A.7: Matriz Respaldo

A.1.8. Ejemplo de Experimento

Se toma como referencia la función descrita anteriormente para generar un estudio de simulación sobre bases aleatorias con 15000 individuos y 85 ítems.

```
205| for (i in 1:(100)){
206|     analisisdif(15000,85,i)
207| }
```

Código A.8: Ejemplo de Experimento

Referencias

AERA, APA y NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, USA: American Educational Research Association.

Alfaro-Rojas, L. y Rojas Torres, L. (2016). Desempeño de personas con la adecuación de tiempo adicional en una prueba estandarizada. *Revista Latinoamericana de Educación Inclusiva*, *10*(1), 215–227. Descargado de http://www.rinace.net/rlei/numeros/vol10-num1/art9.html

Angoff, W. (1993). Perspectives on Differential Item Functioning Metodology. En P. Holland y H. Wainer (Eds.), *Differential Item Functioning* (pp. 3–24). Lawrence Erlbaum Associates, Inc Publishers.

Barrenechea, I. (2010). Evaluaciones estandarizadas: seis reflexiones críticas. *Archivos Analíticos de Políticas Educativas*, *8*(18), 1–27. doi: http://dx.doi.org/10.14507/epaa.v18n8.2010

Bautista Sánchez, E. (2015). La evaluación mediante pruebas de gran escala en México. *Revista Iberoamericana para la Investigación y el Desarrollo*

Educativo, *5*(10), 1–15. Descargado de http://ride.org.mx/index.php/RIDE/article/view/9/41

Borsboom, D., Mellenbergh, G. y van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. doi: http://dx.doi.org/10.1037/0033-295X.111.4.1061

Camilli, G. y Shepard, L. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage Publications.

Clauser, B. y Mazor, K. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31–44.

Dorans, N. y Holland, P. (1993). DIF detection and description: Mantel-Haenszel and Standarization. En P. Holland y H. Wainer (Eds.), *Differential Item Functioning* (pp. 35–66). New York, USA: Lawrence Erlbaum Associates, Inc Publishers. Descargado de https://books.google.es/books?hl=es&lr=&id=6YAXJfswvfYC&oi=fnd&pg=PP2&dq=differential+item+functioning&ots=0vDWAogppW&sig=PS_AgFfol-H-1M9iP-ibWgXKXjc#v=onepage&q&f=false

Ellis, P. D. (2010). Effect sizes and the interpretation of research results in international business. *Journal of International Business Studies*, *1*(41), 1581–1588. doi: 10.1057/jibs.2010.39

Estado de la Educación. (2015). Desigualdades en rendimiento en el sistema educativo costarricense. En *Quinto Informe Estado de la Educación 2015* (pp. 249–294). San José, Costa Rica: Programa Estado de la Nación. Descargado de http://www.estadonacion.or.cr/educacion2015/assets/cap-5-ee-2015.pdf

Holland, P. y Thayer, D. (1985). *An alternative definition of the ETS delta scale of item difficulty (RR-85-43)*. Princeton, NJ: Educational Testing Service.

Hortensius, L. (2012). *Advanced Measurement - Logistic regression for DIF detection*. University of Minnesota, USA. Descargado de http://www.tc.umn.edu/~horte005/docs/Totalmeas.pdf

Jiménez-Alfaro K. y Morales-Fernández, E. (2010). Validez predictiva del promedio de admisión de la Universidad de Costa Rica y sus componentes. *Actualidades en Psicología*, *23*(110), 21–55. doi: http://dx.doi.org/10.15517/ap.v23i110.11

Kim, S.-H. y Cohen, A. (1995). A Comparison of Lord's Chi-Square, Raju's Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning. *Applied Measurement in Education*, *8*(4), 291–312. doi: http://dx.doi.org/10.1207/s15324818ame0804_2

Lord, F. (1977). A study of item bias, using item characteristic curve theory. En Y. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam: Swets & Zeitlinger.

Magis, D., Beland, S. y Raiche, G. (2015). difr: Collection of methods to detect dichotomous differential item functioning (dif) [Manual de software informático]. (R package version 4.6)

Magis, D., Béland, S., Tuerlinckx, F. y Boeck, P. D. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. doi: http://dx.doi.org/10.3758/BRM.42.3.847

Mainieri Hidalgo, A. (2010). *Reconstrucción teórica e histórica de los fundamentos de la Prueba de Aptitud Académica* [Informe final de investigación]. San José, Costa Rica: Programa Permanente de la Prueba de Aptitud Académica, Instituto de Investigaciones Psicológicas, Universidad de Costa Rica.

Martin, M., Mullis, I. y Hooper, M. (Eds.). (2016). *Methods and Procedures in Timss 2015*. Boston College: TIMSS & PIRLS International Study Center. Descargado de http://timssandpirls.bc.edu/publications/timss/2015-methods.html

Martínez Rizo, F. (2009). Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista Electrónica de Investigación Educativa*, *11*(2), 1–18. Descargado de http://redie.uabc.mx/redie/article/view/231/388

Mclaughlin, M. y Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psycological Measurement*, *11*(2), 161–173. Descargado de http://conservancy.umn.edu/bitstream/handle/11299/103980/v11n2p161.pdf;sequence=1

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105–118. Descargado de http://www.jstor.org/stable/1164960 doi: 10.2307/1164960

MEP. (2012). *Programas de Estudio de Matemáticas*. San José, Costa Rica: Ministerio de Educación Pública.

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5–11. Descargado de http://www.jstor.org/stable/1175249

Montero Rojas, E. (2013). Referentes conceptuales y metodológicos sobre la noción moderna de validez de instrumentos de medición: implicaciones para el caso de personas con necesidades educativas especiales. *Actualidades en Psicología*, *27*(114), 113–128. doi: http://dx.doi.org/10.15517/ap.v27i114.7900

Montero-Rojas, E., Castelain, T., Moreira Mora, T., Alfaro-Rojas, L., Cerdas-Núñez, D., García-Segura, A. y Roldán Villalobos, M. (2013). Evidencias iniciales de validez de criterio de los resultados de una Prueba de razonamiento con figuras para la selección de estudiantes indígenas para la Universidad de Costa Rica y el Instituto Tecnológico de Costa Rica. *Revista Educación*, *37*(2), 103–117. doi: http://dx.doi.org/10.15517/revedu.v37i2.12928

Moreira Mora, T. (2008). Funcionamiento diferencial del ítem en pruebas de matemática para educación media. *Actualidades en Psicología*, *22*(109), 91–113. doi: http://dx.doi.org/10.15517/ap.v22i109.16

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo*, *31*(1), 57–66. Descargado de http://www.papelesdelpsicologo.es/pdf/1796.pdf

Nunnally, J. y Bernstein, I. (1995). *Teoría psicométrica* (3.ª ed., J. Velázquez Arellano, Traduc.). McGraw-Hill.

OECD. (2016). *PISA 2015 Results* (Vol. 1). París, Francia: OECD Publishing. doi: http://dx.doi.org/10.1787/9789264266490-en

R Core Team. (2016). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de http://www.R-project.org/

Raju, N. (1988). The area between two item characteristic curves. *Psycometrika*, *53*(4), 495–502. doi: http://dx.doi.org/10.1007/BF02294403

Raju, N. (1990). Determining the significance of estimated signed and unsigned

areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197–207. doi: http://dx.doi.org/10.1177/014662169001400208

Shepard, L. (1982). Definitions of bias. En R. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9–30). Baltimore: Johns Hopkins University Press.

Smith-Castro, V. (Ed.). (2014). *Compendio de Instrumentos de Medición IIP-2014* (Cuaderno Metodológico n.º 6). Universidad de Costa Rica, Costa Rica: Instituto de Investigaciones Psicológicas. Descargado de http://iip.ucr.ac.cr/sites/default/files/cuadernosmetodologicos/cuamet6.PDF

Wasserstein, R. L. y Lazar, N. A. (2016). The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, *10*(2), 129–133. doi: 10.1080/00031305.2016.1154108

Zumbo, B. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF)*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Descargado de http://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf

Zumbo, B. y Hubley, A. (1998). Differential item functioning (DIF) analysis of a synthetic CFAT. En *Technical note 98-4 , Personnel Research Team.* Ottawa ON: Department of National Defense.