Funcionamiento Diferencial del Ítem: una abordaje desde la perspectiva bibliométrica, experimental y empírica

Ángela Berrío Beltrán

TESIS DOCTORAL

FUNCIONAMIENTO DIFERENCIAL DEL ITEM: Una aproximación Bibliométrica, Experimental y Empírica.

ÁNGELA BERRÍO BELTRÁN

UNIVERSIDAD DE BARCELONA
2022

Tesis Doctoral

Funcionamiento Diferencial del Ítem: un abordaje desde la perspectiva bibliométrica, experimental y empírica

Autora: Ángela Berrío Beltrán

Directora: Juana Gómez Benito

UNIVERSITAT DE BARCELONA

Funcionamiento Diferencial del Ítem: un abordaje desde la perspectiva bibliométrica, experimental y empírica

Memoria presentada para optar al grado de doctor por la Universidad de Barcelona

Doctorado en Psicología Clínica y de la Salud

Autora: Ángela Berrío Beltrán

Directora: Juana Gómez Benito

Barcelona, Enero de 2022

Departamento de Psicología Social y Psicología Cuantitativa

Facultad de Psicología

Universidad de Barcelona

UNIVERSITAT DE BARCELONA

Agradecimientos

Finalizar esta tesis doctoral ha implicado un trabajo arduo en el que he aprendido el valor de la persistencia, la tolerancia a la frustración, el trabajo colaborativo y el soporte emocional. Pero esta tesis no es sólo el fruto de mi trabajo individual, sino también del esfuerzo, participación, colaboración y apoyo de diferentes personas a las que quisiera extender mi más sincero agradecimiento.

Quiero agradecer a tod@s los que han hecho parte de este camino porque sin sus palabras, abrazos, consuelo y confianza no hubiese podido enfrentar las dificultades y retos que me ha supuesto culminar la tesis.

Un especial agradecimiento a la Profesora Juana por su gran paciencia ante mis innumerables preguntas y dudas, por su guía y por la confianza que depositó en mí. Ha sido un verdadero placer trabajar bajo su dirección, no sólo por el aprendizaje teórico y del saber hacer, sino también por todo el apoyo recibido. Porque cuando la angustia me nublaba y no podía ver claro el panorama, siempre tuvo la palabra certera para mostrarme el mejor camino y para devolverme la confianza. Gracias por insistirme cuando fue necesario, por exigirme cuando tenía que hacerse y por acompañarme cuando lo necesité.

Quiero agradecer a mis compañeros de doctorado. A Viviana por los primeros encuentros "para tomarnos un Té-Café" que siempre terminaban siendo una sesión catártica. A Gomaa mi compañero de despacho siempre dispuesto a ayudar, una persona trabajadora con gran calidad humana. A Annie que con su alegría y amistad me acompañó en los últimos tramos de la tesis. A Ana, Cristina, Belén, Laura y Estefanía por las comidas en compañía llenas de risa e historias por contar.

A Nidia Herrera y Georgina por sus aportes en los estudios 2 y 3, respectivamente, por sus comentarios siempre tan certeros y el entusiasmo que le dedican al trabajo día a día, ha sido muy enriquecedor contar con la guía y aprender de personas que como ustedes se apasionan con lo que hacen.

A Erika por apoyarme y animarme en los momentos difíciles, gracias por tu insistencia y fortaleza, y por ser parte y testigo de mi historia. Gracias por tu participación como coautora del primer estudio, realmente siempre trabajamos muy bien juntas, especialmente porque sé el esfuerzo que te debió suponer colaborar en ese artículo.

A Angie, Carito, Flora y Bárbara con quienes intercambié otras visiones del mundo y quienes con el pasar de los años se convirtieron en fuente de soporte y gran amistad.

A los miembros del laboratorio de Psicometría de la Universidad Nacional de Colombia.

A mi familia, por acompañarme en la distancia y ser parte de mi caminar.
A la tía Sofía por animarme a conseguir mis metas y trazarme nuevos propósitos, y aunque lamentablemente ya no estarás para compartir este momento, siempre me acuerdo de tu: "Querer es poder".
A Cristty, mi confidente, hermana y amiga, siempre conmigo en la distancia, siempre aquí ayudándome a centrarme en lo fundamental, a reconocer el miedo y la ansiedad que me nublaban.
A Paola, Germán y Catalina; hermanitos, siempre dándome ánimo para seguir adelante, apoyándome en mis decisiones y compartiendo mi felicidad.
A mis papis por todo y por tanto, por su amor y sus miles de enseñanzas, por ser cómplices de mis sueños y vigías de mi camino.
A Paco, gracias por tu comprensión y amor, por escuchar mis dudas (que no son pocas), por creer en mí y darme seguridad, por darme motivos para sonreir en los días difíciles y animarme a continuar cuando me sentía perdida. Gracias porque junto a ti el camino se llena de historias, de color y música.
A Cándida, Charo, Montse, Lucía, Juan Antonio, Rosario y Antonia, gracias por hacerme sentir parte de una familia en estos últimos años lejos de la mia.

Índice de contenido

Índice de Figuras

Índice de Tablas

Lista de Acrónimos

AFC	Análisis Factorial Confirmatorio
APA	Asociación Americana de Psicología – American Pychological Asociation
CAT	Test adaptativo informatizado
CATSIB	Test de Sesgo Simultáneo para Test Adaptativos Informatizados
CCC	Curva Característica de la Categoría de respuesta
CCI	Curva Característica del Ítem
CDM	Modelo de Diagnóstico Cognitivo
DC	Tasas de Detección Correcta
DIF	Funcionamiento Diferencial del Ítem – Differential Item Functioning
DIFT-TRI	Funcionamiento Diferencial del Ítem y del Test bajo la Teoría de Respuesta al ítem
DTF	Funcionamiento Diferencial del Test – Differential Test Functioning
EMD	Método de igualdad de la dificultad media – Equal Mean Difficulty
ERIC	Centro de Información en Recursos de Educación - Education Resources Information Center
ETS	Servicio de Evaluación Educativa – Educational Testing Service
FP	Tasas de Falsos Positivos
GEIMAC	Grupo de Estudios de Invarianza de la Medida y Análisis del Cambio
ICFES	Instituto Colombiano para la Evaluación de la Educación
IFT	Árboles Enfocados en los Ítems – Item-Focused Trees
ITC	Comisión Internacional de Tests – International Test Commission
MH	Mantel-Haenszel
MIDD	Diferencia de la dificultad media del ítem – Mean Item Difficulty Difference
MIMIC	Múltiple Indicador Múltiple Causa
PCM	Modelo de Crédito Parcial – Partial Credit Model
PCM-IFT	Árboles Enfocados en los Ítems bajo el Modelo de Crédito Parcial – Partial Credit Model-Item-Focused Trees
PISA	Programa para la evaluación internacional de estudiantes – Programme for International Student Assessment
POLYSIBTEST	Test de Sesgo Simultáneo para ítems Politómicos
PRISMA	Preferred reporting items for systematic review and meta-

	analyses
Rasch-DIF	Procedimiento de detección de DIF basado en el Modelo Rasch
Rasch-Trees	Árboles basados en Rasch
RL	Regresión Logística
SIBTEST	Test de Sesgo Simultáneo
TCT	Teoría Clásica del Test
TIMMS	Estudio internacional de tendencias en matemáticas y ciencias – Trends in International Mathematics and Science Study
Tree-PCM	Árboles basados en Rasch bajo el Modelo de Crédito Parcial
TRI	Teoría de Respuesta al Ítem
WHODAS	Cuestionario para la Evaluación de la Discapacidad de la Organización Mundial de la Salud – World Health Organization Disability Assessment Schedule
WoS	Web of Science
1PL	Un parámetro logístico
3PL	Tres parámetros logísticos

Resumen

La detección del funcionamiento diferencial del ítem (DIF) en un test constituye una amenaza a la validez de las inferencias que de sus resultados se puedan extraer. Ignorar su presencia o desconocerla puede traer implicaciones o consecuencias sociales relevantes para los evaluados que pertenecen a grupos específicos de la población. Es por ello que hoy en día, el análisis del DIF se ha convertido en un aspecto que requiere ser valorado a la hora de desarrollar nuevos tests o de adaptar y traducir los ya existentes. A este respecto en las guías de la Comisión Internacional de Tests (ITC) sobre la traducción y adaptación de test, el uso del test, la evaluación a gran escala de poblaciones lingüísticamente diversas, y las pruebas realizadas por internet e informatizadas, se pueden encontrar referencias para que los usuarios tengan presente o valoren la presencia de DIF.

Después de casi 4 décadas de desarrollos, los usuarios cuentan con un gran número de métodos que permiten analizar la presencia o no de DIF. Éstos pueden estar basados en distintos modelos teóricos de la medición y/o responder a necesidades propias de los diferentes ámbitos de evaluación. En este contexto, con esta tesis se propuso analizar y proporcionar información sobre las tendencias y desarrollos recientes en métodos de detección de DIF teniendo en cuenta los principales ámbitos de aplicación. Para ello se diseñaron tres estudios: un estudio de revisión sistemática que incluyó algunos análisis bibliométricos, un estudio de tipo experimental con datos simulados y un estudio de tipo empírico.

El estudio de revisión sistemática encontró once tópicos de investigación en los que se ha centrado el interés de los investigadores respecto de los métodos de detección de DIF, señaló los valores más frecuentes que toman algunas variables en los estudios con datos simulados, mostró algunas tendencias respecto a los valores de dichas variables y presentó los métodos estudiados con mayor frecuencia y aquellos que se han estudiado más recientemente. Debido a que el

método de detección de DIF basado en el modelo Rasch fue uno de los más estudiados y presenta desarrollos recientes, se diseñaron dos estudios para analizar algunos de sus procedimientos.

Por un lado, se analizó el funcionamiento del procedimiento diferencia del parámetro de dificultad en un estudio con datos simulados bajo condiciones similares a las pruebas nacionales que evalúan la calidad de la educación y dan acceso a la universidad en Colombia. Se analizó el efecto de: la razón de tamaños, datos simulados bajo modelos de un parámetro logístico (1PL) o tres parámetros logísticos (3PL; cuando existe un peso importante del azar en los aciertos y por tanto los datos están desajustados al modelo Rasch) y diferencias en la distribución del nivel de atributo, sobre las tasas de detección correcta (DC) y tasas de falsos positivos (FP). Los resultados respecto de las tasas de falsos positivos mostraron que el funcionamiento del procedimiento se ve afectado cuando el modelo de simulación de los datos fue 3PL, mientras que las tasas de detección correcta se vieron más afectadas por la razón de tamaños y por las diferencias en la distribución del nivel de atributo. Se discutieron los resultados y las implicaciones prácticas derivadas de ellos.

Por otro lado, se aplicaron dos procedimientos de árboles de Rasch (propuestos más recientemente) para la detección de DIF en la escala WHODAS 2.0 en personas con esquizofrenia. Se incluyeron las respuestas de 280 participantes y se analizaron variables de tipo demográfico y clínico. Los resultados indicaron que sólo uno de los procedimientos detectó la presencia de DIF en un único ítem debido a la variable edad. Se demostró la fortaleza de estos procedimientos al no establecer a priori los puntos de corte de variables continuas y se discutieron las diferencias observadas, así como también la información complementaria que proporcionan. Aunque la evidencia aportada en este estudio está a favor de la validez de la escala, también se identificaron algunas debilidades de esta medida al ser aplicada a personas con esquizofrenia.

Abstract

In general terms, the detection of differential item functioning (DIF) in a test constitutes a threat to the validity of the inferences that can be drawn from its results. Ignoring their presence or not knowing it can bring relevant social implications or consequences for those evaluated who belong to specific groups of the population. That is why today, DIF analysis has become an aspect that needs to be assessed when developing new tests or adapting and translating existing ones. For this reason, the International Test Commission (ITC) contains guidelines for translating and adapting tests, on test use, for large-scale assessment of linguistically and culturally-diverse populations, and on computer-based and internet-delivered testing, references or suggestions, so that users are aware of, or able to value the presence of DIF.

After almost four decades of development, users have a large number of methods that allow the presence or absence of DIF to be analyzed. These can be based on different theoretical models of measurement and/or respond to the needs of the different evaluation fields. In this context, this thesis proposed to analyze and provide information on recent trends and developments in DIF detection methods, taking into account the main fields of application. For this purpose, three studies were designed: a systematic review study that included some bibliometric analysis, an experimental study with simulated data, and an empirical study.

The systematic review study found eleven research topics in which the interest of researchers has focused on DIF detection methods. This review pointed out the most frequent values taken by certain variables in studies with simulated data, showed some trends regarding the values of these variables and presented the most frequently studied methods and those that have been studied more recently. Due to the fact that the DIF detection method based on the Rasch model was one

of the most studied and presents recent developments, two studies were designed to analyze some of its procedures.

On the one hand, the functioning of the Difficulty Parameter Differences procedure was analyzed in a study with simulated data under similar conditions to the national tests that evaluate the quality of education and give access to the university in Colombia. The effect of: the sample size ratio, data simulated under 1PL or 3PL models (when there is a significant influence of chance in the hits), and differences in the distribution of the attribute level, on the rates of correct detection and rates of false positives. In general terms, the results regarding the false positive rates showed that the procedure is affected when the simulation model of the data was 3PL, while the correct detection rates were more affected by the sample size ratio and by the differences in the distribution of the attribute level. The results and the practical implications derived from them were discussed.

On the other hand, two Rasch trees procedures (more recently proposed) were applied for the detection of DIF on the WHODAS 2.0 scale in people with schizophrenia. The responses of 280 participants were included and demographic and clinical variables were analyzed. The results indicated that only one of the procedures detected the presence of DIF in a single item due to age. The strength of these procedures was shown by not establishing the cut-off points for continuous variables *a priori*, furthermore the observed differences were discussed, as well as the complementary information that the procedures provided. Although the evidence provided in this study is in favor of the validity of the scale, some weaknesses were identified when the WHODAS 2.0 scale is applied to people with schizophrenia.

1. Introducción

La medición de un atributo con el fin de estimar el nivel que de éste presenta una persona o un grupo de ellas, es una tarea que ha sido de gran interés en las ciencias sociales y de la salud. Dada la complejidad de los constructos o atributos que se intentan medir (p.e. Inteligencia, Calidad de vida, Aptitudes, Actitudes, Personalidad, Competencias, Conocimientos, entre otros), se han diseñado diversos instrumentos para dar cuenta de algunas facetas de ellos o de aspectos específicos. Entre estos instrumentos uno de los más comúnmente usados son los tests psicométricos (pruebas, inventarios, cuestionarios y escalas) debido a su fácil acceso y aplicación en diversos ámbitos de evaluación.

Todo test psicométrico debe permitir que las conclusiones inferidas de sus resultados sean válidas, es decir, que los resultados no se encuentren influidos por variables externas y espurias al atributo que se está midiendo. Por ejemplo, si se aplica un cuestionario de calidad de vida a un paciente, el resultado obtenido por éste debe permitir concluir claramente sobre el nivel de calidad de vida y no ha de encontrarse afectado por otras variables como género, etnia o capacidad de lectura, que afecten los resultados del test y por tanto conduzcan a conclusiones erróneas sobre el nivel de atributo.

El mayor auge e interés por temas relacionados con la validez comenzó a aparecer hacia mitad del siglo XX cuando la relevancia de temas como equidad e igualdad estaban en un punto álgido. A este respecto, la demostración empírica de que algunos ítems de tests de inteligencia presentaban diferencias culturales (Eells et al., 1951), sentó las bases sobre la discusión de la presencia de sesgo en los ítems. Posteriormente, durante la década de los 60s y 70s se dio inicio a un movimiento en contra del uso de los tests, señalándolos como instrumentos sesgados e inequitativos, puesto que sus resultados presentaban diferencias entre poblaciones pertenecientes a diversos grupos étnicos con claras desventajas para grupos de

minoría étnica. Después de dos décadas de duras críticas sobre el uso del test, en la década de los 80s comenzaron a surgir una serie de aclaraciones teóricas y técnicas que reavivaron el interés en su uso y permitieron la consolidación de la psicometría como técnica necesaria para su adecuado diseño, desarrollo y evaluación.

Las nuevas aproximaciones a la definición de la validez como un juicio sobre la adecuación de las inferencias que se extraen de las puntuaciones de un test (American Educational Research Association [AERA] et al., 1985; Messick, 1980), la diferenciación entre la connotación ética de sesgo entendida como injusticia y la connotación estadística del sesgo y finalmente, la precisión propuesta por Holland y Thayer (1988) entre sesgo y funcionamiento diferencial del ítem, permitieron llegar a la conclusión de que no todas las diferencias entre poblaciones corresponden a un sesgo del instrumento y que dichas diferencias pueden ser debidas a aspectos relevantes del constructo medido o a defectos del test.

Si las diferencias no se deben a aspectos relevantes del constructo se analiza estadísticamente si se trata de un defecto del test y si es así, el análisis del sesgo corresponderá a un juicio informado que tendrá en cuenta además de los resultados estadísticos, el objetivo del test y la información cultural o social de los grupos de examinados, entre otros aspectos.

A partir de estas aclaraciones, las diferencias entre poblaciones recobran su papel fundamental en toda evaluación y ya no se considera como un aspecto que va en detrimento de la medida per se, sino como un aspecto más a tener en cuenta en el análisis de la calidad de las medidas. Durante dos décadas, las discusiones sobre el sesgo en pruebas y la equidad social habían ensombrecido el uso de los tests. Ahora, obtener resultados diferentes entre grupos de examinados planteaba una situación en la que se debía constatar si dichas diferencias eran debidas a diferencias reales en el atributo medido, o si por el contrario estábamos ante el posible funcionamiento diferencial de los ítems o del test.

Así, después de las aclaraciones introducidas por Holland y Thayer (1988), encontrar diferencias entre un grupo de minorías étnicas y otro mayoritario, implica que en un primer momento se evalúe si existen deficiencias en la validez del test o si existen diferencias reales en el atributo medido. Una forma de aproximarse a ello es analizar las diferencias entre los grupos en cada uno de los ítems que conforman el test. Es de suponer que dos examinados con el mismo nivel de atributo, aunque pertenezcan a grupos diferentes, deben presentar un comportamiento igual o similar ante cada uno de los ítems. Si el comportamiento es diferente, el ítem funciona diferencialmente (DIF por sus siglas en inglés: Differential Item Functioning), pero será necesario un análisis más profundo sobre las causas de DIF para poder afirmar que el ítem está sesgado.

Otro concepto que también ha estado en la base de las discusiones iniciales sobre el uso del test hace referencia a la justicia/injusticia. A este respecto, la AERA et al. (2014) aportaron mayor claridad al presentar cuatro connotaciones que puede tomar dicho concepto. La primera se refiere a la igualdad o equidad en el tratamiento que reciben los examinados para tener la misma oportunidad de mostrar su nivel de atributo. Esto se logra con un proceso cuidadoso de estandarización del test, unas condiciones de administración y unas instrucciones de calificación adecuadas. La segunda hace referencia a una aproximación estadística para la detección de la carencia de sesgo en la medición ya sea a través del análisis de DIF, del funcionamiento diferencial del test (DTF por sus siglas en inglés: Differential Test Functioning), de la falta de consistencia del constructo entre subgrupos y del sesgo predictivo. La tercera está relacionada con una oportunidad comparable de acceso al constructo medido y las covariables relacionadas con su ejecución por parte de los examinados, de tal manera que ellos (población objetivo de evaluación) puedan mostrar su nivel en el constructo sin que algunas de las características personales (edad, idioma, discapacidad, raza, etc.) y que son irrelevantes para el atributo produzcan ventajas o desventajas en exceso. Finalmente, la justicia se relaciona con la validez de la interpretación de la

puntuación individual obtenida de acuerdo con los usos previstos de ella. Es relevante tener en cuenta las características individuales y cómo dichas características interactúan con rasgos contextuales de la situación de evaluación al momento de extraer inferencias sobre la habilidad o destreza del examinado.

En este punto es claro que no se trata de que un instrumento de medida sea justo o injusto, sino que existen unas características tanto del test como del proceso de evaluación y de las inferencias que se extraen de la puntuación, que permiten asegurar un tratamiento equiparable para diferentes grupos de examinados durante la evaluación y la interpretación de las puntuaciones.

Teniendo en cuenta lo anterior, el análisis de DIF comenzó a centrar la atención de los investigadores y de los desarrolladores de tests, de tal manera que ha llegado a convertirse en un paso cada vez más indispensable cuando se trata de valorar la calidad psicométrica de los tests, o cuando se diseñan o adaptan éstos.

Adicionalmente, la importancia del estudio de DIF se ha incrementado debido a la globalización que ha permitido tanto las evaluaciones internacionales, por ejemplo en materia de educación, como el acceso a tests diseñados y elaborados en diversos ámbitos culturales y sociales. Pero también debido a la extensión que ha alcanzado la aplicación de tests y el impacto que puede generar para determinados examinados el resultado obtenido, como en pruebas nacionales para el ingreso a la universidad, en pruebas que conforman procesos de selección para acceder a determinados empleos públicos o privados, en tests de clasificación para determinar el acceso a tratamientos o ayudas sociales, etc.

Teniendo en cuenta que el interés por el análisis de DIF continua vigente hoy en día y el gran desarrollo en cuanto a métodos de detección de los últimos años, se consideró relevante realizar una tesis doctoral por compendio de artículos sobre esta temática, focalizando el interés en los métodos de detección de DIF a partir de una perspectiva bibliométrica, experimental y empírica.

1.1. Aspectos conceptuales del funcionamiento diferencial del ítem.

1.1.1. Definición de DIF, impacto y sesgo

1.1.1.1 DIF

En primera instancia vale la pena recalcar que DIF es un indicador de validez al nivel del ítem y que, como se mencionó previamente, se encuentra muy relacionado con el concepto de sesgo, pero también con el concepto de impacto. DIF se define como una violación de la invarianza en un ítem que ocurre cuando la probabilidad de respuesta a dicho ítem es diferente para dos examinados que pertenecen a subgrupos distintos y que tienen el mismo nivel de atributo o que están igualados por el puntaje total del test (Camilli, 1993; Clauser & Mazor, 1998; Dorans & Holland, 1993; Mapuranga et al., 2008; Zumbo, 1999).

Así, el elemento característico y diferenciador de la presencia de DIF es la igualación del nivel de atributo de los examinados que se comparan, por lo que se trata de un análisis de diferencias condicionado. Otro elemento característico hace referencia a la formación de subgrupos dentro de la población objeto de evaluación. Dichos subgrupos se conforman de acuerdo con una variable externa a la medición realizada y que se cree genera diferencias entre las puntuaciones. Por lo general se trata de variables sociodemográficas (género, edad, etnia, idioma, nivel socioeconómico, etc.), pero también pueden ser variables de tipo clínico (estado de salud, nivel de bienestar, funcionamiento, estadio de una enfermedad, etc.) o variables de tipo cognitivo, entre otras.

Las variables más empleadas han sido las sociodemográficas probablemente debido a que el nacimiento del mismo concepto viene dado por la pregunta sobre la equidad en las pruebas psicométricas y la evaluación igualitaria en población considerada como de minoría étnica, racial o social. Sin embargo, debe tenerse en

cuenta que la selección de los subgrupos a comparar ha de obedecer a consideraciones teóricas del constructo a medir (Teresi et al., 2008). Por ejemplo, si un test mide la magnitud de dolor agudo en una fase inicial de una enfermedad específica, tendría poco sentido establecer subgrupos por tipo de dolor o incluso por enfermedades, pero sería interesante analizar DIF para subgrupos definidos por cultura y género.

Otro elemento a resaltar es que la presencia de DIF se considera un error sistemático que produce una respuesta diferente entre subgrupos, condicionada al nivel de atributo y que se vincula directamente con la carencia de validez y más específicamente con la ausencia de invarianza. Es de esperar que un ítem mida de igual manera el aspecto del constructo para el que se ha diseñado entre diferentes subpoblaciones, por tanto, las propiedades de una medida deben ser independientes de las características de las personas a las que se aplica, pero no de aquellas que hacen parte del objeto de medida (Millsap, 2007). Así se reconoce que todo proceso de medida está constituído tanto por características, variables o dimensiones objetivo como por otras secundarias, espurias, poco o nada relacionadas con el constructo a medir. La presencia de DIF dependerá de la injerencia que este segundo grupo pueda tener en la respuesta a los ítems. En este sentido, de Ayala (2009) resaltó la relevancia de la detección de DIF puesto que puede indicar la presencia de sesgo de tal forma que la varianza de un constructo irrelevante para la medición tiene un efecto diferencial en los examinados que pertenecen a diferentes subgrupos.

En términos formales y adaptando las definiciones de autores como Lord (1980), Millsap y Meredith (1992) y Osterlind y Everson (2009) la presencia de DIF puede expresarse como:

$$P(Y|\theta, G = 1) \neq P(Y|\theta, G = 2),$$

donde P(•) representa la probabilidad, Y corresponde a la variable observable o la respuesta dada al ítem, θ se refiere al atributo o constructo que se está midiendo, que no es directamente observable (variable latente) y G representa al subgrupo. Así, la probabilidad de respuesta en un ítem condicionada al nivel de atributo es diferente para cada uno de los grupos que se comparan y por tanto no se puede afirmar que Y y G son independientes. Lo contrario, es decir la igualdad entre los dos términos de la expresión anterior, correspondería a la invarianza del ítem.

1.1.1.2 Impacto

Por su parte, el concepto de impacto refleja una diferencia real entre subgrupos debida al nivel de atributo o lo que es lo mismo, corresponde a las diferencias válidas en la probabilidad de acertar o adoptar una opción de respuesta como consecuencia de una magnitud de atributo diferencial (Clauser & Mazor, 1998; Fidalgo et al., 1999; Herrera et al., 2005; Hidalgo & Gómez-Benito, 2003). En este caso dos examinados de diferentes subgrupos presentan una probabilidad de respuesta diferente en un ítem determinado, debido a que su nivel de atributo es distinto. En una evaluación educativa, por ejemplo, se esperaría que, a mayor nivel de atributo del examinado, su respuesta a un ítem sea un acierto o sea una respuesta correcta; mientras que lo contrario se esperaría para un examinado con menor nivel de atributo.

Por lo general, el impacto al nivel del ítem se reporta como una diferencia en la respuesta a dicho ítem debida a la diferencia en el nivel de atributo que pretende medir el ítem, mientras que el impacto en el test se define como una diferencia entre subgrupos respecto del desempeño en el test (puntuación total o estimación del nivel de atributo) causada por diferencias válidas (reales) en el atributo que se está midiendo (Ackerman, 1992). La presencia de impacto sugiere que, debido a la preexistencia de diferencias en las puntuaciones, se requerirá generar normas de puntuación diferenciales.

1.1.1.3 Sesgo

Finalmente, el concepto de sesgo se refiere a la búsqueda e identificación de las causas o razones por las cuales los ítems funcionan diferencialmente. Tal como han afirmado Padilla y Benítez (2017) el estudio del sesgo en los ítems además de ayudar a entender los resultados obtenidos en el análisis de DIF, también ayuda a conocer por qué ocurre este sesgo incorporando variables contextuales y personales como parte del modelo explicativo. Sin embargo, a pesar de las muchas décadas de desarrollo en DIF, el estudio del origen o de la razón por la cual se presenta ha sido de menor trascendencia para los investigadores.

La menor frecuencia de estudios sobre sesgo, pese a la importancia empírica que conllevan, puede haberse producido porque la gran mayoría de métodos clásicos, que han sido los más estudiados, están centrados únicamente en detectar la existencia o no del funcionamiento diferencial y no en proporcionar información referente a las causas, fuentes o naturaleza de DIF, privilegiando por tanto, el primer tipo de estudios y no la determinación de sus causas. Así mismo, la reciente inclusión del análisis de DIF como una técnica que recoge evidencia de validez de la estructura interna de un test que se encuentra en los estándares de pruebas (AERA et al., 2014) tampoco proporciona una clara directiva que centre la atención de los investigadores en el estudio de sesgo versus el estudio de detección de DIF.

Como ya se ha mencionado previamente, el estudio de sesgo en ítems debe incluir, además de la información estadística de la detección de DIF, un estudio riguroso de sus posibles causas o naturaleza. Una de las estrategias más frecuente ha sido emplear técnicas cualitativas para conocer el juicio de expertos, mediante el diseño de un plan riguroso que permita obtener información precisa respecto del funcionamiento diferencial y cómo éste puede estar relacionado con la presencia de otras variables o dimensiones irrelevantes. Este plan puede incluir una serie de preguntas que se formulan a los expertos orientadas a recabar información sobre la posible existencia de características del ítem que hacen intervenir otras variables

o dimensiones distintas para responder, aspectos de la redacción del ítem que pueden estar en el origen del funcionamiento diferencial o incluso aspectos relacionados con el contenido temático del ítem que puedan actuar como una fuente de DIF.

1.1.2. Conceptos propios del análisis de DIF

Antes de iniciar un análisis de DIF se deben definir varios aspectos que se consideran relevantes a la hora de seleccionar el método de detección que se empleará. Entre estos aspectos cabe destacar principalmente, los grupos que se van a comparar, el tipo de DIF que se pretende analizar o que se espera encontrar y el tipo de ítems que serán sujeto del análisis.

1.1.2.1 Grupos a comparar

Como ya se sugirió anteriormente, el análisis que se lleva a cabo para la detección de DIF requiere de la comparación entre grupos y clásicamente esta comparación se realiza entre dos grupos. Un grupo que recibe el nombre de "referencia" y otro que recibe el nombre de "focal".

El grupo de referencia representa a la población general o aquella que se piensa es aventajada, mientras que el grupo focal representa a la población de interés sobre la que se espera que se produzcan efectos adversos. Se debe tener en cuenta que el grupo focal no siempre corresponde a una población subrepresentada, por lo menos en términos de su proporción demográfica. Estos dos grupos se pueden definir por la pertenencia o presencia de determinados valores de variables que pueden ser de tipo sociodemográfico (género, etnia, idioma, raza, etc.), cognitivo (atención, memoria, lenguaje) o clínico (variables de tipo médico o psicológicas).

1.1.2.2 Tipo de DIF

Las diferencias detectadas entre los grupos cuando se analiza la presencia de DIF indican que la respuesta a un ítem está relacionada con la pertenencia a un grupo

determinado y pueden ser uniformes o no uniformes y unidireccionales o bidireccionales. De acuerdo con Hessen (2003) el DIF uniforme se presenta cuando la relación entre la respuesta a un ítem y la pertenencia a un grupo es constante en todos los valores del nivel de atributo medido. Mientras que el DIF no uniforme se presenta cuando dicha relación no es constante a lo largo del continuo del atributo medido.

Una aproximación gráfica de DIF puede hacerse a través de la comparación de la curva característica del ítem (CCI) de cada grupo. Esta aproximación es ideal puesto que relaciona la probabilidad de respuesta en un ítem con el nivel de atributo para cada grupo y por tanto permite identificar el tipo de DIF que se presenta.

En la figura 1.1 se han trazado las curvas características de un ítem para cada uno de los grupos (referencia y focal). Claramente se observa que las CCI's presentan una probabilidad de respuesta diferente que es constante en todos los valores del nivel de atributo. Por ello se puede afirmar que el ítem presenta DIF uniforme.

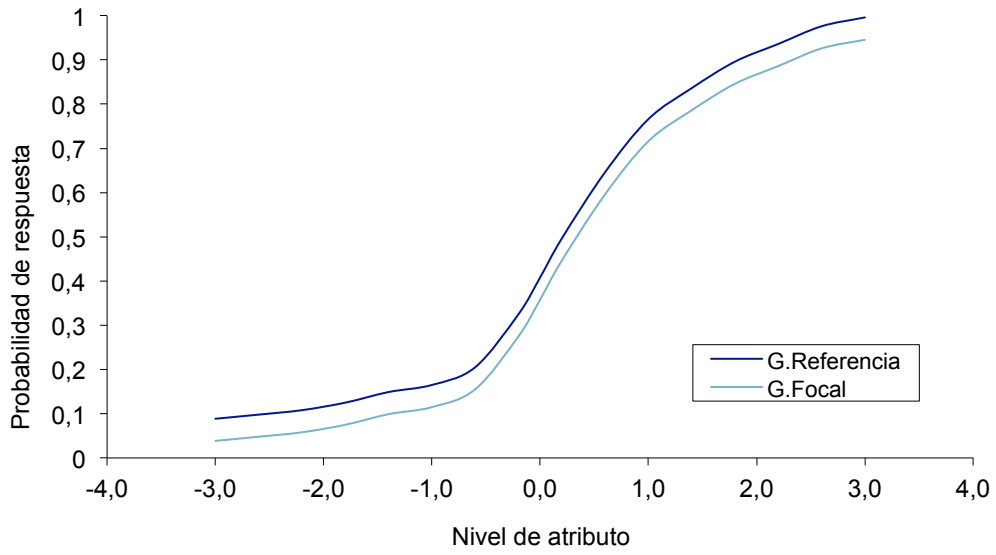


Figura 1.1. Curva característica de un ítem con DIF uniforme

Mientras que en la figura 1.2 se presentan las CCI's de un ítem que claramente tiene DIF no uniforme. En este caso, las diferencias en la probabilidad de respuesta para los grupos varían en distintos puntos del nivel de atributo. Nótese que tanto en los valores bajos del nivel de atributo como en los altos las diferencias son menores que en los niveles intermedios.

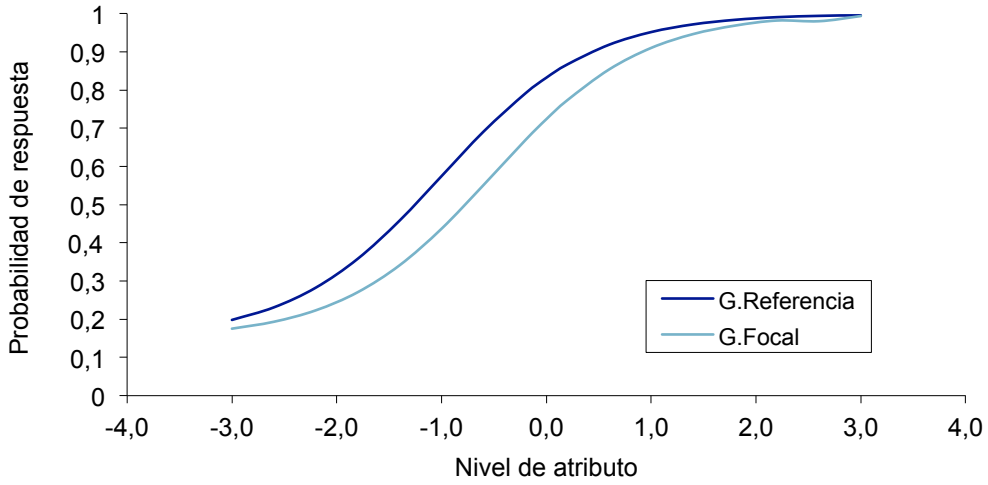


Figura 1.2. Curva Característica de un ítem con DIF no uniforme

Otro tipo de diferencias que pueden presentarse se refieren a la dirección de la relación y a este respecto se presenta DIF unidireccional o bidireccional (Shealy & Stout, 1993). El primero se observa cuando las CCI's de los grupos no se intersectan, tal como ocurre en las figuras 1.1 y 1.2. En ellas, para todos los niveles de atributo, la probabilidad de respuesta del grupo focal es siempre inferior a la del grupo de referencia y por tanto no hay intersección entre las CCI's. El DIF bidireccional, por su parte, ocurre cuando las CCI's se intersectan en al menos un punto del nivel de atributo. En la Figura 1.3 se observa que en los niveles bajos de atributo el grupo focal presenta una mayor probabilidad de respuesta, pero en los niveles más altos esta dirección cambia a favor del grupo de referencia.

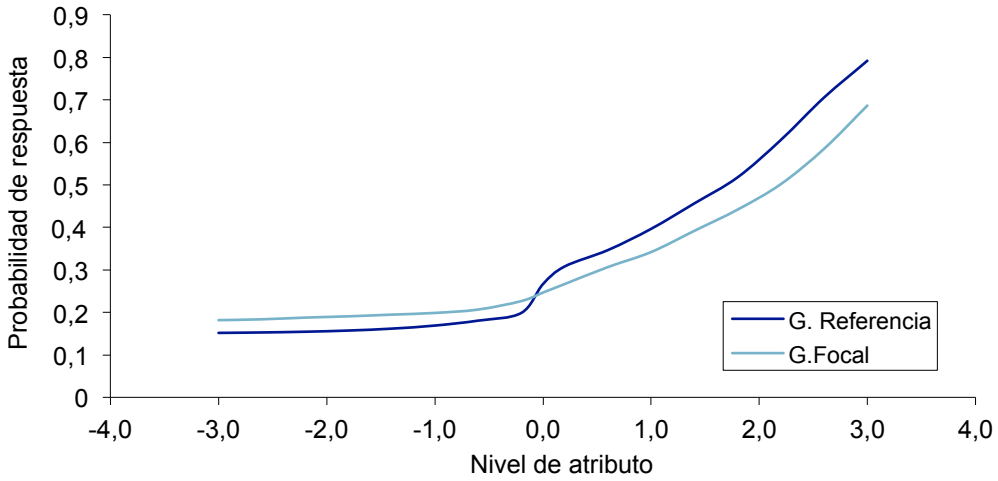


Figura 1.3. Curva Característica de un ítem con DIF bidireccional

Como ya se ha anticipado al hablar de DIF unidireccional, los distintos tipos de DIF presentan alguna relación. De acuerdo con Hessen (2003) DIF uniforme siempre será unidireccional, mientras que DIF unidireccional puede presentarse junto con DIF no uniforme. Así, DIF uniforme no puede ser además DIF bidireccional porque son conceptos mutuamente excluyentes, mientras que DIF no uniforme puede ser también DIF unidireccional o bidireccional. Esto último se puede observar en las Figuras 1.2 y 1.3 respectivamente.

Es importante mencionar que desde el punto de vista de la Teoría de Respuesta al Ítem (TRI), la presencia de DIF uniforme ocurre cuando el parámetro de dificultad del ítem es diferente entre los grupos, mientras que DIF no uniforme presenta dos variantes, éstas corresponden a DIF bidireccional, es decir que las CCI's presentan una intersección. De acuerdo con Hidalgo et al. (2005) ocurre DIF no uniforme simétrico cuando el parámetro de discriminación es diferente entre los grupos manteniendo constante el parámetro de dificultad, mientras que DIF no uniforme asimétrico ocurre cuando tanto el parámetro de dificultad como el de discriminación son diferentes entre los grupos.

Un ejemplo gráfico de DIF no uniforme simétrico se presenta en la figura 1.4, la intersección de las curvas ocurre en niveles medios del atributo medido. Por lo

general, este tipo de DIF se caracteriza porque la intersección de las CCI's ocurre en el valor del parámetro de dificultad.

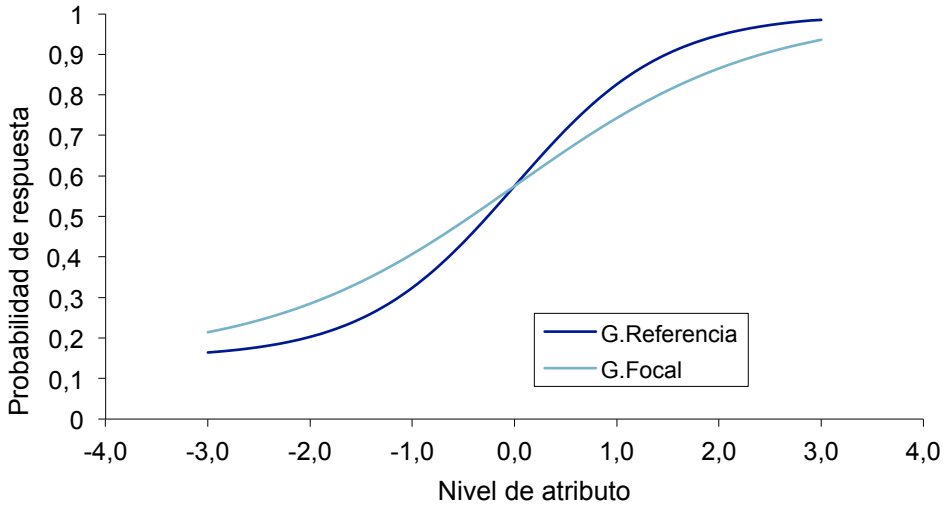


Figura 1.4. Curva Característica de un ítem con DIF no uniforme simétrico

Por su parte, en la figura 1.5 se presenta DIF no uniforme asimétrico. En este caso la intersección entre las CCI's puede ocurrir en cualquier punto del nivel de atributo, en este caso en concreto la intersección ocurre en niveles bajos del atributo medido.

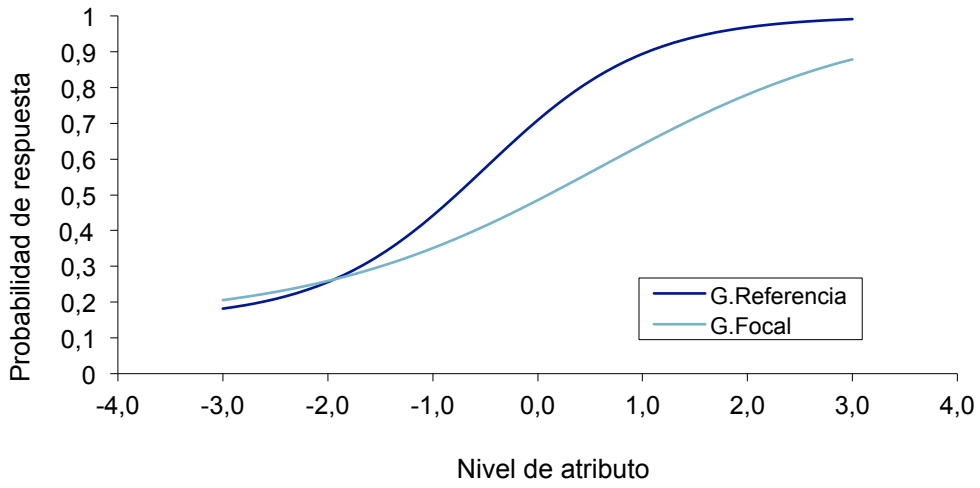


Figura 1.5. Curva Característica de un ítem con DIF no uniforme asimétrico

1.1.2.3 Tipo de ítems

Teniendo en cuenta la escala de medición, los ítems pueden ser de tipo dicotómicos o politómicos. El análisis psicométrico de ítems dicotómicos se considera más sencillo puesto que sólo existen dos categorías de respuesta, de las cuales sólo una proporciona información sobre el nivel de atributo y por tanto puede estudiarse la presencia de DIF por la diferencia entre grupos en la CCI del ítem que se está analizando.

Por su parte, el análisis de DIF en ítems politómicos es más complejo porque este tipo de ítems contienen más de dos categorías de respuesta que proporcionan información sobre el nivel de atributo. Para cada categoría de respuesta existe una curva característica (CCC: Curva Característica de la Categoría de respuesta) y la CCI corresponde a la suma ponderada de las CCC (Ali et al., 2015). Se debe tener en cuenta que los parámetros para un ítem politómico contemplan la existencia de un valor de dificultad para cada categoría de respuesta y el mismo parámetro de discriminación en cada categoría de respuesta (Masters, 1982; Samejima, 1969).

De acuerdo con Penfield (2010) existen dos aproximaciones al DIF con ítems politómicos, DIF global y DIF neto. DIF global se detecta cuando la diferencia entre grupos en la probabilidad condicional asociada con cualquier nivel de puntuación del ítem es diferente de cero. Es decir, si en algún nivel de puntuación del ítem se detecta DIF esto será suficiente evidencia para afirmar que dicho ítem presenta DIF. Por su parte, DIF neto evalúa el efecto neto agregado a través de todos los niveles de puntuación del ítem, por lo que se basa en una diferencia condicional entre los grupos con signo. En este caso, aunque se detecte DIF en niveles particulares de puntuación del ítem, su efecto puede cancelarse debido a que en algunos casos el funcionamiento diferencial puede estar favoreciendo al grupo de referencia y en otros al grupo focal (Penfield, 2010; Penfield et al., 2009).

1.2. Métodos de detección de DIF.

Desde finales de la década de los 80s se han propuesto numerosos métodos para la detección de DIF. Dichos métodos cada vez presentan desarrollos con mayor refinamiento y sofisticación de la mano de los avances en la tecnología, para dar respuestas a las necesidades que desde la práctica o los estudios empíricos se han ido planteando.

Este apartado no pretende ser un compendio exhaustivo de todos los métodos de detección de DIF, por el contrario, lo que se pretende es proporcionar las características más notorias de los métodos que tradicionalmente se han empleado en el análisis de DIF, y posteriormente se presentarán con un poco más de detalle los métodos estudiados en dos de las publicaciones que componen la tesis.

1.2.1. Esquemas de clasificación de los métodos de detección de DIF

En la literatura pueden encontrarse algunas propuestas que permiten clasificar los métodos de detección de DIF. Uno de los primeros esquemas de clasificación fue el propuesto por Camilli y Shepard (1994) quienes de acuerdo con la fundamentación teórica del método propusieron tres categorías:

1. Métodos basados en la Teoría Clásica del Test (TCT) y en el análisis de varianza.
2. Métodos basados en las Tablas de contingencia.
3. Métodos basados en la Teoría de Respuesta al Ítem (TRI).

Otra propuesta de clasificación de los métodos corresponde a la de Sireci y Ríos (2013) que proponen que los métodos de detección de DIF se clasifican de acuerdo con dos categorías:

1. Cálculo del nivel de atributo (puntaje observado / puntaje basado en TRI)

2. Tipo de técnica estadística (Aproximaciones de estadística descriptiva / Salida gráfica / Tablas de contingencia / Modelos de regresión / Métodos basados en TRI).

Esta propuesta incluye un nivel adicional para la categorización de los métodos en comparación con la propuesta de Camilli y Shepard (1994), pero puede resultar poco ilustrativa para el usuario final puesto que no se mencionan aspectos propios del test que orienten la decisión sobre qué método emplear.

Otro esquema de clasificación que permite un mayor grado de discriminación entre los métodos, y que incluye un aspecto observable directamente en el test es el propuesto por Potenza y Dorans (1995) quienes proponen una clasificación basada en tres variables:

1. Tipo de ítems que se van a analizar (dicotómicos /politómicos).
2. El criterio que se emplea para la igualación de los grupos (variable observada / variable latente).
3. El modelo empleado para establecer la relación entre el puntaje del ítem y la variable de igualación (no paramétrico / paramétrico).

Basándose en la propuesta de Potenza y Dorans (1995), Hidalgo y Gómez-Benito (2010) actualizaron el esquema agregando los métodos desarrollados después de la propuesta original. En la tabla 1.1 se presenta una adaptación y actualización de dicha propuesta. Respecto a los métodos no parámetricos listados cabe destacar lo siguiente:

- MH se refiere al procedimiento Mantel-Haenszel estándar propuesto por Holland y Thayer (1988) ya sea que la detección de DIF se realice por medio del estadístico ji-cuadrado de MH (χ^2 MH) o mediante el estadístico delta de MH (MH D-DIF).
- MH-Bayesiano hace referencia a la aplicación del método bayesiano empírico (Zwick et al., 2000) y el método bayesiano completo (Sinharay et al., 2009) en la clasificación del DIF basado en la detección de DIF con MH.

- Cochran-MH hace referencia al procedimiento Cochran-Mantel Haenszel que es aplicable tanto para ítems dicotómicos como politómicos (Meyer et al., 2004).
- Liu Agresti corresponde a la aplicación del estimador de Liu Agresti en la detección de DIF en ítems dicotómicos y politómicos (Penfield & Algina, 2003).
- Delta plot contiene la propuesta que mejora la definición del umbral para la detección de DIF realizada por Magis y Facon (2012) y el desarrollo del paquete deltaPlotR para su aplicación por medio del software R (Magis & Facon, 2014).

Tabla 1.1. Clasificación de los métodos de detección de DIF basada en la propuesta de Hidalgo y Gómez-Benito (2010)

Tipo de ítem	Puntaje observado		Variable latente	
	No paramétrico	Paramétrico	No paramétrico	Paramétrico
Dicotómico	MH MH-Bayesiano Cochran-MH Liu Agresti Estandarización Delta-Plot	RL Modelos log-lineales Modelos de clase latente	SIBTEST CATSIB	IRT log-likelihood ratio χ^2 de Lord Medidas del área DIFT-TRI AFC MIMIC Modelo Bayesiano-TRI CDM Modelo jerárquico lineal RASCH-DIF
Politómico	MH Generalizado MH Ordinal Cochran-MH Liu Agresti Mantel Estandarización	RL multinomial RL discriminante Modelos log-lineales Modelos de clase latente	POLYSIBTEST	IRT log-likelihood ratio χ^2 de Lord Medidas del área DIFT-TRI AFC MIMIC Modelo jerárquico lineal RASCH-DIF

Nota: MH = Mantel Haenszel; RL = Regresión Logística; IRT log-likelihood ratio = Logaritmo de la razón de verosimilitud en la Teoría de Respuesta al Ítem, DIFT-TRI = Funcionamiento Diferencial del Ítem y del Test basado en la Teoría de Respuesta al Ítem; AFC = Análisis Factorial Confirmatorio; MIMIC = Múltiple Indicador Múltiple Causa; CDM = Modelo de Diagnóstico Cognitivo.

Por su parte, y respecto de los métodos paramétricos que se listan en la tabla 1.1 cabe mencionar lo siguiente:

- AFC incluye métodos como Análisis factorial confirmatorio multi-grupo (Meredith, 1993; Vandenberg & Lance, 2000), Análisis factorial restringido (Barendse et al., 2010; Oort, 1998), y Análisis factorial de clase latente (Kankaras et al., 2011; Magidson & Vermunt, 2001).

- El Modelo Bayesiano-TRI hace referencia a la aplicación de la aproximación bayesiana para la detección de DIF basada en la estimación de los parámetros del ítem basada en TRI.
- El método basado en CDM, aunque corresponde a un campo específico de aplicación en la detección de DIF, contiene algunos estadísticos particulares que vale la pena mencionar como: el test de wald estándar y el test de wald basado en la matriz de información de producto cruzado, en la matriz de información observada, y en la matriz de covarianzas tipo sándwich (Liu et al., 2019).
- El modelo jerárquico lineal contiene el modelo lineal jerárquico generalizado (Chen et al., 2014; Cheong, 2006) y su aplicación para ítems politómicos (Williams & Beretvas, 2006).
- Finalmente, en Rasch-DIF se incluyen además del procedimiento clásico diferencia del parámetro de dificultad (Dabra, 1977; Gattamorta et al., 2012), el análisis con algunos índices de ajuste (Seol, 1999) y procedimientos basados en árboles de Rasch (Strobl et al., 2015; Tutz & Berger, 2016). En este punto cabe mencionar que los métodos empleados tanto en el estudio empírico como en el experimental están incluidos en este grupo.

1.2.2. Métodos de detección de DIF

En este apartado se describirán muy brevemente algunos métodos clásicos para detectar DIF, entre ellos se encuentran: Mantel Haenszel, Regresión logística y SIBTEST. Posteriormente se describirán con más detalle los procedimientos empleados en los estudios de la tesis: diferencia del parámetro de dificultad y árboles de Rasch.

1.2.2.1 Mantel Haenszel (MH)

Como método para la detección de DIF fue propuesto por Holland y Thayer (1988) y se basa en el análisis de tablas de contingencia. Ha sido uno de los métodos más estudiados y aplicados, especialmente cuando se trata de DIF uniforme (Clauser & Mazor, 1998). Lo anterior ha estado muy influenciado por su bajo coste computacional, sencillez en cuanto a su cálculo y supuestos del modelo.

MH construye *k* tablas de contingencias de 2x2 para cada ítem, donde *k* hace referencia al número de intervalos o estratos en los que se divide el puntaje total del test. En dicha tabla se compara la respuesta al ítem entre el grupo de referencia y el focal tal como se muestra en la tabla 1.2. Las letras A, B, C, y D corresponden al número de examinados que cumplen con las condiciones tanto de la puntuación del ítem como de pertenencia a un grupo en un nivel *k* de puntuación total. Los valores marginales de columna corresponden al número total de examinados que responden correctamente el ítem en el nivel *k* (N_{1k}) y los que no lo responden correctamente en el mismo nivel (N_{0k}). Por su parte los valores marginales de fila corresponden al número total de examinados en el grupo de referencia con un nivel *k* (N_{Rk}) y en el grupo focal con el mismo nivel (N_{Fk}). Finalmente, el valor N_k representa el total de examinados en el nivel de puntuación *k*.

Tabla 1.2. Tabla de contingencia 2x2 para un determinado nivel de puntuación total (k)

	Puntuación en el ítem		Total
	Correcta = 1	Incorrecta = 0	
Grupo Referencia	A_k	B_k	N_{Rk}
Grupo Focal	C_k	D_k	N_{Fk}
Total	N_{1k}	N_{0k}	N_k

MH proporciona una serie de indicadores estadísticos para determinar la presencia de DIF: un estadístico que está asociado con una prueba de significación y que sigue una distribución χ^2 con un grado de libertad (χ^2_{MH}) y una medida basada en el cociente de razones común (α_{MH}) cuyos valores pueden oscilar entre 0 e infinito (Guilera et al., 2007). Además cuenta con una medida de clasificación de la magnitud de DIF llamada MH D-DIF (Holland & Thayer, 1985).

Este método presenta algunas limitaciones relacionadas con la detección de DIF no uniforme, la detección de DIF con ítems politómicos y con el criterio para establecer los estratos. Aunque existe una variación de este procedimiento para detectar DIF no uniforme (Clauser et al., 1994) los resultados indican una menor potencia que los métodos paramétricos (Hidalgo & López-Pina, 2004), la detección

con ítems politómicos se afecta por las diferencias entre grupos en la distribución del nivel de atributo medido (Wang & Su, 2004). El criterio para establecer los estratos o niveles de atributo medido son arbitrarios y esto puede afectar la detección de DIF (Hidalgo & Gómez-Benito, 2010).

1.2.2.2 Regresión logística (RL)

El método regresión logística (RL) para detectar DIF fue propuesto por Rogers y Swaminathan (1993). La detección de DIF se realiza comparando el ajuste del modelo cuando la variable grupo y/o la interacción entre grupo y el nivel de atributo (calculado a través de la puntuación total observada en el test o estimada a través de algún modelo TRI) se adicionan. RL proporciona un modelo que predice la probabilidad de responder correctamente un ítem en función del nivel de atributo y de la pertenencia o no a un grupo determinado, así:

$$p(y = 1|\theta) = \frac{e^Z}{1 + e^Z}$$

donde $p(y = 1|\theta)$ corresponde a la probabilidad condicional de responder el ítem correctamente y Z contiene la expresión lineal de las variables predictoras (Gómez-Benito et al., 2009) y representa los tres modelos que se comparan:

$$Modelo\ 3. \qquad Z = \beta_0 + \beta_1\theta + \beta_2 G + \beta_3\theta G$$

$$Modelo\ 2. \qquad Z = \beta_0 + \beta_1\theta + \beta_2 G$$

$$Modelo\ 1. \qquad Z = \beta_0 + \beta_1\theta$$

donde θ es el nivel de atributo de los examinados y G representa el grupo al que pertenecen. El modelo 3 es el modelo completo y permite evaluar la presencia de DIF uniforme y no uniforme simultáneamente. El segundo modelo permite evaluar la presencia de DIF uniforme y el modelo 1 evalúa la ausencia de DIF. Para detectar DIF se compara entre estos tres modelos el que mejor se ajusta a los datos de acuerdo con el principio de parsimonia. La comparación del modelo 1 con el modelo 2 permite detectar DIF uniforme, mientras que la comparación del modelo 2 con el modelo 3 permite detectar DIF no uniforme y la comparación del modelo 1 con el modelo 3 es un indicador de la presencia de DIF uniforme y no uniforme simultáneamente (Gómez-Benito et al., 2009).

Este método ha sido muy utilizado puesto que es bastante robusto para evaluar DIF no uniforme, permite incluir como variable de estudio aquellas de tipo continuo e incluye diversas medidas de tamaño del efecto que reducen el error tipo I (Zumbo, 1999). Pese a lo anterior también se han señalado una serie de desventajas al emplear este método. En comparación con otros métodos clásicos RL requiere un mayor cumplimiento de supuestos del modelo, una baja variabilidad en los ítems puede producir una mayor tasa de error tipo I (Millsap & Everson, 1993) y el tamaño de muestra depende del número de categorías de respuesta y la asimetría de los ítems (Hidalgo & Gómez-Benito, 2010).

1.2.2.3 Test de sesgo simultáneo (SIBTEST)

Este método es una modificación del método de estandarización que aplica una aproximación del modelo multidimensional de sesgo bajo la TRI y fue propuesto por Shealy y Stout (1993). Al contrario de la TRI, que impone un modelo de respuesta específico al ítem, SIBTEST no especifica una forma o modelo funcional a los ítems y por tanto se considera que corresponde a un método no paramétrico. Además, permite detectar simultáneamente DIF, DTF y el efecto de cancelación o amplificación de DIF.

En términos generales, SIBTEST evalúa la presencia de DIF por medio del cálculo de diferencias ponderadas. Las diferencias se ponderan por la proporción de examinados del grupo focal que obtuvieron una puntuación total determinada (*j*) y la diferencia se calcula entre el promedio de la puntuación obtenida, en un ítem o en un grupo de ellos, por parte de los examinados que obtuvieron la puntuación *j* del grupo de referencia y los del grupo focal (Magis et al., 2010).

Este método ha suscitado algunos desarrollos específicos para abarcar las características más comunes en la detección de DIF. Así, se ha desarrollado un procedimiento para valorar el DIF en ítems politómicos (POLYSIBTEST; Chang, et al., 1996), y para detectar DIF no uniforme (Crossing-SIBTEST; Li & Stout, 1996) para lo cual ha demostrado tener una potencia estadística alta.

Una de las desventajas de este método es que no es completamente robusto ante la presencia de impacto o diferencias en la distribución del nivel de atributo medido. Además, puede no resultar muy adecuado con muestras pequeñas (Bolt, 2002), la detección de DIF no uniforme con ítems politómicos es problemática y para aplicarlo se requiere un software específico.

1.2.2.4 Diferencia del parámetro de dificultad del ítem basada en el modelo Rasch

De acuerdo con Andrich y Hagquist (2012), bajo el modelo Rasch se parte del siguiente supuesto:

$$Pr\{X_{ni} = x_{ni}\} = \frac{e^{(x_{ni}(\beta_n - \delta_i))}}{1 + e^{(\beta_n - \delta_i)}}$$

donde $Pr\{X_{ni} = x_{ni}\}$ representa la probabilidad de que una persona (n) que responde el ítem (i) obtenga una puntuación de 0 o 1 en el caso de ítems dicotómicos. β_n corresponde al parámetro estimado para la persona (n) y δ_i se refiere al parámetro estimado para el ítem (i), estos dos parámetros se obtienen bajo el mismo continuo de nivel de atributo medido o rasgo latente.

De lo anterior se extrae que la respuesta a un ítem bajo este modelo depende del nivel de atributo de la persona (β) y del nivel de dificultad del ítem (δ). Así, para un ítem y persona específicos la probabilidad de responderlo correctamente puede expresarse por medio de la siguiente ecuación:

$$P_{ni}\,(x = 1) = f(\beta_n - \delta_i)$$

De acuerdo con Paek y Wilson (2011), la forma general del modelo Rasch para DIF puede expresarse como:

$$P(x_{ni} = 1 | \beta_n, g) = \beta_n - \delta_i + \gamma_i G$$

Esta ecuación se lee como: la probabilidad de responder correctamente un ítem condicionada a un nivel de atributo para uno de los grupos (g) referencia (R) o focal (F), es igual al nivel de atributo menos el parámetro de dificultad del ítem

más el índice de DIF (γ_i) o la diferencia del parámetro de dificultad entre el grupo focal y el de referencia ($\gamma_i = \delta_F - \delta_R$). $G = 1$ si $g = R$ y $G = 0$ si $g = F$.

El algoritmo general que se sigue con este procedimiento consta de 4 pasos: 1) se realiza un análisis conjunto de todos los examinados en el que se anclan los valores del parámetro de los ítems y se calcula el nivel de atributo de los examinados. 2) se calcula el parámetro de los ítems para el grupo de referencia manteniendo las estimaciones del nivel de atributo obtenidas en el primer paso y anclando el valor de la dificultad de todos los ítems menos el que se está analizando. 3) se calcula el parámetro de dificultad de los ítems para el grupo focal de la misma manera que en el segundo paso. 4) se calculan las diferencias entre los parámetros de dificultad ($\delta_F - \delta_R$) (Hauser & Kingsbury, 2004).

Para la detección de DIF, γ_i indica la magnitud de DIF y se puede calcular el estadístico t-student para probar la hipótesis nula de igualdad de la dificultad entre los grupos.

$$t_i = \frac{y_i}{SE\ (y_i)} = \frac{\delta_F - \delta_R}{\sqrt{s_{if}^2 + s_{ir}^2}}$$

Además de la prueba estadística de significación, la magnitud de la diferencia entre grupos en los parámetros de dificultad se ha sugerido como una medida de tamaño del efecto. Así, una diferencia mayor o igual a 0.5 logits junto con un resultado significativo de la diferencia entre parámetros ha sido recomendado como criterio para considerar la existencia de DIF (Linacre, 2021). Un criterio similar había sido propuesto por Lai et al. (2005) quienes afirmaron que una diferencia de 0.5 logits era suficiente para la detección de DIF y adicionalmente, propusieron una categoría intermedia, en la que se considera una diferencia entre 0.35 y 0.5 como DIF probable.

Por otro lado, Paek y Wilson (2011) propusieron aplicar una métrica similar a las reglas de clasificación del Servicio de Evaluación Educativa (ETS por sus siglas en

inglés: Educational Testing Service), es decir las categorías A, B y C de DIF basadas en MH D-DIF (Holland & Thayer, 1985). Esta propuesta se origina en la premisa de que la puntuación total del test bajo el modelo Rasch es un estadístico suficiente como nivel de atributo y que por tanto $\alpha_{MH} = \alpha = e^{\delta_F - \delta_R}$, que también fue demostrado por Linacre y Wright (1987), entonces:

$$\Delta_{MH} = -2.35 \ln(\alpha_{MH}) = -2.35 (\delta_F - \delta_R) = -2.35 \gamma$$

Partiendo de lo anterior estos autores proponen los siguientes puntos de corte:

- Categoría A si $|\gamma| \leq 0.426$ o si la diferencia en el parámetro de dificultad no es significativa al nivel de 0.05.
- Categoría B si $0.426 < |\gamma| < 0.638$ y si la diferencia entre los parámetros de dificultad es significativa al nivel de 0.05.
- Categoría C si $|\gamma| \geq 0.638$ y si la diferencia entre los parámetros de dificultad es significativa al nivel de 0.05.

Debido a que este procedimiento se basa en estadísticos inferenciales, la variable tamaño de muestra ha tomado un papel relevante en el análisis de su adecuado funcionamiento. Por lo general se tiene la creencia que con métodos paramétricos se requieren amplios tamaños de muestra, aproximadamente entre 1,000 y 2,000 para obtener tasas adecuadas de detección correcta (Ankenmann et al., 1999; French & Miller, 1996; Hidalgo & Gómez, 2006; Hidalgo-Montesinos & Gómez-Benito, 2003).

Al analizar el impacto general del tamaño de muestra sobre la detección de DIF de este método (Bernstein et al., 2013; Lai et al., 2005; Paek & Wilson, 2011; Rouquette et al., 2019) se ha concluido, como con otros métodos, que cuanto mayor es el tamaño de muestra mayor será la potencia estadística del procedimiento. Pero, ¿cuál es el tamaño mínimo de muestra admisible que permite una adecuada detección de DIF con este procedimiento? La respuesta a esta pregunta parece estar en torno a 100 personas para cada grupo.

Schulz et al. (1996) compararon la detección de DIF, para la variable género, con MH y con diferencia del parámetro de dificultad en muestras empíricas extraídas de una población de 60,000 estudiantes quienes contestaron 46 ítems. Los tamaños de muestra total analizados fueron 200, 600 o 2,000. Los resultados indicaron que el procedimiento diferencia del parámetro de dificultad es más sensible y fiable que MH cuando hay tamaños de muestra pequeños si la distribución del nivel de atributo entre los grupos es igual, es decir si no hay impacto.

Paek y Wilson (2011) compararon el funcionamiento del procedimiento diferencia del parámetro de dificultad con MH mediante un estudio de simulación con muestras pequeñas (de 100 a 300 examinados por grupo) y tests cortos (de cuatro a 39 ítems). En general, sus resultados indicaron que el procedimiento diferencia del parámetro de dificultad mostró un buen funcionamiento tanto en la detección correcta de DIF como en el control de tasas de falsos positivos. En comparación con MH, las tasas de detección correcta fueron más altas en todas las condiciones simuladas para el procedimiento diferencia del parámetro de dificultad.

Otros aspectos que se han estudiado sobre el funcionamiento de este procedimiento tienen que ver con el efecto que diversos procedimientos de anclaje tienen sobre la detección de DIF (Huang, 2014; Kopf et al., 2015; Wang, 2004), la detección del funcionamiento diferencial del umbral entre categorías (Gattamorta et al., 2012), el efecto de tamaños pequeños del grupo focal (Bernstein et al., 2013) y el efecto de la diferencia en la distribución del nivel de atributo entre grupos cuando los examinados responden correctamente por azar (DeMars & Jurich, 2015).

Respecto a los métodos de anclaje y la detección de DIF bajo el modelo Rasch, Wang (2004) encontró que cuando se empleó el método con 1 ítem constante se obtuvieron estimaciones no sesgadas, un error tipo I controlado y potencia razonable (en condiciones de 20% y 50% ítems con DIF), mientras que cuando se empleó el método de igualdad en la dificultad media (EMD por sus siglas en

inglés: Equal Mean Difficulty) o el método de todos los otros ítems (all-other anchor item method) se obtuvieron estimaciones sesgadas, inflación del error tipo I y potencia inapropiada si la diferencia de la dificultad media del ítem (MIDD por sus siglas en inglés: Mean Item Difficulty Difference) entre los grupos de referencia y focal no fue de cero (0). Otro hallazgo fue que a mayor número de ítems conformen el anclaje con el método constante, más controlado estará el error tipo I y mayor será la potencia. Además, Huang (2014) encontró que cuando se aplicó este método de anclaje los resultados no parecieron afectarse por el tamaño de muestra (15,000 o 20,000), la longitud del test y el tipo de ítems (dicotómicos: 40 o 60, politómicos: 30 o 50). Por otro lado, Kopf et al. (2015) confirmaron que el método de todos los otros ítems (all-other anchor item method) presentó una elevada inflación del error tipo I cuando una gran proporción de DIF favorece al grupo de referencia, que el método de anclaje en formato iterativo mostró una relativa robustez en dicha condición y que la detección de DIF se vio afectada por el tamaño de la muestra, la proporción de ítems con DIF, la dirección del DIF y el número de ítems en el anclaje.

En cuanto a la detección del funcionamiento diferencial del umbral entre categorías, Gattamorta et al. (2012) demostraron una adecuada detección por parte del software Winsteps en ítems politómicos bajo el Modelo de Crédito Parcial (PCM, por sus siglas en inglés: Partial Credit Model) y que en comparación con procedimientos no paramétricos basados en Odds-Ratio (Razón de Odds) los resultados obtenidos son similares en la detección del funcionamiento diferencial y en su magnitud.

Por otro lado, Bernstein et al. (2013) encontraron que no había diferencias en las tasas de error tipo I debidas al tamaño del grupo focal o a la diferencia en el atributo medido. Sin embargo, las tasas de error tipo II si fueron diferentes para los dos grupos focales más pequeños (20 examinados o menos). Además, estos autores comprobaron que el procedimiento diferencia del parámetro de dificultad presenta resultados similares a MH con tamaños de grupo que van desde 10 hasta

1,000 examinados. Los dos procedimientos presentaron una inadecuada detección de DIF cuando el grupo focal fue pequeño y aunque MH resultó algo más hábil bajo estas circunstancias, su tasa de error tipo I se incrementó.

Finalmente, en relación con la diferencia en la distribución del nivel de atributo entre grupos, DeMars y Jurich (2015) concluyeron que cuando las diferencias en la media del nivel de atributo fueron grandes y se presentaron respuestas correctas por azar, el error tipo I aumentó y que cuando los grupos son iguales en el nivel de atributo la detección de DIF en ítems sin DIF fue adecuada y en ítems con DIF se presentó un relativo sesgo. Además, estos autores propusieron como estrategia para mitigar el efecto de la presencia de respuestas al azar, su transformación a datos perdidos. Sin embargo, advirtieron de la reducción del tamaño de muestra para la estimación de DIF y su correspondiente incremento en el error estándar de las estimaciones.

1.2.2.5 Árboles de Rasch

Este tipo de procedimientos se proponen como una alternativa que permite incluir en el análisis de DIF varias variables a la vez y cuando se trata de variables de tipo continuo no es necesario determinar a priori los grupos de referencia y focal, a diferencia de los procedimientos clásicos, que sí requieren establecer un criterio a priori para definir los grupos. Por ejemplo, si se analiza DIF por la variable edad con un método como MH será necesario que el investigador defina previamente un punto de corte para conformar cada uno de los grupos. Para ello, puede emplear criterios teóricos o estadísticos como la mediana de la distribución de la variable, que es un criterio ampliamente usado (Strobl et al., 2015). Sin embargo, se debe tener en cuenta que sea cual fuere el criterio que el investigador emplee, éste puede no corresponderse con el valor empírico de la variable que causa el funcionamiento diferencial. Así, la aproximación a priori para la definición de los grupos puede causar pérdida de información ya que no se evalúan todos los posibles valores de la variable (Strobl et al., 2015).

Otro aspecto diferencial entre los procedimientos clásicos y los árboles de Rasch es que al incluir en el análisis varias variables y evaluar sus valores posibles, la detección de DIF en un ítem incluye combinaciones de variables y todos sus valores (Tutz & Berger, 2016). Por tanto, la detección de DIF por una variable no significa que todos los examinados que pertenecen a un subgrupo estén consistentemente en ventaja o desventaja puesto que puede presentarse interacción entre las variables (Chen & Jiao, 2014).

Los procedimientos árboles de Rasch emplean el modelo basado en técnicas de partición recursiva que es adaptado de la econometría y se emplea para evaluar el cambio estructural detectando diferencias en los parámetros de un modelo estadístico entre grupos de participantes definidos por variables o por una combinación de ellas (Strobl et al., 2015). La partición recursiva se basa en la segmentación sucesiva de la muestra; los datos se dividen siempre que los diferentes grupos presenten valores sustancialmente diferentes en el parámetro estimado (Kopf, 2013). Nótese que, a diferencia de técnicas como árboles de clasificación y árboles de regresión, esta técnica particiona los datos en aquellos valores de las variables en los cuales el parámetro de un modelo paramétrico varía entre grupos. En pocas palabras, la aplicación del modelo basado en la partición recursiva comienza con la formulación de un modelo paramétrico, posteriormente éste se analiza por el algoritmo del modelo basado en la partición recursiva y se determina si otras covariables podrían alterar los parámetros del modelo de interés, que para el caso del modelo Rasch corresponde al parámetro de dificultad (Kopf, 2013).

Respecto a la detección de DIF, árboles de Rasch incluye dos procedimientos: Árboles basados en Rasch (Rasch-Trees; Strobl et al., 2015) y Árboles Enfocados en los Ítems (IFT por sus siglas en inglés: Item-Focused Trees; Tutz & Berger, 2016). Algunos términos a tener en cuenta en el análisis de los árboles mediante partición recursiva son: *árbol* representa la partición jerárquica del espacio predictor o de la muestra, *nodo* del árbol corresponde a un subconjunto o submuestra, *raíz* es el

nodo principal que contiene la muestra completa, *nodos terminales u hojas* son las subregiones finales (Tutz & Berger, 2016).

1.2.2.5.1 Procedimiento Rasch-Trees

Este procedimiento fue propuesto originalmente por Strobl et al. (2015) para ítems dicotómicos y posteriormente, Komboz et al. (2018) propusieron una extensión para evaluar ítems politómicos (Tree-PCM).

De acuerdo con Strobl et al. (2015) este procedimiento se lleva a cabo en 4 pasos:

1. Se estima el parámetro de dificultad para todos los ítems y para toda la muestra. La estimación de los parámetros de dificultad se realiza mediante la Máxima Verosimilitud Condicional puesto que se parte del supuesto de que bajo el modelo Rasch la puntuación observada es un estadístico suficiente para estimar el nivel de atributo.
2. Se evalúa la estabilidad de los parámetros de los ítems con relación a las variables de interés. Tomando las estimaciones del parámetro de dificultad para toda la muestra, las desviaciones individuales de dichas estimaciones se ordenan con respecto a los valores de una covariable, se espera que en promedio estas desviaciones individuales sean cero (0). Si existe DIF para esta covariable se observará un cambio sistemático de las desviaciones individuales en algún o algunos valores de la variable. Si por el contrario no existe DIF los valores de las desviaciones individuales cambiarán aleatoriamente. La ventaja que tiene esta aproximación es que las desviaciones individuales son las mismas para todas las covariables, lo único que cambia es su orden y por tanto no requiere que se reestimen cada vez que se analiza una covariable.
3. Cuando se encuentra inestabilidad o diferencias significativas en los parámetros de los ítems respecto a una variable, la muestra se divide tomando como referencia la variable con la inestabilidad más fuerte y en el punto de corte que proporciona la mayor mejora del ajuste del modelo. La prueba estadística de la inestabilidad del parámetro para cada covariable lleva

asociado un *p* valor que es corregido mediante el ajuste de Bonferroni. La covariable con el *p* valor más pequeño será la que produzca la primera división (o nodo).

4. Se repiten los pasos 1 a 3 para las submuestras resultantes hasta que no se detectan inestabilidades o hasta que el *p* valor del paso 3 excede el nivel de significación, o hasta que la submuestra es demasiado pequeña. Una vez definidas las covariables que generan inestabilidad, se determina el punto de corte óptimo como aquel valor que maximiza la verosimilitud logarítmica particionada de dos modelos (uno para todos aquellos con valores en el punto de corte y menor, y otro para todos aquellos con valores superiores al punto de corte), sobre todos los puntos de corte posibles.

El procedimiento proporciona información sobre los grupos de participantes que presentan DIF. Es decir, que su resultado informa al usuario sobre aquellas submuestras que son afectadas por la presencia de DIF en los ítems de un test en su conjunto. Por ello se evalúa la hipótesis nula que representa la igualdad del parámetro de dificultad de todos los ítems de un test. Nótese que para aceptar la hipótesis nula es necesario que todos los ítems estén libres de DIF, pero para rechazarla debe existir suficiente inestabilidad en los parámetros de los ítems.

Respecto a la eficacia de este procedimiento con relación a otros más comúnmente utilizados, Strobl et al. (2015) llevaron a cabo una serie de estudios con datos simulados en los que compararon las tasas de error tipo I, la potencia estadística y el recobro de los grupos simulados de procedimientos como Rasch-Trees, y Test de Razón de Verosimilitud.

En el primer estudio se emplearon dos tipos de covariables, una binaria y la otra numérica, pero el análisis de DIF sólo empleó una a la vez. Para la variable numérica se definieron dos puntos de corte uno ajustado en la mediana de la variable (en un valor de 50) y otro ajustado en un valor de 80. Debido a que el Test de Razón de Verosimilitud requiere la especificación a priori de los grupos en la

variable numérica, el punto de corte empleado para ello fue la mediana. Los resultados indicaron que Rasch-Trees no presentó inflación del error tipo I, obtuvo una potencia estadística similar al Test de Razón de Verosimilitud cuando el punto de corte fue conocido, pero una potencia estadística mucho más alta y mejor recobro de los grupos simulados cuando el punto de corte fue desconocido.

En el segundo estudio Strobl et al. (2015) ilustraron el efecto de la presencia de impacto sobre el error tipo I y la potencia estadística de Rasch-Trees y de Test de Razón de Verosimilitud cuando el análisis se aplicó con una variable binaria únicamente. Los resultados indicaron que la presencia de impacto no afectó el comportamiento de los procedimientos de detección de DIF y que tanto las tasas de error tipo I y la potencia estadística de los dos procedimientos fueron comparables.

El tercer estudio analizó el funcionamiento de dichos procedimientos bajo condiciones en las que la conformación del grupo focal y de referencia obedece a patrones no estándar (cuando la mediana no es el criterio de partición, patrones en forma de U e interacciones entre covariables). Para ello se incluyeron en los análisis variables de tipo binario y numérico que fueron analizadas por separado o simultáneamente. Además se compararon los resultados cuando se aplicó o no el ajuste de Bonferroni. Los resultados indicaron que Rasch-Trees no presentó inflación de las tasas de error tipo I aún cuando se analizaron dos covariables a la vez y que cuando la conformación de los grupos no se corresponde con la establecida a priori, Rasch-Trees obtuvo tasas de detección correcta más altas.

Por su parte y respecto al procedimiento Tree-PCM, Komboz et al. (2018) encontraron que los procedimientos Tree-PCM y Test de Razón de Verosimilitud presentaron tasas de error tipo I entorno al nivel nominal (0.05) con covariables de tipo binario y numérico. Sin embargo, el procedimiento Tree-PCM siempre obtuvo tasas más bajas que 0.05. Respecto a los resultados de la potencia estadística con la covariable binaria, los resultados fueron similares entre los dos procedimientos,

pero con la covariable numérica y cuando la conformación de los grupos se correspondió con la establecida a priori, la potencia estadística fue menor para Tree-PCM. Por otro lado, cuando la conformación de los grupos se realizó por patrones complejos (interacción entre covariables, la mediana no es el criterio de partición o patrones en forma de U), Tree-PCM presentó tasas más altas de detección de DIF y un mejor recobro de la estructura de los subgrupos. Debido a que en la práctica no se puede tener total certeza de que el criterio establecido a priori para definir los grupos de referencia y focal sea el correcto, se puede concluir que tanto Rasch-Trees como Tree-PCM han mostrado un adecuado funcionamiento en el control de las tasas de error tipo I, en la detección correcta de DIF y que muestran resultados superiores al Test de Razón de Verosimilitud.

1.2.2.5.2 Procedimiento Árboles Enfocados en los ítems –IFT-

Tutz y Berger (2016) propusieron el procedimiento IFT y posteriormente Bollmann et al. (2018) presentaron una aplicación de este procedimiento para ítems politómicos (PCM-IFT). A diferencia del procedimiento anterior esta propuesta evalúa la presencia de funcionamiento diferencial en cada uno de los ítems de un test uno por uno, por lo que se puede obtener un árbol distinto para cada ítem que sea detectado con DIF y dicho árbol puede tener una estructura distinta. Por otro lado, similar al procedimiento anterior éste también emplea el modelo basado en la partición recursiva para la detección de DIF. Así al comenzar el análisis se tiene en cuenta la estructura completa de la base de datos (incluyendo todos los ítems y la muestra) y se van evaluando sucesivamente cada una de las variables de interés y todos sus valores como posibles puntos de corte.

Tanto IFT como PCM-IFT se llevan a cabo en 3 fases (Bollmann et al., 2018; Tutz & Berger, 2016):

1. Estimación: se estiman los parámetros de todos los ítems para la muestra total por medio de la Máxima Verosimilitud Conjunta.
2. Selección: para todos los ítems se consideran todas las variables y sus posibles puntos de corte examinando en cada uno de ellos la hipótesis nula de

igualdad del parámetro del ítem que se evalúa mediante el Test de Razón de Verosimilitud. Si para todos los posibles puntos de corte de la variable la hipótesis nula se mantiene se puede afirmar que el ítem no presenta DIF respecto de dicha variable. Si por el contrario se detecta DIF en determinados valores se selecciona la combinación de ítem, variable y punto de corte que produce el *p* valor más pequeño, de tal manera que se selecciona el modelo más ajustado.

3. Decisión: se decide si la partición se debe realizar o no. Para ello existen dos criterios, uno basado en pruebas de significación y el otro basado en el tamaño de la muestra para cada nodo. Con el primer criterio se analiza la dependencia del ítem y la variable analizada, es decir la hipótesis nula de que la variable no afecta al ítem. Se toma un ítem y una variable y se aplica un test de permutación aleatoria de dicha variable que se controla por el nivel de significación. La toma de decisión de aplicar la partición o no se basa en el *p* valor y el punto de corte será aquel en el que el Test de Razón de Verosimilitud es máximo. Por otro lado, el segundo criterio que guía la toma de decisión sobre la presencia o no de una partición, se refiere a la presencia de un tamaño de muestra mínimo que permita una adecuada estimación del parámetro en cada nodo final. Komboz et al. (2018) siguieron como criterio mínimo para el tamaño de muestra en cada nodo aquel que corresponda a 10 veces el número de parámetros por ítem.

Se debe tener en cuenta que, debido a las múltiples pruebas realizadas, este procedimiento aplica un ajuste de Bonferroni para valorar la significación en cada permutación con un nivel inicial de significación de α/m (donde m corresponde al número de variables). Este ajuste se va adaptando en la medida que el análisis avanza y se van descartando variables. Así después de analizar todos los pasos para una primera variable, el siguiente análisis contará con una significación de α/m-1 y así sucesivamente.

En relación a su efectividad en la detección de ítems con DIF, Tutz y Berger (2016) compararon los resultados obtenidos por IFT con los de métodos previamente establecidos (MH, RL y X^2 de Lord) en una serie de condiciones experimentales distintas. En la primera condición se definió la presencia por separado de DIF inducido por una variable binaria y por una variable ordinal y se comparó IFT con MH, RL y X^2 de Lord. Los resultados indicaron que IFT obtiene resultados muy similares a MH y RL en cuanto a tasas de falsos positivos y de detección correcta tanto para la variable binaria como para la ordinal. Para esta última los resultados de IFT fueron levemente mejores.

La segunda condición experimental probó la presencia de DIF inducido por una variable continua y se comparó IFT con LR. Los resultados mostraron que IFT obtuvo tasas de falsos positivos por debajo de 0.05 mientras que para RL dichas tasas fueron algo mayores. Por su parte, las tasas de detección correcta fueron algo similares entre los dos procedimientos, aunque algo más altas para RL.

En la tercera condición experimental se definió la presencia de DIF inducido por variables de tipo binario, continuo y por su interacción comparando la ejecución de IFT con LR. Los resultados indicaron que con IFT se obtienen tasas de falsos positivos mucho menores (0.0263 a 0.0313) que LR (0.0563 a 0.0619) cuando DIF es inducido por variables binarias, continuas o por su interacción. Respecto a las tasas de detección correcta los resultados indicaron que IFT presentó tasas por encima de 0.83 cuando la magnitud de DIF fue media o fuerte, pero cuando ésta fue débil las tasas cayeron por debajo de 0.65, mientras que LR presentó tasas por encima de 0.80 cuando la magnitud de DIF fue media o fuerte y tasas menores a 0.70 cuando fue débil.

Finalmente, Tutz y Berger (2016) compararon la detección de DIF entre Rasch-Trees e IFT. Para ello generaron una condición experimental en la que dos variables (una binaria y la otra continua) generaron DIF y seis ítems de un total de 20 se simularon con DIF. En cinco ítems, la variable binaria indujo DIF y en el

sexto ítem tanto la variable binaria como la continua indujeron DIF. Al comparar las tasas de detección correcta se encontró que Rasch-Trees detectó el 100% de las veces la variable binaria, pero sólo un 16% de las veces la variable continua por tanto su tasa de detección correcta fue de 0.84. Los autores mencionaron que dicho resultado puede deberse a que la variable binaria al inducir DIF en un mayor número de ítems se hizo más detectable. Por su parte, IFT presentó una tasa de detección correcta de 0.963 y de falsos positivos de 0.03. Este procedimiento detectó adecuadamente la presencia de DIF inducido conjuntamente por la variable binaria y continua, así como el caso de la presencia de DIF inducido únicamente por la variable binaria.

Bollmann et al. (2018) compararon las tasas de falsos positivos y tasas de detección correcta entre PCM-IFT y Test de Razón de Verosimilitud, y entre PCM-IFT y Tree-PCM cuando una variable binaria indujo DIF y el número de ítems era ocho o 20. Los resultados indicaron que las tasas de falsos positivos fueron menores para PCM-IFT y que una inflación de éstas por encima del nivel nominal se presentó para Test de Razón de Verosimilitud cuando la magnitud de DIF fue fuerte. Respecto a las tasas de detección correcta los dos procedimientos mostraron el mismo patrón: tasas por encima de 0.80 cuando la magnitud de DIF fue media o fuerte y cuando fue débil tasas por debajo de 0.40. En cuanto a la comparación entre los procedimientos árboles de Rasch, las tasas de falsos positivos fueron considerablemente mayores para PCM-IFT (0.39 a 0.68) que para Tree-PCM (0.04 a 0.08). Sin embargo, los autores mencionaron que las tasas fueron calculadas al nivel de la variable porque Tree-PCM sólo aporta información en este nivel y por tanto dichas tasas para PCM-IFT fueron calculadas controladas al nivel del ítem (es decir, en la misma variable para cada ítem independientemente) por lo que estos resultados fueron esperables teniendo en cuenta que si la probabilidad de detectar un ítem erróneamente con DIF es de 0.05, la probabilidad de detectar al nivel de variable uno o más, de ocho ítems, será de 0.337 y de 20 será de 0.642. Respecto a las tasas de detección correcta éstas fueron similares (por encima de 0.90) entre los dos procedimientos cuando la magnitud de DIF fue fuerte y

mayores para PCM-IFT cuando la magnitud fue media (PCM-IFT: 0.88 a 1.0; Tree-PCM: 0.280 a 0.610) o débil (PCM-IFT: 0.49 a 0.85; Tree-PCM: 0.10 a 0.14).

Posteriormente estos autores simularon datos en los cuales las variables que inducían DIF eran de tipo binario, ordinal o numérico y cada una de ellas era la responsable de la presencia de DIF en un ítem. Para PCM-IFT las tasas de falsos positivos a nivel del ítem o a nivel combinado de ítems y variable estuvieron entorno a 0.05 y las tasas de detección correcta fueron considerablemente altas (0.977 a 1.00) con una magnitud de DIF fuerte, adecuadas (0.368 a 0.597) con una magnitud media e inadecuadas (0.096 a 0.124) con una magnitud débil. Al comparar los resultados de los dos procedimientos basados en árboles de Rasch los resultados indicaron que PCM-IFT obtuvo tasas altas de falsos positivos a nivel de la variable (0.075 a 0.131) mientras que Tree-PCM fue mucho más conservador (0.013 a 0.035). Respecto a las tasas de detección correcta PCM-IFT obtuvo tasas más altas que Tree-PCM independientemente de la magnitud de DIF. Con DIF fuerte PCM-IFT obtuvo tasas entre 0.977 y 1.00, mientras que para Tree-PCM las tasas oscilaron entre 0.70 a 0.97. Con DIF débil PCM-IFT obtuvo tasas entre 0.153 y 0.202, mientras que Tree-PCM obtuvo tasas entre 0.03 y 0.06. Finalmente, estos autores concluyeron que PCM-IFT presentó un buen funcionamiento frente a procedimientos clásicos, especialmente cuando había pocos ítems con DIF, como ocurre en la mayoría de escenarios de evaluación, y que Tree-PCM fue apropiado cuando la mayoría de ítems tenían DIF o al evaluar la presencia de DIF en un test sin detectar los ítems individuales.

2. Enfoque de trabajo y objetivos

Después de varias décadas del surgimiento del concepto de DIF, múltiples desarrollos se han generado en este ámbito y en los últimos años la relevancia de su estudio a nivel empírico se ha visto incrementada como una herramienta que permite aportar evidencia de la validez de las inferencias extraídas de los resultados del test.

Numerosos estudios se han orientado a probar la eficacia de uno u otro método de detección de DIF bajo determinadas condiciones, otros estudios han propuesto nuevos métodos o han presentado un refinamiento de los ya existentes, otros se han encargado de la aplicación empírica para determinar si un test está afectado por el funcionamiento diferencial, y otros desde una perspectiva metaanalítica han analizado las variables que pueden o no afectar el adecuado funcionamiento de algunos métodos.

Ante la enorme producción científica que versa sobre el funcionamiento diferencial, se consideró relevante hacer una síntesis analítica respecto de los métodos que se han propuesto para la detección de DIF tomando como objeto de estudio los artículos que con datos simulados han propuesto o analizado la efectividad de dichos métodos. Esta síntesis fue un punto de partida que permitió analizar los diferentes tópicos de investigación que han centrado el interés de los investigadores cuando realizan estudios con datos simulados. Además, permitió describir las variables, aspectos propios de los estudios con datos simulados y presentar algunos desarrollos recientes y últimas tendencias. Para nuestro conocimiento, hasta el momento no se ha realizado ninguna síntesis de este tipo en cuanto a métodos de detección de DIF.

Al observar que el método basado en el modelo Rasch era de los que más interés había despertado entre los investigadores y que adicionalmente estaba incluido en aquellos métodos de desarrollos recientes, se diseñaron dos estudios independientes con perspectivas y ámbitos de aplicación diferentes, pero que pretendían proporcionar nueva evidencia de su robustez y aplicabilidad.

Además, para la realización de esta tesis han concurrido diferentes elementos contextuales que moldearon el enfoque y diseño de la tesis: (a) la necesidad de contar con una síntesis y análisis de la información referente a los métodos de detección de DIF, (b) el hecho de que la gran mayoría de estudios sobre dichos métodos se han realizado desde el ámbito educativo y el de la salud, (c) la experiencia profesional y predoctoral, (d) la vinculación y colaboración entre los grupos de investigación "Métodos e Instrumentos para la Investigación en Ciencias del Comportamiento" de la Universidad Nacional de Colombia y el "Grupo de Estudios de Invarianza de la Medida y Análisis del Cambio – GEIMAC" de la Universidad de Barcelona (ver figura 2.1).

Los cuatro elementos mencionados previamente se integraron en el enfoque de esta tesis doctoral de tal manera que se decidió que el abordaje metodológico incorporaría una perspectiva histórico-contextual, una experimental y otra empírica. La perspectiva histórico-contextual se cubrió con una revisión de la literatura, mientras que para las perspectivas experimental y empírica se diseñaron estudios independientes cada uno de los cuales respondió a necesidades e intereses particulares relacionados con los elementos b, c, y d, mencionados en el párrafo anterior, teniendo en cuenta algunos resultados obtenidos en el estudio de perspectiva histórico-contextual.

En la perspectiva experimental tuvo un peso fundamental la trayectoria del grupo de investigación de la Universidad Nacional de Colombia con los estudios de DIF en el ámbito educativo principalmente en pruebas de aplicación masiva, mi experiencia profesional previa también en la evaluación en el ámbito educativo

(Ministerio de Educación Nacional de Colombia y en el Instituto Colombiano para la Evaluación de la Educación –ICFES) y la necesidad de probar el funcionamiento del procedimiento que emplea el ICFES para detectar DIF en las pruebas nacionales que dan acceso a la universidad y que pretenden evaluar la calidad de la educación. Para ello se simularon las principales características de dichos exámenes teniendo en cuenta las diferencias existentes entre población que pertenece a grupos indígenas y población no indígena.

Por su parte, en la perspectiva empírica tuvo un peso fundamental la trayectoria del grupo GEIMAC, en especial todo lo relacionado con la evaluación de la funcionalidad en esquizofrenia, mi experiencia predoctoral en dicho grupo, las características típicas de la evaluación en el ámbito de la salud, y el interés de ilustrar la aplicación de procedimientos propuestos recientemente para la detección de DIF a fin de estimular su uso en la evaluación en salud.

De esta manera, la tesis doctoral proporciona información novedosa relacionada con las tendencias y desarrollos recientes en métodos de detección de DIF, no sólo a través de la revisión de la literatura, sino también con los estudios independientes. Mientras que el estudio experimental aporta información novedosa respecto al funcionamiento de uno de los métodos más empleados en evaluación educativa, el estudio empírico aporta información novedosa respecto del análisis de DIF en el Cuestionario para la Evaluación de la Discapacidad de la Organización Mundial de la Salud 2.0 (WHODAS 2.0 por sus siglas en inglés: World Health Organization Disability Assessment Schedule 2.0) en personas con esquizofrenia.

El objetivo principal de la tesis fue analizar y proporcionar información referente a tendencias y desarrollos recientes en métodos de detección de DIF teniendo en cuenta sus principales ámbitos de aplicación. Para ello se definieron tres objetivos específicos que guiaron la realización de cada uno de los estudios que componen esta tesis doctoral.

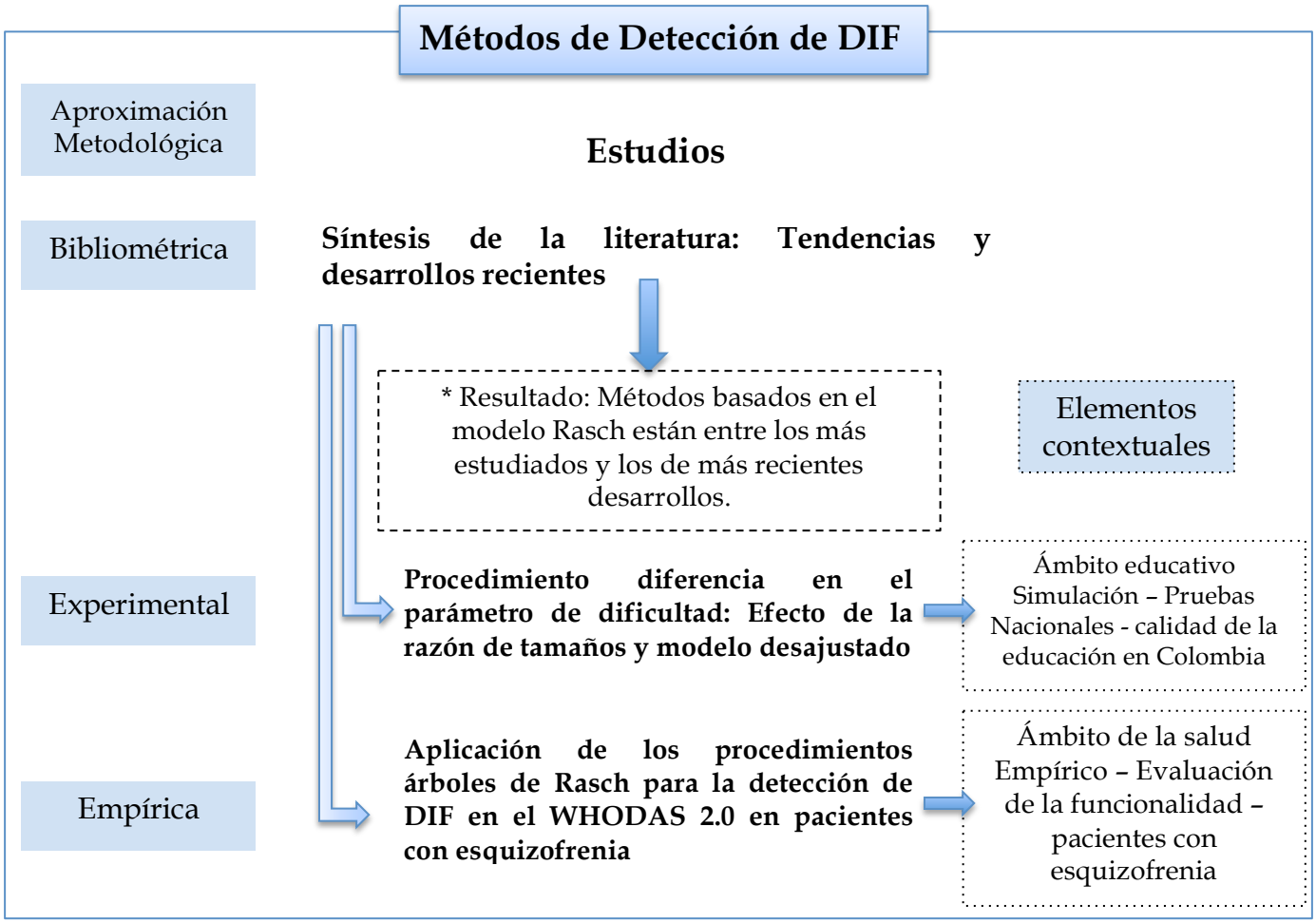


Figura 2.1. Esquema resumen del enfoque y contexto de trabajo.

Para comenzar, el primer objetivo específico fue analizar y sintetizar la información publicada disponible sobre los métodos de detección de DIF centrando la atención en los estudios con datos simulados mediante una revisión de la producción científica de los estudios que pretendieron evaluar la eficacia o proponer nuevos métodos. Dicha revisión giró entorno a la descripción de los principales tópicos o líneas de investigación en este ámbito de estudio, la descripción de los últimos desarrollos y tendencias, y la descripción de las características de los estudios con datos simulados.

En segundo lugar, se pretendió analizar experimentalmente el funcionamiento de uno de los métodos más empleados para la detección de DIF mediante el análisis

del efecto de variables como la razón de tamaños de muestra y del modelo desajustado, es decir cuando los datos se simularon con un modelo 3PL, en condiciones que simula pruebas de aplicación masiva y de altas consecuencias.

En tercer lugar, se pretendió ilustrar la aplicación de procedimientos de desarrollo reciente para la detección de DIF en un conjunto de datos empíricos empleando covariables de diferente índole y escala de medición.

Teniendo en cuenta el contexto de base en el cual se desarrolló esta tesis doctoral y los objetivos planteados, se decidió estructurar una tesis como compendio de publicaciones con tres estudios. Cada uno de los estudios derivó en un artículo científico publicado en revistas de impacto.

El primer estudio proporcionó un abordaje de tipo histórico-contextual, se trató de una revisión sistemática que cubrió el primer objetivo. A partir de los hallazgos respecto a los métodos más empleados y los más recientes, se definieron las características particulares del segundo y tercer estudio. El segundo estudio proporcionó una mirada experimental, se trató de un estudio con datos simulados sobre el funcionamiento de un procedimiento de detección de DIF basado en el Modelo Rasch (diferencia del parámetro de dificultad) en el ámbito de las pruebas de aplicación masiva más comúnmente empleadas en educación. El tercer estudio aportó el abordaje empírico ilustrando la aplicación de dos procedimientos basados en el Modelo Rasch y que se han propuesto en la última década para la detección de DIF, estos procedimientos se aplicaron a un conjunto de datos obtenidos en el ámbito de la salud mental. En la figura 2.2 se ilustra gráficamente la estructura de la tesis.

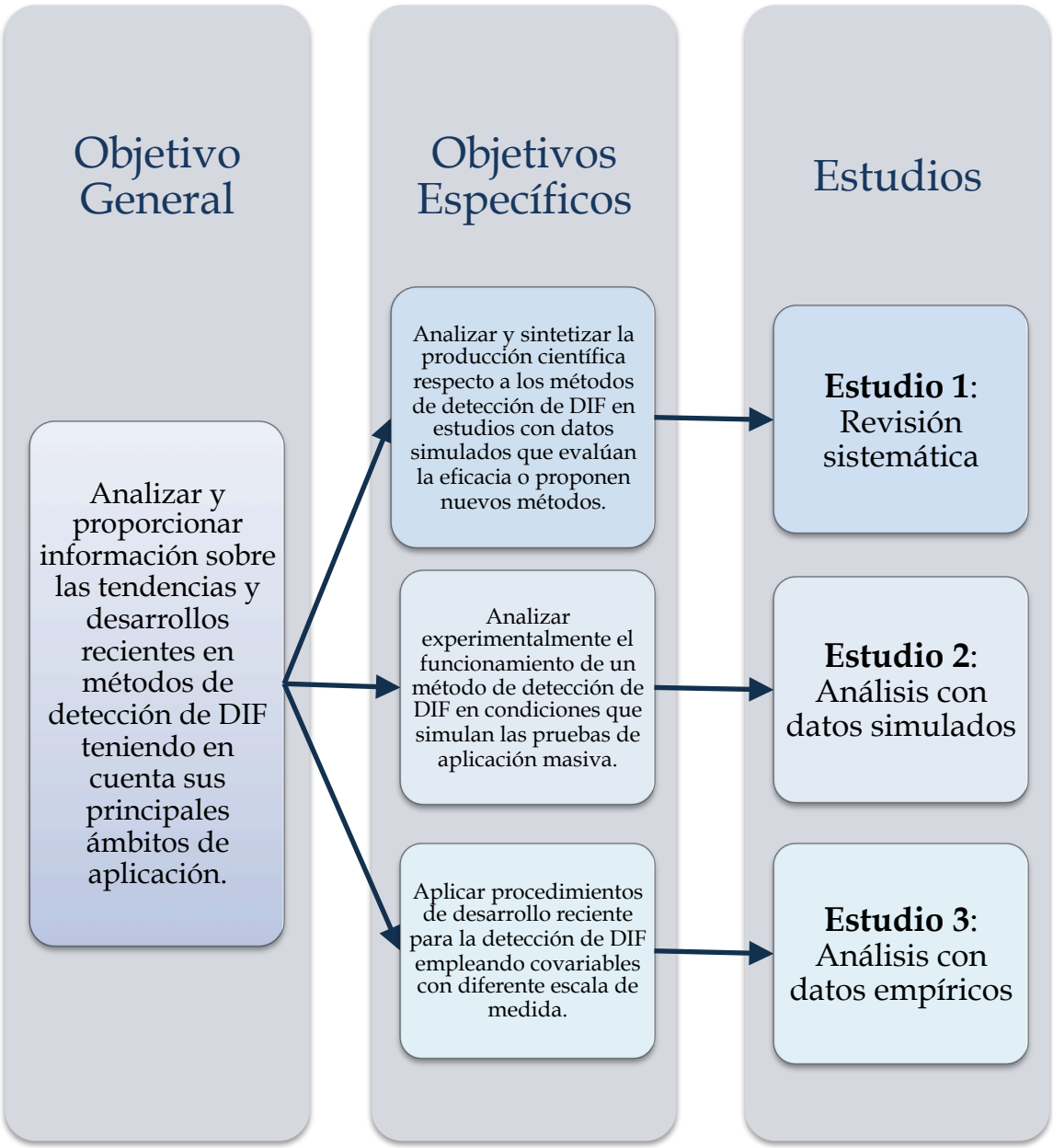


Figura 2.2. Estructura general de la tesis.

3. Resultados

A continuación se presentarán cada uno de los tres estudios que componen esta tesis doctoral. Tal y como se indicó anteriormente, cada uno de los estudios corresponden a un único objetivo específico y cada estudio produjo un manuscrito que fue publicado en revistas de alto impacto.

Para cada uno de los manuscritos se mostrará la siguiente información: en primer lugar los datos identificativos del manuscrito, es decir la revista en la que se ha publicado, el título, autores, datos de factor de impacto, rango y posición de la revista, y DOI del artículo. En segundo lugar se presentará un breve resumen de los principales resultados del estudio. En tercer lugar y dado que los tres manuscritos se han publicado, se anexará la versión postprint del manuscrito, es decir la versión aceptada previa al proceso de maquetación por parte de la revista.

3.1. Estudio 1: Revisión sistemática

3.1.1. Datos identificativos

ELSEVIER

Educational Research Review

Volume 31, November 2020, 100340

Developments and trends in research on methods of detecting differential item functioning

Ángela I. Berrío [a, b], Juana Gómez-Benito [a, b], Erika Margarita Arias-Patiño [a]

[a] Department of Social Psychology and Quantitative Psychology, University of Barcelona, Barcelona, Spain
[b] Group on Measurement Invariance and Analysis of Change (GEIMAC), Institute of Neuroscience, University of Barcelona, Barcelona, Spain

Título: Developments and trends in research on methods of detecting differential item functioning.

Autoras: Ángela I. Berrío, Juana Gómez-Benito & Erika Margarita Arias-Patiño

Año: 2020

Revista: Educational Research Review

Métrica según Journal Citation Reports para el año 2020:

- Factor de Impacto: 7,803
- Rango por factor de impacto: 4 de 264, Cuartil: Q1 (D1), Categoría: Education & Educational Research (SSCI).
- Rango por Journal Citation Indicator (JCI): 27 de 722, Cuartil: Q1 (D1), Categoría: Education & Educational Research (SSCI).

Material Suplementario: se puede consultar en el anexo 1.

DOI: https://doi.org/10.1016/j.edurev.2020.100340

3.1.2. Resumen

Con este estudio se pretendió analizar y sintetizar los desarrollos clásicos y actuales en métodos de detección de DIF cuando se aplican estudios con datos simulados. Dichos desarrollos se analizaron a través de tres aspectos: la identificación de los tópicos en torno de los cuales los investigadores han centrado su atención en este campo de estudio, la descripción de los métodos más estudiados y los que contienen propuestas más recientes y finalmente, la descripción de las variables simuladas en los estudios y sus valores. Complementario a esto se proporcionó información sobre patrones de colaboración entre autores y revistas científicas. Teniendo en cuenta lo anterior, las preguntas que guiaron este estudio fueron: ¿qué métodos de detección de DIF se han analizado a través de datos simulados? Además de métodos como MH, RL y SIBTEST, ¿cuáles han sido los más estudiados? ¿Cómo se han diseñado los estudios con datos simulados? ¿Qué variables y cuáles valores son los más comunes? ¿Qué tópicos de investigación diferentes se pueden encontrar en el estudio de los métodos de detección de DIF con datos simulados?

Para dar respuesta a estas preguntas y cumplir con el objetivo trazado se diseñó una revisión sistemática que siguió la guía PRISMA (Preferred reporting items for systematic review and meta-analyses; Moher et al., 2010). La búsqueda de la literatura se centró en artículos originales revisados por pares cuyo objeto de estudio fue proponer o analizar el funcionamiento de uno o varios métodos de detección de DIF usando datos simulados. Dicha búsqueda se llevó a cabo en las bases de datos Web of Science (WoS) de Clarivate Analytics, PsycInfo de la Asociación Americana de Psicología (APA) y Education Resources Information Center (ERIC) del Instituto de Ciencias de la Educación. Para el análisis de los datos se emplearon técnicas de visualización de similaridades para construir los mapas de patrones de coocurrencias y de colaboración mediante el software VOSViewer 1.6.9 (Van Eck & Waltman, 2018) y análisis descriptivos de las variables codificadas.

291 artículos originales se incluyeron en este estudio. Se encontraron once tópicos de investigación en torno de los cuales los investigadores han centrado su interés y que pueden categorizarse en dos grandes grupos:

1. Tópicos centrados en métodos de detección de DIF:
   a) Detección de DIF mediante la comparación de modelos (primer cluster).
   b) Métodos paramétricos clásicos para detectar DIF no direccional (segundo cluster).
   c) SIBTEST (sexto cluster).
   d) Modelo Rasch (séptimo cluster).
   e) Estadísticos para la clasificación de DIF y detección en Test Adaptativos Informatizados (CAT por sus siglas en inglés: Computarized Adaptative Test; octavo cluster).
   f) Funcionamiento diferencial del test (noveno cluster).
   g) Mantel Haenszel (decimo cluster).
2. Tópicos centrados en variables relacionadas con la evaluación y sesgo:
   a) Modelos y procedimientos relacionados con la agrupación de ítems (tercer cluster).
   b) Detección de DIF en ítems politómicos (quinto cluster)
   c) Modelos para respuesta múltiple e invarianza de la medición o sesgo (cuarto cluster).
   d) Estudios de sesgo (undécimo cluster).

Los resultados relacionados con los métodos de detección de DIF indicaron que los cinco más estudiados fueron MH, RL, SIBTEST, IRT-LR-DIF y los basados en el modelo Rasch, los cuales se analizaron en el 63,5% de los estudios incluidos. En la categoría de métodos basados en el modelo Rasch se incluyeron procedimientos como índice de ajuste basado en Rasch, modelo Rasch para testlet, índice de similaridad de información y diferencia del parámetro de dificultad. Por su parte, entre los cinco métodos estudiados más recientemente se encontraron las aplicaciones a modelos de diagnóstico cognitivo, delta plot, modelo jerárquico linear, modelo Rasch y MIMIC.

Finalmente, con relación a los valores de las variables que comúnmente se emplean en estudios con datos simulados cuando se analizan métodos de detección de DIF, se encontró que el número de réplicas más empleado fue 100 con un rango entre 1 y 20,000 réplicas. 1,000 fue el tamaño de muestra más frecuente y 500 participantes en cada uno de los grupos fue el tamaño de muestra de grupo más frecuente. En la mitad de las condiciones experimentales simuladas los tamaños de muestra de los grupos fueron iguales y en la otra mitad no lo fueron. Entre estas últimas, la razón de tamaños de muestra más frecuente fue 2, 3, 4 o 5. Las razones de tamaño de grupos más grandes empleadas en los estudios con datos simulados han sido 26.7, 30, 40, 50 y 100. La longitud del test mostró una tendencia cada vez más reciente a considerar test con menos ítems. Los ítems dicotómicos son los que se han simulado más frecuentemente en las condiciones experimentales y entre los ítems politómicos lo más frecuente ha sido el uso de cinco categorías de respuesta. Sólo un 21.7% de las condiciones simularon diferencias entre grupos en la distribución del nivel de atributo, 13.4% emplearon procedimientos de purificación del atributo medido y 14.5% emplearon medidas de tamaño del efecto para estimar la magnitud de DIF.

3.1.3. Versión postprint

Developments and Trends in Research on Methods of Detecting Differential Item Functioning

Ángela I. Berrío[a, b, *], Juana Gómez-Benito[a, b], & Erika Margarita Arias[a]

[a] Department of Social Psychology and Quantitative Psychology, University of Barcelona, Barcelona, Spain
[b] Group on Measurement Invariance and Analysis of Change (GEIMAC), Institute of Neuroscience, University of Barcelona, Barcelona, Spain
[*] **Corresponding author.** Quantitative Psychology Unit, Department of Social Psychology and Quantitative Psychology, Faculty of Psychology, University of Barcelona, Passeig Vall d'Hebron 171, 08035 Barcelona, Spain. E-mail address: aiberriob@ub.edu (Á. I. Berrío)

Abstract

Differential item functioning (DIF) has been considered as an aspect of special relevance in assessment, mainly in educational assessment. After almost three decades of research by means of simulation studies on the detection of DIF, around a hundred methods and statistics have been proposed. This is the first systematic review of published studies that have analyzed DIF detection methods using simulated data. The primary goal was to provide information about the overall structure of the field, its main developments, and current trends. A total of 1400 articles were identified, of which 291 were finally included. The results showed a growth in the number of experimental studies, a field of study that is organized around 11 lines of research, and a greater output of articles involving methods such as the Mantel-Haenszel, logistic regression, SIBTEST, Item Response Theory-likelihood ratio test for DIF, and methods based on the Rasch model. In discussing the results, the paper focuses particularly on trends in experimental DIF studies and the latest developments in terms of methods.

Keywords: DIF, DIF methods, simulated data, systematic review, distance-based maps.

1. Introduction

Measuring constructs by means of tests is a common practice in processes of assessment. In empirical studies the reliability of a test is usually assessed by calculating Cronbach's alpha coefficient, while its validity is examined by analyzing the association between test scores and external criteria. All this is necessary but not sufficient to support the psychometric quality of a test and the possibility of making valid inferences from the scores obtained by examinees, which is the objective of any assessment. More specifically, any test that is used for such a purpose must permit an adequate assessment of the construct being measured and an adequate comparison of its results between observations, which can be analyzed by studying invariance in the instrument's metric properties and, more particularly, by examining differential item functioning (DIF).

The concept of DIF, proposed by Holland and Thayer (1988), refers to the phenomenon whereby two examinees with the same ability level but from different groups have a different probability of responding correctly to an item. The presence of DIF may produce invalid test results, preventing comparisons between populations and demonstrating that the instrument does not provide an adequate measure of the target construct.

The interest in DIF in the context of differences between groups made it necessary to distinguish the concept from others such as impact and bias. Whereas DIF always refers to the situation whereby examinees who are matched on ability level have a different probability of answering an item correctly, impact is defined as a difference in means on the latent trait or ability, such that groups or examinees are classified as more and less able. As for bias, this refers to an informed judgment about the behavior of an item that takes into account the purpose of the test, the relevant experiences of certain subgroups of examinees, and statistical information about the item (Holland & Wainer, 1993). In short, if we find a difference between examinees in their scores on a given item, we should ask ourselves what this difference indicates. If it merely reflects a difference in skill level, in other words, that a given examinee or group of examinees is genuinely more or less able on the aspect being measured by the item, then we are dealing with impact. However, if two examinees with the same level of ability differ in their probability of endorsing the item, then the difference is the result of measurement error and we speak of DIF. If DIF exists, the researcher can analyze bias by examining the reasons why the item leads to different responses in different examinees or groups. This is done by taking into account the purpose of the test, the experience of respondents, and statistical information regarding item analysis or response categories, among other aspects.

During more than 50 years of research into DIF, various methods have been proposed for detecting it. The early methods sought to identify biased items by means of analysis of variance (Angoff & Sharon, 1974; Cardall & Coffman, 1964; Cleary & Hilton, 1968), but these were soon superseded by other more precise or robust techniques. The first generation of methods were applied to dichotomous items and unidimensional tests with the aim of detecting uniform DIF, but the need to analyze DIF in different instrument formats led to the development of new techniques or modifications of existing ones so as to take into account, for example, the presence of polytomous items and non-uniform DIF.

The wide range of methods for detecting DIF has led some authors to classify them according to their main characteristics. Noteworthy examples of classification schemes are those of Camilli and Shepard (1994) and Potenza and Dorans (1995). The former groups detection methods into three categories: those based on analysis of variance and classical test theory, those based on item response theory (IRT), and those based on the analysis of contingency tables. The latter classification scheme takes into account the type of item (dichotomous for those with two response

options versus polytomous for items with more than two possible responses), the matching criterion for the groups (observed score for procedures in which the raw score obtained in a test is used to determine which subjects have the same level on the measured construct versus the latent variable for procedures in which the construct of interest is not directly measurable and is therefore estimated through indicators), and the model used to establish the relationship between item score and the matching variable (parametric, non-parametric). Hidalgo and Gómez-Benito (2010) proposed an updated classification of DIF detection methods that was based on the scheme of Potenza and Dorans (see Table 1). It can be seen in the table that the standardization method, the Mantel-Haenszel (MH) statistic, logistic regression (LR), log-linear models, and latent class models are based on the observed score. Here it is worth noting that Millsap and Everson (1993) suggested that observed score is an inadequate indicator as a substitute for latent trait and that DIF analysis should not be based on the observed scores unless the Rasch model fits. Returning to Table 1, the SIBTEST family of methods corresponds to the latent trait and the non-parametric model, whereas under the parametric model we find a variety of methods: those based on IRT, including area measures, Lord's chi-squared, the IRT-likelihood ratio test for DIF (IRT-LR-DIF), and those based on factor analysis and structural equation modeling, for example, the multiple indicators multiple causes (MIMIC) model.

Table 1

Classification of DIF detection methods.

Model	Observed score		Latent trait	
	Dichotomous	Polytomous	Dichotomous	Polytomous
Non-parametric	Mantel-Haenszel (MH) Standardization	MH – Generalized MH – Ordinal Standardization	SIBTEST CATSIB	POLYSIBTEST
Parametric	Logistic regression	Multinomial logistic regression	IRT log-likelihood ratio	IRT log-likelihood ratio
	Log-linear models	Log-linear models	Lord's chi-squared statistic	Lord's chi-squared statistic
	Latent class models	Latent class models	Area measures	Area measures
		Discriminant logistic regression	IRT: Test of differential item functioning	IRT: Test of differential item functioning
			Confirmatory factor analysis	Confirmatory factor analysis
			Multiple indicators multiple causes (MIMIC) model	Multiple indicators multiple causes (MIMIC) model

DIF: Differential item functioning; IRT: Item response theory

One aspect to take into account in IRT-based methods is the model used for item analysis: the Rasch model, or one-, two-, or three parameter models. In essence, the Rasch and one-parameter models state that the probability of responding correctly to an item is a function of the relationship between the item's difficulty and the person's ability level, and that the item maintains

a constant discrimination parameter. As for the two-parameter model, this takes into account not only the difficulty but also the discrimination of the item, while the three-parameter model also incorporates the pseudo-guessing parameter.

In general, any DIF detection method uses as its matching variable the IRT ability estimates or the total test score, which may be contaminated by the presence of DIF. One way of mitigating this is to apply purification procedures, whereby the items flagged as showing DIF when first applying the detection method are then excluded from the calculation of the matching variable in a second application. This is referred to as two-stage purification (Holland & Thayer, 1988). An alternative is what is known as iterative purification (French & Maller, 2007), consisting of the successive application of the DIF detection method and the exclusion of items showing DIF from the calculation of the matching variable until the same items are detected in two consecutive iterations. Several studies have demonstrated that DIF detection methods perform better when they include purification stages (Clauser, Mazor, & Hambleton, 1993; Fidalgo, Mellenbergh, & Muñiz, 2000; French & Maller, 2007; Miller & Oshima, 1992; Navas-Ara & Gómez-Benito, 2002; Park & Lautenschlager, 1990; Wang & Su, 2004).

Another issue to consider when applying DIF detection methods concerns the use of statistics based on differences in means, which may be affected by sample size. Ways of addressing this include measures of effect size or metrics that, when combined with the statistical significance, allow the magnitude of DIF to be classified. An example of such a measure is the Mantel-Haenszel delta metric (Holland & Thayer, 1985).

The current interest in DIF is due, in part, to the enormous social implications that the presence of DIF in a test may have, not only for individual examinees but also at the broader level of public policy making and educational quality assessment. In addition, the analysis of DIF is an important source of information that impacts on all the agreed sources of validity evidence in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In this respect, Gómez-Benito, Sireci, Padilla, Hidalgo, and Benitez (2018) have argued that a DIF study is by definition a study of validity since it can provide information about all five sources of validity evidence. For example, a DIF study will analyze whether the DIF is associated with factors that are relevant or irrelevant to the construct, or whether there are negative consequences associated with the presence of DIF, and so on.

Given that developments in research on DIF are linked to the identification of the conditions under which detection methods function adequately, the use of simulated data has become very popular in this field as it allows researchers to assess in a more systematic and controlled way whether or not certain variables have an effect on the statistical properties of DIF methods. The present study aims to provide a critical review of current developments in DIF

detection methods and trends in research using simulated data. In a previous study, Gómez-Benito, Hidalgo, Guilera, and Moreno (2005) offered a bibliometric analysis of general scientific activity in the field of DIF, and two subsequent studies focused specifically on the Mantel-Haenszel method. One of these took a bibliometric perspective (Guilera, Gómez-Benito, & Hidalgo, 2009), while the other involved a meta-analysis of Type I error and power of the Mantel-Haenszel statistic (Guilera, Gómez-Benito, Hidalgo, & Sánchez-Meca, 2013). To our knowledge, however, no previous studies have analyzed the whole of the literature involving the use of simulated data to examine DIF detection methods. By doing so, this study provides a broader overview of simulated methodological studies on DIF methods and reveals trends within the field.

2. Method

2.1. Information Sources

Articles were obtained in accordance with the PRISMA guidelines (Preferred reporting items for systematic review and meta-analyses; Moher, Liberati, Tetzlaff, Altman & The PRISMA Group, 2010) and through a search of the following online scientific databases: Web of Science (WoS), maintained by Clarivate Analytics; PsycInfo, produced by the American Psychological Association; and Education Resources Information Center (ERIC), sponsored by the Institute of Education Sciences. In WoS we consulted the Medline database and the core collection. These databases were used as they contain extensive information for the main fields in which DIF has been studied, namely education, psychology, and the clinical area. It is important to point out that in education scope is where the biggest developments have taken place. This is mainly reflected in the developments derived from research carried out by institutions such as the Educational Testing Service (ETS), the European Educational Research Association (EERA), and the National Assessment Program (NAP), as well as in large-scale national and international educational assessment initiatives (e.g., PISA - Programme for International Student Assessment, and TIMSS - Third International Mathematics and Science Study, among others). All of this has contributed to the development of and research into methods for detecting DIF, as well as to their application and importance for cross-cultural assessments at the national and international level.

2.2. Search Strategy

Potentially relevant articles were identified through a systematic search that used the following keywords and Boolean operators in the topic or theme: DIF OR "differential item functioning" AND Simulat*. The search covered the period from 1990 to April 6, 2018. The results of this search were complemented with studies on measurement invariance.

2.3. Criteria for Inclusion and Exclusion of Articles

After searching the aforementioned databases we first eliminated any retrieved documents that were duplicates or which were not original peer-reviewed articles (dissertations, books, book chapters, among others). A second exclusion stage involving screening of content was then

required because the term DIF also appears in areas of knowledge that are unrelated to educational and psychological measurement and related fields. During this content review we also identified those articles whose aim was to test, propose, or compare DIF detection methods using simulated data. This was done by two experts who independently classified a random sample of 50 articles, assigning them to one of two categories (include or exclude). Once inter-rater agreement of 90% was achieved, one of these experts classified the remaining articles. In the subsequent coding stage we included solely those articles written in English or Spanish, as coding requires fluency in the article's language.

2.4. Data Collection

For each article we coded the following variables: number of replications, sample size, sample size ratio, test length, type of items, type of DIF, method used to detect DIF, use of purification, use of effect size measures, Type I error, and power. Two independent coders recorded data for 20% of the articles and achieved 97.5% agreement (95% CI [96.1, 98.8]). The remaining articles were then coded by one of the raters.

2.5. Data Analysis

Because we were interested in identifying the main developments and trends in research on DIF, we used techniques for visualizing similarities, constructing distance-based maps (Van Eck & Waltman, 2010) using VOSViewer 1.6.9 (Van Eck & Waltman, 2018). This software produces visual maps that establish relationships between different items (documents, authors, journals), yielding three important aspects for analysis: The distance between items, which reflects the magnitude of the relationship between them (the smaller the distance, the stronger the relationship); the grouping of items to form a cluster (distinguished by an assigned color), which represents a research topic; and the size of the item, which indicates its relative importance in the field of study, depending on the type of analysis conducted (Sassetti, Marzi, Cavaliere, & Ciappei, 2018). For readers interested in learning more about the technique, we recommend (in addition to the articles cited above) the paper by Van Eck and Waltman (2007).

In order to observe the different lines of research, we conducted a co-occurrence analysis of author keywords, as these ensure that the results are more representative of the content of articles and of the concepts that authors have considered important (Stuart, 2018). Before constructing a co-occurrence map of author keywords we first had to filter terms in order to identify those referring to the same concept, for example, DIF and differential item functioning; item response theory and IRT; Mantel-Haenszel, Mantel Haenszel, and Mantel-Haenszel procedure; CAT and computerized adaptive test, among others. Of the 365 concepts obtained, we included in the analysis the 120 with the greatest link strength so as to ensure that we counted not only the most frequently occurring terms but also those which, despite occurring less frequently, were likely to be relevant for identifying current research trends. The map was constructed using

the full counting method and a cluster resolution of 1.00, with a minimum number of 7 words per cluster.

We also analyzed journal and author collaboration patterns using the bibliographic coupling approach in order to produce distance-based maps for each. This approach (bibliographic coupling) was used as it is the one best able to represent current fronts in a research field, as opposed to a historical taxonomy (Boyack & Klavans, 2010; Klavans & Boyack, 2017). It should be noted that when bibliographic coupling is applied, the degree to which two articles are related depends on the similarity of the references cited in each one.

In order to characterize research on DIF detection methods using simulated data, we carried out descriptive analyses of the variables coded for each of the articles included.

3. Results

Of the 1400 documents retrieved, 519 were duplicates and 168 were eliminated as they were not original journal articles (i.e., book chapters, dissertations or other types of document). After reviewing the abstract or full text of the remaining articles we eliminated a further 422 records because they were not related to the topic of interest: either they were focused solely on the simulation of computerized adaptive tests or they did not study methods of DIF detection. A total of 291 articles were therefore included in the review. Fig. 1 shows a PRISMA flow chart for the four stages of study inclusion and exclusion (Moher et al., 2010).

3.1. Study Characteristics

Of the 291 articles analyzed, 59 (20.3%) included both empirical and simulated data, whereas the remaining 232 studies (79.7%) only used simulated data. Regarding article production by year, there was an increasing trend across decades: 20.6% of the articles were published between 1990 and 1999, 31.6% between 2000 and 2009, and 47.8% between 2010 and April 6, 2018. The years with the highest number of articles published on DIF detection methods were 2009 and 2015 (with 20 articles each), followed by the years 2011, 2012, and 2013 (with 19 articles each).

The journals that published the highest numbers of articles on DIF detection methods were *Applied Psychological Measurement* and *Educational and Psychological Measurement* (56 each), followed by the *Journal of Educational Measurement* (37 articles), the *Journal of Educational and Behavioral Statistics* (20 articles), *Applied Measurement in Education* (16 articles), and *Psicothema* and *Psychometrika* (11 articles each). Table 2 shows the number of articles published by journal for each of the journals that have published at least two papers.

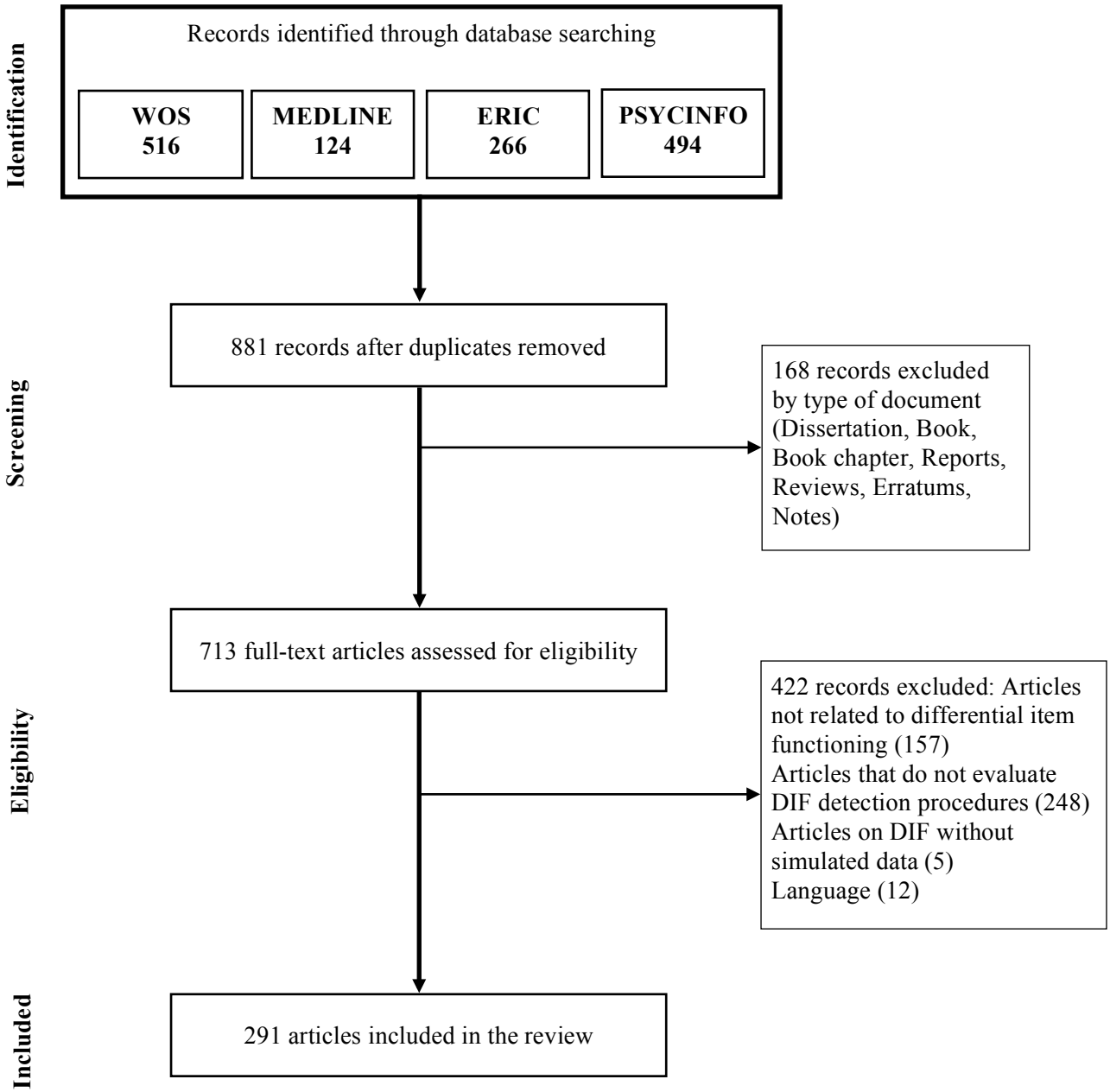


Fig. 1. PRISMA flow chart for the process of study inclusion and exclusion.

Table 2

Number of articles published in each journal.

Name of Journal	No. of articles
Applied Psychological Measurement	56
Educational and Psychological Measurement	56
Journal of Educational Measurement	37
Journal of Educational and Behavioral Statistics	20
Applied Measurement in Education	16
Psicothema	11
Psychometrika	11
International Journal of Testing	8
Quality & Quantity	7
Structural Equation Modeling: A Multidisciplinary Journal	6
Psicológica	5
Alberta Journal of Educational Research	4
Journal of Experimental Education	4
British Journal of Mathematical & Statistical Psychology	3
Journal of Applied Measurement	3
Methodology: European Journal of Research Methods for the Behavioral and Social Sciences	3
Multivariate Behavioral Research	3
Computational and Mathematical Methods in Medicine	2
European Journal of Psychological Assessment	2
Journal of Applied Psychology	2
Methods of Psychological Research	2
Organizational Research Methods	2

3.2. Main Topics and Lines of Research

A total of 120 concepts were included in this analysis, and the distribution of concepts by clusters is shown in Fig. 2.

A total of 11 different lines of research or topics were obtained. The first line of research (red cluster), which we labeled "Detection of DIF through model comparison", included the methods IRT-LR-DIF, confirmatory factor analysis, multiple indicators multiple causes, and, of less relevance, ordinal logistic regression. There were also other terms related to these methods, with two of the most important being scale purification and measurement equivalence.

The second research line (green cluster) was labeled "Classical parametric methods of detecting non-directional DIF" and included methods such as logistic regression, the log-linear model, and different area measures, with associated concepts including Monte Carlo simulation and non-directional DIF.

The third topic (blue cluster) was defined as "Models and procedures related to item grouping" and involved the following concepts: testlets, linking, generalized graded unfolding model, and unfolding, as well as concepts of less relevance, including POLYSIBTEST, single-peaked preference functions, and multi-group bifactor model.

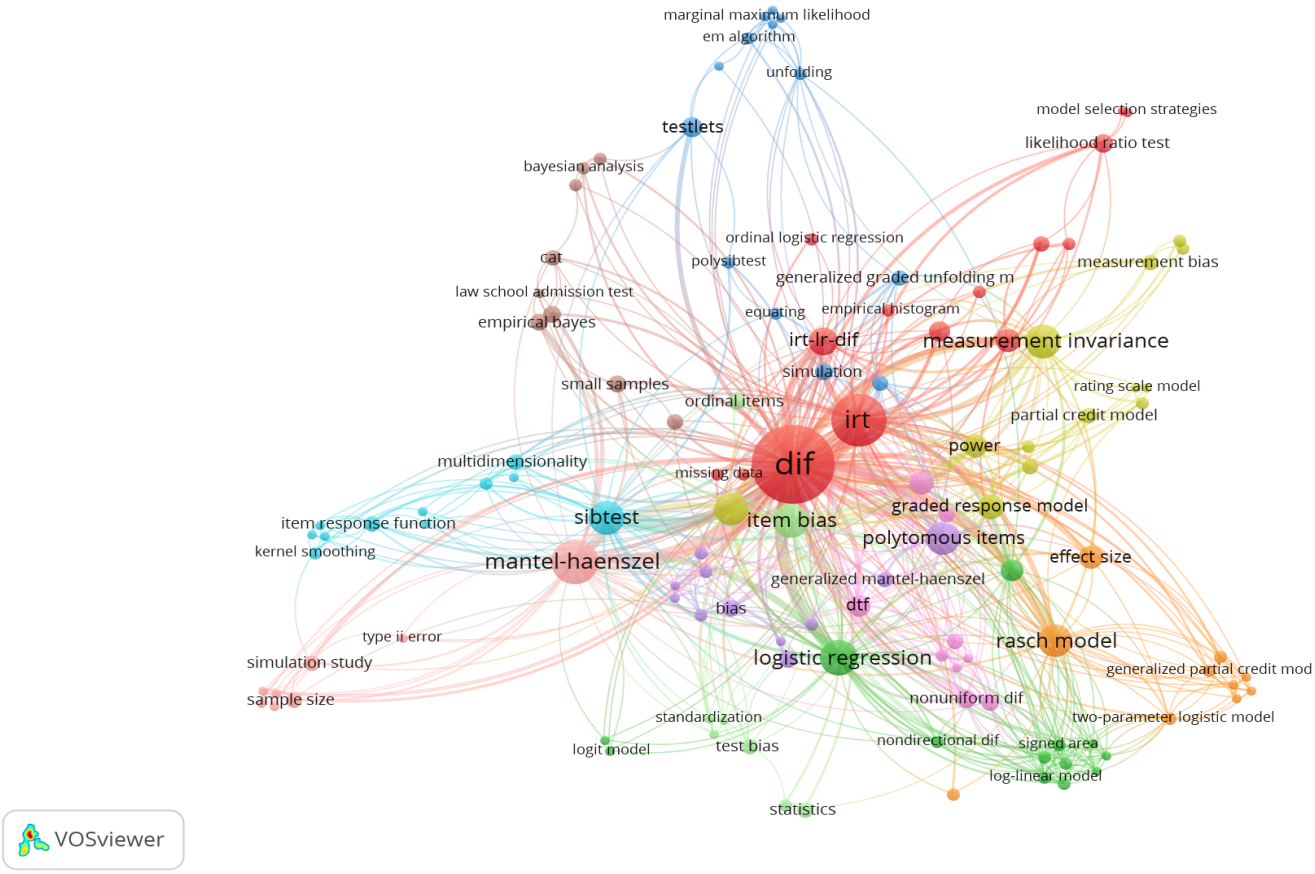


Fig. 2. Keyword co-occurrence map.

The fourth topic (yellow cluster) was labeled "Multiple response models and measurement invariance or bias" and comprised the graded response model, the partial credit model, and model-based recursive partitioning, among the most relevant. It also contained terms related to measurement adequacy such as Type I error and power.

The fifth line of research (purple cluster) was "Detecting DIF in polytomous items" and the main terms it included were polytomous items, DIF detection, generalized Mantel-Haenszel, bias, unidirectional DIF, and crossing DIF.

The sixth topic (turquoise cluster) corresponded to "Simultaneous item bias procedures" and contained concepts including SIBTEST, item response function, multidimensionality, and regression correction.

The seventh topic (orange cluster) referred to studies related to the "Rasch model" and included terms such as Rasch model, two-parameter logistic model, effect size, generalized partial credit model, and model fit, among others.

The eighth topic (brown cluster) was labeled "Statistics for classifying DIF and detection in computerized adaptive tests (CATs)" and contained keywords including empirical Bayes, loss function, and CAT.

The ninth line of research (pink cluster) referred to studies of "Differential test functioning" and included keywords such as DFIT, DTF, uniform DIF, nonuniform DIF, and three-parameter logistic model.

The tenth topic (salmon pink cluster) involved studies focusing on the "Mantel-Haenszel" procedure and contained the keywords Mantel-Haenszel, simulation study, and sample size, as well as more recent terms such as ability distribution, international assessment, thick matching, and thin matching.

Finally, the eleventh line of research (olive green cluster) referred to "Studies of bias" and included terms related to item bias and test bias. Supplementary material 1 lists all the author keywords for each cluster, along with their frequency of occurrence, strength, and mean year of use.

3.3. Journal and Author Collaboration Patterns

For those readers who are interested, supplementary material 2 offers a graphical illustration, based on network analysis, of journal and author collaboration patterns, showing their relative importance. In general, the journals formed four clusters: journals primarily related to psychology and social sciences, journals related to education, journals in mathematics and statistics, and journals with more general interest reflecting the multidisciplinarity.

Regarding author collaboration patterns, authors are grouped into four clusters: authors with an interest in methods based on item response theory (IRT) and the MIMIC method, authors who have mainly studied classical methods (Mantel-Haenszel and SIBTEST), and some applications of these in computerized adaptive test (CATs), authors with an interest in the application of methods to polytomous items, and finally, the interest in methods of multilevel data analysis.

3.4. DIF Detection Methods in Simulation Studies

The 291 articles reviewed corresponded to a total of 358 studies, since some papers reported two, three, or even four separate studies. It should be noted that the total frequency for some of the variables analyzed may be higher than the total number of studies, depending on the different conditions that were simulated.

3.4.1. Number of Replications

The number of replications performed ranged between 1 and 20000, with the most frequent values being 100 (37.5%), 1000 (15.2%), 500 (9.2%), 50 (7.1%), and 200 (5.7%). In order to examine changes over time we established three decade intervals: 1990 to 1999, 2000 to 2009, and 2010 until April 2018. In general terms, there were differences between the number of replications

performed by studies published in the first decade (μ = 140.80; 95% CI [92.45, 189.15]) compared with both the second (μ = 943.34; 95% CI [417.80, 1468.89]) and third decades (μ = 1044.02; 95% CI [690.21, 1397.83]), but not between those published in the latter two periods.

3.4.2. Sample Size

The sample size considered in the studies reviewed ranged widely between 2 and 40000 examinees. The most frequent sample sizes were 1000 (18.8% of simulated conditions), 2000 (13.9%), 500 (9.2%), and 600 and 1500 (both 4.6%). The overall mean sample size was 1751.36 examinees (SD = 3135.8). Analysis of changes over time in this variable revealed no differences between the three decades (μ = 1666.68; 95% CI [1444.95, 1888.40], μ = 1811.73; 95% CI [1344.58, 2278.89], μ = 1750.11; 95% CI [1475.84, 2024.38], respectively), although we did find that the range for this variable was lower during the first decade (100 to 8000) in comparison with the other two decades (range between 100 and 40000 in the second decade and between 2 and 40000 in the third).

Regarding differences in sample size between research fields (taking as a reference the four clusters obtained among the journals that publish this type of study), no significant differences were found ($F(3, 760) = 2.568$, $p = .053$, $\eta 2 = .010$). However, when analyzing the mean differences between the psychology and social sciences cluster versus the education cluster, significant differences were found ($t = -0.207$, $p = .0028$), with a larger sample size in education (μ = 1993.66) than in psychology (μ = 1416.25).

3.4.3. Sample Size Ratio

The mean size of the reference group and the focal group was, respectively, 952.3 (SD = 1573.03) and 694.4 (SD = 1379.1). The most frequent reference group sizes were 500 (25.3%), 1000 (23.3%), 250 (6.9%), 100 (4.9%), 200 (4.8%), and 1500 (4.0%), while for the focal group the most frequent values were 500 (23.4%), 1000 (14.7%), 250 (11.9%), 100 (9.9%), 200 (7.1%), and 50 (4.7%). With respect to the sample size ratio, 49.5% of simulated conditions involved groups of equal sizes, while the remaining 50.5% simulated differences in group size. In the large majority of the latter conditions the reference group was larger than the focal group, and in only 3.7% was the focal group the largest. Among conditions that simulated differences in group size the most common ratios of reference to focal group size were 2 (22.1%), 3 (11.8%), 4 (11.4%), and 5 (8.1%). Thus, when groups differed in size the most frequent scenario was a reference group twice as large as the focal group. It is also worth noting that the largest sample size ratios (i.e., when the simulated difference between groups was greatest) were 26.7 (0.4%), 30 (0.4%), 40 (1.1%), 50 (0.4%), and 100 (0.4%).

3.4.4. Test Length and Type of Items

The articles reviewed simulated both short and long tests. The former ranged between 2 and 15 items, while the latter comprised tests of between 200 and as many as 400 items. In terms of changes over time in this variable, more recent studies tended to consider shorter tests. Specifically, the mean test length in the first decade was 42.4 items (95% CI [36.3, 48.4]), compared with 32.1 items in the second (95% CI [26.4, 37.8]) and 29.3 items in the third decade (95% CI [25.9, 32.7]).

With regard to test length and the field of research (considering the four clusters resulting from the journals that publish this type of study; see Table 3), significant differences were found ($F_{(3, 460)} = 5.116$, $p = .002$, $\eta^2 = .032$). Specifically, the descriptive statistics show that the main differences were between the area of psychology and social sciences, the area of education, and the area of mathematics and statistics (Table 3). Although the minimum number of test items does not vary much across the different areas, the maximum number shows considerable variation.

Table 3.

Descriptive statistics for the variable 'test length' with respect to fields of research.

Journal clusters	Mean	SD	95% CI		Minimum	Maximum
			Lower bound	Upper bound		
Psychology and social science	26.85	15.95	24.21	29.50	3	80
Education	34.05	26.51	30.57	37.53	2	266
Mathematics and statistics	44.66	59.64	30.74	58.57	5	400
Multidisciplinarity	33.46	23.82	23.40	43.52	4	100

Regarding the type of items considered, 64.5% of simulated conditions involved dichotomous items, 25.5% polytomous items, and 0.5% other types of item (e.g., continuous items). The type of item simulated was not specified in 9.4% of conditions. Among simulated polytomous items, 46% contained five response categories, 37% four categories, 7% three, 3% seven, and 1% six. The number of item response categories was not specified in the remaining 6% of cases. When comparing the research fields according to the type of items used when studying DIF with simulated data, we found that the highest percentage of polytomous items (36.5%) corresponded to psychology and social science, as compared with education (24.9%), mathematics and statistics (6.3%), and multidisciplinarity (33.3). The simulation of dichotomous items was most commonly observed in the areas of mathematics and statistics (79.4%) and education (72.8%).

3.4.5. Type of DIF

In 8.7% of simulated conditions, DIF was not simulated, in 5.6% the type of DIF simulated was not reported, and in 85.8% a specified type of DIF was simulated. Among the latter conditions, 53.8% simulated uniform DIF, 35.5% both uniform and nonuniform DIF, 8.8% nonuniform DIF,

and 0.7% simulated DIF in the pseudoguessing parameter in addition to uniform and nonuniform DIF. Finally, 1.3% of conditions simulated DIF in a spurious dimension.

3.4.6. Impact

Regarding differences between groups in the distribution of the attribute, 26.5% of conditions did not simulate impact, 21.7% did simulate it, and 51.8% compared the presence and absence of impact.

3.4.7. Purification and Measures of Effect Size

Purification procedures were employed in 13.4% of the studies reviewed, and 14.5% used measures of effect size to estimate the magnitude of DIF. The number of studies using some kind of purification procedure showed a slight upward trend over the last eight years: 27.1% of these studies were published during the first decade, 31.3% in the second, and 41.7% in the most recent period considered. A similar pattern was observed for the use of effect size measures: 17.3% of studies that applied a measure of effect size were published during the first decade, 28.8% in the second, and 53.8% in the third.

3.4.8. Power and Type I Error

With regard to outcome variables, power was examined in 72.9% of the studies reviewed and Type I error in 79.9%. The other main measures of outcome used were Type II error or false negatives, recovery of DIF magnitude, bias in DIF estimates, means and standard deviations of the DIF estimate, standard error of the DIF estimate, root mean square error, rates of correct classification of non-DIF items, and the correlation between the magnitude of the estimated DIF and simulated DIF.

3.4.9. DIF Detection Methods

In order to facilitate the analysis of the large number of methods proposed for detecting DIF, while maintaining a certain level of specificity, we classified them into 23 categories. Our aim in doing so was to group the methods into broad families that would highlight in each case the core procedure used to detect DIF. For example, we classified as Mantel-Haenszel methods all such procedures regardless of how they were specifically referred to by researchers, for instance, classical Mantel-Haenszel or standard Mantel-Haenszel. Each of the categories also included all the statistics that have been associated with the DIF detection method in question. Thus, the Mantel-Haenszel category, for example, also includes the chi-square statistic and odds ratio, as well as the continuity correction for Mantel-Haenszel. We also included new developments associated with each method, for example, the generalized Mantel-Haenszel and the Mantel test. In the specific case of methods based on IRT models, we distinguish between the most widely known and oft-cited (e.g., those based on the Rasch model, Lord's χ^2, IRT-LR-DIF, and differences in the difficulty parameter) and the remaining IRT-based methods and statistics that aim to detect DIF on the basis of other IRT parameters. This latter category also includes some of the most recent

proposals. Table 4 shows a description of the 22 categories, along with the percentage of corresponding studies.

Table 4

Categories for classifying DIF detection methods.

Category	Description	%
Mantel-Haenszel and related methods	This category includes all those methods that estimate odds ratios based on the Mantel-Haenszel procedure, the Cochran Mantel-Haenszel test, and the Liu-Agresti estimator, among others, and for both binary and polytomous items.	25.8
Logistic regression	Logistic regression and its various adaptations (related methods) for working with dichotomous and polytomous items, and using simple or multiple grouping variables (hierarchical models). Includes: ordinal logistic regression, hierarchical ordinal logistic regression, discriminant logistic regression, generalized logistic regression, binary logistic regression, logistic regression adapted for CAT, multinomial logistic regression, and polytomous logistic regression.	13.1
SIBTEST-based methods	Includes SIBTEST and SIBTEST with corrections and all its derivations: Crossing SIBTEST, SIBTEST modified for multiple booklets, POLYSIBTEST, CATSIB, and MULTISIB.	10.5
IRT-LR-DIF	Includes the IRT-LR-DIF procedure and some of its variants such as free baseline IRT-LR-DIF, graded response model IRT-LR-DIF, and IRT-LR-DIF with or without data imputation.	7.1
Methods based on the Rasch model	Includes all methods based on the Rasch model, such as tree analysis, mixture Rasch model, Rasch DIF model, Rasch fit index, Rasch testlet model, and the information similarity index (ISI).	7.0
Other methods based on IRT	IRT-based methods other than the classical and most well-known approaches are classified here. These include semi-parametric IRT-based models (Cochran's Z and kernel smoothing), and those that require model estimation but without including the interest parameter, (Lagrange multiplier test). Also included are methods that establish differences between the item parameters, the IRT-DIF model, the mixture IRT model, analysis of residuals derived from parameter differences, multi-group IRT-DIF, the iterative logit model, other methods that compare models, multidimensional methods, IRT with covariates, and the fixed effects IRT model.	6.3
SEM models	This category includes mean and covariance structure methods (MACS), confirmatory factor analysis, restricted factor analysis, multilevel confirmatory factor analysis, and multi-group extension of the aforementioned methods.	5.2
Lord's χ^2	Includes Lord's χ^2 and some of its minor derivations, such as Lord's χ^2 with Wald-1 statistics, with Wald-2 statistics, and with the iterative Wald approach.	4.2
DTF and DFIT methods	Includes all methods for detecting differential test functioning and differential functioning of items and tests.	3.7
Bayesian approach	Methods for detecting DIF based on the Bayesian approach, for example, application of the empirical Bayes method and subsequent distribution of model parameters.	3.3

Category	Description	%
MIMIC	Includes the multiple indicators multiple causes method and its variants for multidimensional data, with the iterative approach, with pure anchor, and with scale purification.	3.0
Area measures	Includes all methods for comparing areas between item characteristic curves: exact unsigned area, exact signed area, SOS1, SOS3, standardized unsigned area, and standardized signed area, among others.	2.4
Standardized method	This refers to the classical standardized mean difference (SMD) procedure.	1.9
Odds ratio approach	Refers to methods based on the odds ratio for detecting DIF using the nested logit model, nominal response model, applications to differential distractor functioning, and the Breslow-Day method.	1.9
Log-linear model	Refers to the method based on log-linear models.	0.9
Hierarchical linear models	Includes methods related to multilevel analysis, the original hierarchical generalized linear procedure, the polytomous hierarchical generalized linear model, and the bifactor mixture IRT-DIF model.	0.9
Difficulty difference method	Refers to the method of detecting DIF based on the difference between the difficulty parameters, under the IRT model.	0.7
Cognitive diagnosis model	Refers to proposed methods for detecting DIF based on the cognitive diagnosis model.	0.7
Empirical histogram method	This category includes the method based on histograms, which has been proposed as a variant of the IRT-LR-DIF approach.	0.5
ANOVA	Refers to the classical approach to detecting DIF by means of ANOVA.	0.3
Delta plot	Includes the classical delta plot method and its recent modifications.	0.3
Other DIF methods	Includes methods based on partial correlations and the application of other statistical indices that do not correspond to any of the other categories.	0.3

Analysis of the use of DIF detection methods by category over time showed that the methods with a most recent mean year of studies were the cognitive diagnosis model (CDM; mean year of publication 2014.50), delta plot (2012.50), hierarchical linear models (HLM; 2012), the Rasch model (2011.65), MIMIC (2011.35), and the empirical histogram (EMH) method (2010). Table 5 shows mean values for some of the variables of interest in relation to these more recently used methods. It can be seen that the highest mean number of replications corresponds to the delta plot, and the lowest to HLM. The highest mean values for sample size, size of reference group, and size of focal group all correspond to the Rasch method, whereas the mean sample size ratio is highest for the MIMIC method, followed by the Rasch model. As regards test length, the highest mean value corresponds to the delta plot, and the lowest to the MIMIC method.

Table 5

General characteristics of the studies that employed more recent methods.

	CDM	Delta plot	HLM	Rasch	MIMIC	EMH
No. replications	750	1000	72.73	729.48	251.09	137.50
Sample size	2038.46	333.33	2240.91	4369.71	1033.09	2000
Reference group	916.67	216.67	1318.18	2719.88	740.11	1500
Focal group	750	116.67	922.73	2356.52	287.59	500
Sample size ratio	1.24	2.58	1.70	3.84	7.10	3
Test length	36.15	40	25.27	30.37	17.59	23.25

Our analysis also showed that the CDM method has mainly been studied in China and the USA using dichotomous items and, in the majority of cases, simulating both uniform and nonuniform DIF. Recent examinations of the delta plot approach have been carried out in Belgium with dichotomous items and uniform DIF. Studies on the HLM method have been conducted in Turkey, China, and the USA and have analyzed uniform DIF, mostly with dichotomous items, and to a lesser extent with three-category polytomous items. The most recent studies involving the Rasch method have been carried out in Germany, Austria, Switzerland, China, the USA, and Spain, primarily with dichotomous items, followed by four-category polytomous items, with uniform DIF being the most widely studied type. Studies using the MIMIC method have been conducted in Canada, Iran, the USA, and China, mostly with dichotomous items, followed by polytomous items with five, three, or seven categories; in this case, both uniform and nonuniform DIF have been widely studied, followed to a lesser extent by DIF produced by a spurious ability in multidimensional environments. Finally, studies involving the EMH method have been conducted in the USA with dichotomous items and five-category polytomous items, and simulating both uniform and non-uniform DIF.

In the particular case of IRT models, it is worth highlighting the recent use of statistics such as the Lagrange multiplier test (mean year 2013.20, range 1998-2018) and others including kernel smoothing and Cochran's Z (mean year 2009.75, range 2006-2013), primarily by researchers in the USA, UK, and Canada.

4. Discussion

The results of this study are of value both to researchers and users of DIF detection methods. The article is the first to present a compendium of simulation studies on proposed methods for detecting DIF, doing so by means of an exhaustive literature review and the application both of techniques based on the visualization of similarities and quantitative statistical analysis. This approach has enabled us to provide a general overview of the field by considering its

main lines of research, highlighting in the process the ongoing interest in DIF analysis, both from a more methodological perspective and through methods that increasingly seek to reflect real-life assessment scenarios; the latter include methods which are less rigid in the definition of comparison groups (e.g., those based on the Rasch trees) and those that are more inclusive in terms of the data structure (e.g., methods which include a multilevel structure and those which can address categorical or ordinal variables, or small samples).

As a result of technological advances the use of simulation studies has become further consolidated in the field of DIF research, with new methods being proposed or improvements being made to existing techniques, and this has allowed researchers to address aspects that had not previously been studied due to computational difficulties. This is clearly reflected in the number of articles published per year: since 2009, an average of 17 papers have been published each year, compared with just six articles per year during the 1990s and eight per year between 2000 and 2008. This is consistent with a point made by Bauer (2017), who notes how the empirical relevance of this field has grown in tandem with a notable increase in scientific output between 1990 and 2014. In addition, we detected a core network of collaboration between journals that primarily focus on topics related to education, psychology, statistics, and mathematics, and a broader network that includes multidisciplinary areas of the social sciences. Another notable characteristic of the journals that have published the most articles in this field of study is their low impact factor, something which is common to methodological research in many fields. Here the top seven journals in terms of the number of publications have an impact factor ranging between 0.938 and 2.743.

The author collaboration patterns revealed four networks: Classical DIF detection methods, those based on IRT and the MIMIC method, methods for analyzing polytomous items, and the incorporation of multilevel data. This indicates how the pattern of collaboration is due not only to researchers' interest in particular methods but also to the characteristics of assessment or the instruments used.

Our analysis has also identified the latest trends in research on DIF detection using simulated data, as well as the latest developments in terms of methods. The trend over the past year has been towards a greater number of replications, the simulation of ever shorter tests, and the use of samples with as few as 2, 5, or 50 examinees, or as many as 20000 or even 40000. This shows how simulated data studies are becoming increasingly complex as a result of advances in hardware and software, and that they are seeking to address the realities of empirical contexts. For example, empirical assessment in the clinical and psychological field requires short tests, and it is not always possible to recruit large numbers of participants when validating a new instrument. In the field of education, by contrast, national and international assessment systems are increasingly common, and hence samples are large and tests generally have, on average, between 25 and 40 items. The latter is

confirmed by the results obtained regarding the differences in test length and simulated sample size between research fields such as psychology and education.

Our findings regarding the number of replications, a key aspect of simulation studies, show a rising trend. During the first decade we considered, the few guidelines that were available encouraged researchers to use a minimum number of replications. For example, the recommendation in the study by Harwell, Stone, Hsu, and Kirisci (1996) was to use at least 25 replications, with a maximum of 100 replications being considered adequate when the sample and test length were sufficiently large. A common feature of almost all the studies published during the first two decades we reviewed, and especially those from the 1990s, was that authors did not give a reason for the number of replications they chose. Subsequently, researchers used and continue to use a practical criterion that, as a rule, is based on the computational time required (Feinberg & Rubright, 2016), and this, coupled with technological advances, likely led to an increase in the number of replications performed during the second and third decades. Another important aspect to consider during this period is the criteria established by journal peer reviewers. It should also be noted, however, that in recent years researchers in psychometrics have shown increasing interest in studies that use simulated data, with efforts being made to establish guidelines for their most important aspects. Two key developments in this respect are the 2012 annual meeting of the National Council on Measurement in Education (NCME), which included a coordinated session entitled "Design Issues in Equating Studies using Simulation and Resampling", and in 2016 the Instructional Topics in Educational Measurement Series (ITEMS) module published by Feinberg and Rubright. The latter document includes a proposal for calculating the minimum acceptable number of replications depending on the degree of precision required and the variation in the estimated parameter between replications (Feinberg & Rubright, 2016).

With respect to recent trends in DIF detection methods, mention should be made of the following: application of cognitive diagnosis models by means of the DINA model and use of the Wald statistic; recent approaches to the classical delta plot procedure, with adjustments to improve the adequacy of the DIF detection thresholds (Magis & Facon, 2012); application of methods with small samples and purification procedures; the use of hierarchical linear models with latent classes and multilevel structure; and applications of the Rasch model, specifically as regards the Rasch trees method.

As for the latest developments in the field, the main focus of research interest has been centered around the application of methods such as the family of confirmatory factor analytic methods for multilevel and multigroup data structures, logistic regression and the IRT-based methods for data with a multilevel structure, the MIMIC procedure and its extension for multidimensional data, and the Rasch trees method with polytomous items. The advantage of these methods is that the groups do not need to be identified prior to conducting the analysis, thanks to

the combination of observed covariables, and hence these are particularly useful approaches in the applied field. New proposals have also appeared in relation to the two-parameter IRT model, based on stochastic processes of casewise derivatives of the likelihood function, which employ the Lagrange multiplier for categorical, nominal, and ordinal variables. Other proposals include a procedure based on non-linear regression, an extension of logistic regression for detecting bias in the pseudoguessing parameter or the three-parameter IRT model, and application of the Bayesian updating procedure for polytomous items as a way of aggregating information from previous test administrations, including a historical perspective on the item.

The results obtained have enabled us to describe the general field of simulation research on DIF methods and its structure as an area of study. Regarding the latter, we identified 11 topics or lines of research, some of which were focused more on methods of detection, while others were more concerned with assessment-related variables. A distinct topic within the former group is the Mantel-Haenszel method, which is unsurprising given that it has been considered the gold standard approach. However, in more recent studies with simulated data this procedure has been used primarily to compare new proposals, and no new developments of the method itself have been reported. Specifically, the most recent focus of interest in the Mantel-Haenszel method mainly concerns the effects of the ability distribution, thick and thin matching, and its application in the context of international assessment.

Particular mention should be made of the line of research on detection methods that is concerned with the comparison of models, specifically as regards the MIMIC method, the multiple imputation of missing data, and studies with multilevel data, as this line of investigation encompasses many of the topics that underpin the latest developments in the field. Also noteworthy are developments in the use of DIF classification methods, such as Bayesian analysis and application of penalized maximum likelihood estimation, which may be relevant to the field of educational assessment. Among the more classical or less recent lines of research, there are studies based on the SIBTEST methods, those aimed at detecting DIF in the test (DFIT and DTF), and those which more broadly address item and test bias.

In the second group of lines of research, concerning relevant variables for DIF detection, two recent topics are the study of multiple response models and item grouping. Studies on the first topic primarily examine the effect of these response models on measurement invariance, and the most recent concepts include the rating scale model, model-based recursive partitioning, and tree-based methods, which are related to another of the recently developed topics in the area. The second topic of research concerns models and methods of items grouping, simulating testlets, and addressing aspects related to equating.

Generally speaking, methodological research on DIF detection using simulated data has shown particular interest in the size of the groups simulated. Half of the conditions analyzed in this

review used groups of equal size, while among the remainder the most common sample size ratios simulated involved a reference group that was two, three or four times the size of the focal group; the largest sample size ratios simulated were 40, 50, and 100. As regards the size of each group, the most frequent values for both the reference and the focal group were 500, 1000, 250, 100, and 200. The popularity of using sample size ratios lower than 5 may be related to two factors. The first is the idea that different group sizes affects DIF detection results, as shown in studies such as that by Guilera et al. (2013), who found that Type I error increased and the power of the Mantel-Haenszel statistic decreased in line with increasing sample size ratio — although it should be noted that this effect was not very marked in their study. The second factor to consider is that the majority of empirical studies have focused on the gender variable for detecting DIF, and this may have influenced how variables of this type are simulated, insofar as group sizes do not differ widely.

Another relevant variable when simulating data to study DIF is test length. Here, our results suggest a recent trend towards the simulation of ever shorter tests, of between 2 and 15 items. In fact, since 2011 the maximum test length established in the largest proportion of simulated conditions is 15 items, followed by a maximum of 20 and 30 items. Mention should also be made of the growing interest in the application of purification procedures and the use of effect size measures to estimate the magnitude of DIF.

With respect to DIF detection methods, we proposed a categorization that aimed to group the different approaches into broad families characterized by a core procedure. This approach allowed us to include under each category any corrections to a given procedure which have been implemented, as well as its latest developments for addressing different aspects of assessment and tests. It should be noted that our aim was not to develop a new classification of DIF detection methods, but simply to find a way of managing the large amount of information obtained in the review. One of the main findings was that the methods most widely studied using simulated data are the Mantel-Haenszel procedure, logistic regression and all its derivations, the SIBTEST and its derivations, IRT-LR-DIF, and all those methods based on the Rasch model.

Finally, it should be noted that we only included articles that analyzed DIF detection methods using simulated data because our aim was to conduct an evidence-based review of the effectiveness of existing methods and current trends in their usage, and hence we did not consider any theoretical-methodological or empirical-methodological research. Future research could explore the latter avenues of DIF research and compare their evolution with the results obtained in the present study. A further point to bear in mind is that this systematic review shares the limitations of all such studies, since the information has been obtained from specific databases and, therefore, some potentially relevant publications may have been missed.

5. Conclusions

In summary, this paper offers a general overview of methodological research on DIF detection methods using simulated data that may be of value not only to both experienced and new researchers but also to users in the empirical field. The information gathered in relation to the structure of the field has allowed us to organize the main topics of interest in this specific area of study and shows how different concepts have been linked in order to meet the needs of applied psychometrics. The analysis of methods and simulated variables also reveals how simulation research on DIF has been conducted, which variables and simulated conditions have been considered, and how the field has evolved over time. The emphasis on latest developments and trends helps to shed light on the challenges and gaps in knowledge that remain to be addressed, while also indicating those aspects which are currently the subject of researchers' attention. In this respect, one can see how, in recent years, research on DIF detection methods using simulated data has focused more on the analysis of multilevel data, polytomous items, short tests, and small samples. In addition, researchers have incorporated classification methods that allow a more flexible definition of groups or latent classes and which do not, in contrast to earlier more rigid methods, need to be applied prior to the analysis of DIF. Other recent developments include the use of categorical, nominal, and ordinal covariables and the inclusion of multidimensionality as a relevant variable in the detection of DIF. Methods that include these types of variables are likely to attract considerable research interest in coming years. However, classical methods and those which have been most widely studied, such as the Mantel-Haenszel, logistic regression, the Rasch DIF model, SIBTEST, and IRT-LR-DIF, among others, continue to be of value as the accumulated knowledge regarding their effectiveness under various simulated conditions means that they can be used to establish criteria for comparing new developments and proposals. Regarding these classical methods, it will be interesting to see how proposals such as the Bayesian updating procedure for polytomous items evolve and whether, as a result, they can be integrated within the Mantel-Haenszel method, the Rasch trees procedure, ordinal logistic regression, hierarchical ordinal logistic regression, and non-linear logistic regression. Also of interest will be developments in studies on the empirical histogram method, the MIMIC method, the recently proposed delta plot, the cognitive diagnosis model, and methods based on SEM models for multilevel data and categorical variables.

Funding

Juana Gómez-Benito has received research grant from the Agency for the Management of University and Research Grants of the Government of Catalonia [grant 2017SGR1681]

Credit authorship contribution statement

Ángela I. Berrío: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing-Original draft, Visualization, Supervision, Project administration. **Juana Gómez-Benito**: Conceptualization, Writing-review & editing, Supervision, Project administration. **Erika Margarita Arias-Patiño**: Formal analysis, Resources, Writing-Original draft, Writing-review & editing.

Declaration of competing interest

None.

References

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

Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement*, 34, 807-816. https://doi.org/10.1177/001316447403400408.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22(3), 507-526. http://dx.doi.org/10.1037/met0000077.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389-2404. https://doi.org/10.1002/asi.21419.

Camilli, G., & Shepard, L. (1994). *Methods of Identifying Biased Test Item*. Newbury Park, CA: Sage Publications, Inc.

Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different groups on the items in a test*. Princeton, NJ: Educational Testing Service.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, 6, 269-279.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement*, 28, 61-75.

Feinberg, R. A., & Rubright, J. D. (2016), Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, 35(2), 36-49.

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, 5, 43-53.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, 67, 373-393. https://doi.org/10.1177/0013164406294781.

Gómez-Benito, J., Hidalgo, M. D., Guilera, G., & Moreno, M. (2005). A bibliometric study of differential item functioning. *Scientometrics*, 64, 3-6. https://doi.org/10.1007/s11192-005-0234-y.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential Item Functioning: Beyond validity evidence based on internal structure. *Psicothema*, 30(1), 104-109. https://doi.org/10.7334/psicothema2017.183.

Guilera, G., Gómez-Benito, J., & Hidalgo, M. D. (2009). Scientific production on the Mantel-Haenszel procedure as a way of detecting DIF. *Psicothema*, 21(3), 492-498.

Guilera, G., Gómez-Benito, J., Hidalgo, M. D., & Sánchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods*, 18(4), 553-571. https://doi.org/ 10.1037/a0034306.

Harwell, M., Stone, C. A., Hsu, T-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, 20(2), 101-125. https://doi.org/10.1177/014662169602000201.

Hidalgo, M. D., & Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd edition). USA: Elsevier-Science & Technology.

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp 129-145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (eds.) (1993). *Differential Item Functioning*. Hillsdale, NJ, US: Lawrence Erlbaum Associates Inc.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the American Society for Information Science and Technology*, 68(4), 984-998. https://doi.org/10.1002/asi.23734.

Magis, D., & Facon, B. (2012). Angoff's delta method revisited: improving DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, 65, 302-321. https://doi.org/10.1111/j.2044-8317.2011.02025.x.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample sizes, number of biased items and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, 16, 381-388. https://doi.org/10.1177/014662169201600410.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Assessment, 17*(4), 297-334.

Moher D., Liberati A., Tetzlaff J., Altman D., and The PRISMA Group. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*, 8; 336-341. https://doi.org/10.1016/j.ijsu.2010.02.007.

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment*, 18(1), 9-15. http://dx.doi.org/10.1027//1015-5759.18.1.9.

Park, D., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement*, 14, 163-173. https://doi.org/10.1177/014662169001400205.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, 19(1), 23-37. https://doi.org/10.1177/014662169501900104.

Sassetti, S., Marzi, G., Cavaliere, V., & Ciappei, C. (2018). Entrepreneurial cognition and socially situated approach: a systematic and bibliometric analysis. *Scientometrics*, 116, 1675-1718. https://doi.org/10.1007/s11192-018-2809-4.

Stuart, D. (2018). Open bibliometrics and undiscovered public knowledge. *Online Information Review*, 42(3), 412-418. https://doi.org/10.1108/OIR-07-2017-0209.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3.

Van Eck, N. J., & Waltman, L. (2018). *VOSviewer software version 1.6.9*. Universiteit Leiden & CWTS: Meaningful Metrics

Wang, W. C., & Su, Y. H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17*, 113-144.

3.2. Estudio 2: Análisis con datos simulados

3.2.1. Datos identificativos

THE JOURNAL OF EXPERIMENTAL EDUCATION
2019, VOL. 87, NO. 3, 367–383
https://doi.org/10.1080/00220973.2018.1435502

Routledge
Taylor & Francis Group

Check for updates

Effect of Sample Size Ratio and Model Misfit When Using the Difficulty Parameter Differences Procedure to Detect DIF

Ángela I. Berrío [a,b], Aura N. Herrera [a], and Juana Gómez-Benito [b]

[a]Universidad Nacional de Colombia-Sede Bogotá, Bogotá, Colombia; [b]University of Barcelona, Barcelona Spain

Título: Effect of sample size ratio and model misfit when using the difficulty parameter differences procedure to detect DIF.

Autoras: Ángela I. Berrío, Aura N. Herrera & Juana Gómez-Benito

Año: 2019

Revista: Journal of Experimental Education

Métrica según Journal Citation Reports para el año 2019:

- Factor de Impacto: 2,107
- Rango por factor de impacto: 21 de 60, Cuartil: Q2, Categoría: Psychology, Educational (SSCI).
- Rango por Journal Citation Indicator (JCI): 13 de 59, Cuartil: Q1, Categoría: Psychology, Educational (SSCI).

Material Suplementario: se puede consultar en el anexo 2.

DOI: https://doi.org/10.1080/00220973.2018.1435502

3.2.2. Resumen

Teniendo en cuenta que la categoría de métodos de detección de DIF basados en el modelo Rasch fue una de las más estudiadas en el primer estudio, que uno de los procedimientos que hicieron parte de esta categoría fue la diferencia del parámetro de dificultad, y que este procedimiento es el que se ha empleado en Colombia para analizar DIF en las pruebas nacionales que evalúan la calidad de la educación y dan acceso a la universidad (a través del software Winsteps); este estudio se diseñó para analizar el funcionamiento de dicho procedimiento en condiciones experimentales que simularon las pruebas nacionales en Colombia. Las preguntas que orientaron este estudio fueron: ¿Qué efectos tendrán variables como razón de tamaños y ajuste/desajuste del modelo de simulación de datos en el funcionamiento del procedimiento diferencia del parámetro de dificultad? ¿Existirán diferencias en el funcionamiento de dicho procedimiento cuando se emplea la detección de DIF tomando como criterio la prueba de significación versus cuando se aplica un criterio de magnitud de DIF? ¿Qué recomendaciones respecto del uso de este procedimiento se pueden hacer para su aplicación en evaluaciones nacionales de carácter masivo?

El aporte original de este estudio consistió en el análisis del efecto de variables como razón de tamaños de muestra con valores altos (comenzando en 20 y terminando en 1,000), gran tamaño de muestra (130,000 examinados) y análisis del efecto del ajuste/desajuste del modelo de simulación de los datos en el funcionamiento del procedimiento diferencia del parámetro de dificultad.

Para analizar el funcionamiento del procedimiento bajo condiciones similares a las de la evaluación masiva cuando se emplea el software Winsteps, se simuló la matriz de respuesta de 130,000 examinados a una prueba unidimensional de 30 ítems dicotómicos. En total fueron 120 condiciones experimentales, en cada una de ellas se cruzaron los valores de cuatro variables independientes: razón de tamaños de muestra (20, 100, 250, 500 o 1,000), modelo de simulación de los datos (1PL o 3PL c = 0.15), impacto (grupos iguales, diferencias en la media, diferencias en la

varianza o diferencias en la media y varianza) y porcentaje de ítems con DIF (0%, 10% o 20%). Cada condición experimental se replicó 500 veces y en total se obtuvieron 60,000 matrices de datos. Para valorar el adecuado funcionamiento del procedimiento se definieron como variables dependientes: el error tipo I, analizado a través de las tasas de falsos positivos, y la potencia para detectar DIF, medida a través de las tasas de detección correcta. Además, se simularon ítems con DIF uniforme y no uniforme y se emplearon dos criterios para la detección de DIF: la detección de DIF sólo por la prueba de significación o la combinación de la prueba de significación y una medida de magnitud de DIF de 0.5 logits siguiendo las propuestas de Lai et al. (2005) y Linacre (2012).

Los resultados indicaron que el procedimiento presentó buen control del error tipo I cuando se empleó como criterio de detección el test de significación sólo o en combinación con la medida de magnitud de DIF y el modelo de simulación de datos fue 1PL, en todas las condiciones de razón de tamaños de muestra, impacto y porcentaje de ítems con DIF, excepto cuando se empleó sólo el test de significación, la razón de tamaños fue 20 y además: (a) 10% de los ítems tenían DIF y los grupos presentaban igual distribución del nivel de atributo, o (b) 20% de los ítems tenían DIF y los grupos presentaban igual distribución del nivel de atributo o habían diferencias en la media o en la varianza de la distribución de atributo.

Respecto a la potencia del procedimiento para detectar DIF, los resultados indicaron que cuando 10% de los ítems tenían DIF y los datos se habían simulado con el modelo 1PL, empleando como criterio de detección el test de significación sólo o en combinación con la medida de magnitud de DIF, las tasas de detección correcta fueron moderadas (entre 0.62 y 0.76) con razón de tamaños 20, 100 o 250 para todas las condiciones de impacto simuladas. Por su parte, cuando 20% de los ítems tenían DIF y el modelo de simulación fue 1PL, empleando como criterio de detección el test de significación sólo o en combinación con la medida de magnitud, las tasas de detección correcta fueron moderadas (0.62 a 0.78) cuando la razón de tamaños fue 20 para todas las condiciones de impacto y cuando la razón

de tamaños fue 100 para grupos con igual distribución del nivel de atributo o grupos con diferentes varianzas en la distribución del nivel de atributo. Como era de esperar la detección correcta de DIF uniforme bajo la condición de modelo ajustado (1PL) con 10% de ítems con DIF mostró tasas de detección correcta entre 0.66 y 1.00 en todas las condiciones de impacto y razón de tamaños simuladas, excepto para la razón de tamaños 1,000. Con 20% de ítems con DIF las tasas variaron entre 0.50 y 1.00 para: (a) razones de tamaño 20 o 100 en todas las condiciones de impacto y (b) razón de tamaño 250 en todas las condiciones de impacto excepto cuando había diferencias en la media de la distribución del nivel de atributo.

3.2.3. Versión postprint

Effect of Sample Size Ratio and Model Misfit when Using the Difficulty Parameter Differences Procedure to Detect DIF

Ángela I. Berrío[1,2], Aura N. Herrera[1], and Juana Gómez-Benito[2,3]

[1] Universidad Nacional de Colombia-Sede Bogotá-, Facultad de Ciencias Humanas, Departamento de Psicología.
[2] Quantitative Psychology Unit, Department of Social Psychology and Quantitative Psychology, Faculty of Psychology, University of Barcelona, Spain
[3] Group on Measurement Invariance and Analysis of Change (GEIMAC), Institute of Neurosciences, University of Barcelona, Barcelona, Spain
Corresponding author. Aura N. Herrera, Universidad Nacional de Colombia, Sede Bogotá, Facultad de Ciencias Humanas, Departamento de Psicología, Laboratorio de Psicometría, Ciudad Universitaria, Edificio 212, oficina 229, Bogotá, 111321, Colombia. E-mail: **anherrerar@unal.edu.co**

Abstract

This study examined the effect of sample size ratio and model misfit on the Type I error rates and power of the Difficulty Parameter Differences procedure using Winsteps. A unidimensional 30-item test with responses from 130,000 examinees was simulated and four independent variables were manipulated: sample size ratio (20/100/250/500/1000); model fit/misfit (1PL and 3PL c = .15 models); impact (no difference/mean differences/variance differences/mean and variance differences); and percentage of items with uniform and nonuniform DIF (0%/10%/20%). In general, the results indicate the importance of ensuring model fit to achieve greater control of Type I error and adequate statistical power. The manipulated variables produced inflated Type I error rates, which were well controlled when a measure of DIF magnitude was applied. Sample size ratio also had an effect on the power of the procedure. The paper discusses the practical implications of these results.

Keywords: DIF, Difficulty Parameter Differences procedure, model misfit, sample size ratio, Winsteps.

Introduction

In the context of research that seeks to detect bias (Hambleton, Swaminathan, & Rogers, 1991), the concept of differential item functioning (DIF), as proposed by Holland and Thayer (1988), refers to the empirical evidence that a test item functions differently in two groups that are matched on ability level but that differ in some other characteristics, such as gender, ethnicity, or cultural background. The presence of DIF is a necessary but not sufficient condition for stating that an item is biased, since DIF is a metric characteristic of the item. Whereas bias refers to the theoretical explanation for its presence, DIF may be due to an aspect that is of interest or relevant to the measured variable, in which case the item would not be biased, even though it functions differently in two groups (Clauser & Mazor, 1998).

In general, DIF involves an interaction between the group and the item (Linacre & Wright, 1987), and one way of detecting DIF is by examining the difference in item difficulty, specifically by applying the Difficulty Parameter Differences procedure to compare the item parameters for the groups of interest (Angoff, 1993). The Difficulty Parameter Differences procedure using Winsteps is based on the premise that the probability of a correct response to an item depends on the difference between the person's ability level and the item difficulty, it being assumed that there are no differences in the discrimination and guessing parameters (Rasch, 1960). The procedure for detecting DIF begins by estimating ability (*B*) for each examinee in each group on a common interval scale, and then estimating the difficulty parameter (*D*). Once the parameters have been estimated, the probability that a person *n* will respond correctly to item *i* (P_{ni}) can be expressed as described by Bond and Fox (2007),

$$f(\mathrm{B} - \mathrm{D}) = \mathrm{P_{ni}}(\mathrm{x} = 1) \quad (1)$$

since the estimation of the difficulty parameter is the logistic transformation of the probability of correct responses (p_r and p_f for the reference (*r*) and focal (*f*) groups, respectively) divided by the probability of incorrect responses (q_r and q_f for the *r* and *f* groups). Given that D_r and D_f are the difficulty estimates for the reference and focal groups, then for any pair of examinees with the same ability level, the difference in item difficulty can be expressed as:

$$\mathrm{D_f} - \mathrm{D_r} = \ln\frac{p_\mathrm{r}}{q_\mathrm{r}} - \ln\frac{p_\mathrm{f}}{q_\mathrm{f}} \quad (2)$$

Each difficulty estimate has an associated standard error, and thus the *t* statistic can be calculated to test the hypothesis of equal difficulty across two groups.

$$t_\mathrm{i} = \frac{\mathrm{D_{ir}} - \mathrm{D_{if}}}{\sqrt{\mathrm{S_{ir}^2} + \mathrm{S_{if}^2}}} \quad (3)$$

where D_{ir} and S_{ir} are, respectively, the difficulty estimate and standard error for item *i* in the reference group, while D_{if} and S_{if} are, respectively, the difficulty estimate and standard error for item *i* in the focal group. A more detailed description of the procedure can be found in Draba (1977), Hauser and Kingsbury (2004), Linacre and Wright (1987), Schulz, Perlman, Rise, and Wright (1992), and Wright, Mead, and Draba (1976). One of the advantages of the Difficulty Parameter Differences procedure using Winsteps is that it can be applied to relatively small samples and it has therefore become popular in the empirical context. However, although it is capable of analyzing DIF in the difficulty parameter (uniform DIF), the procedure does not detect differences in the discrimination parameter or simultaneous differences in both the difficulty and discrimination parameters, which from the point of view of item response theory (IRT) would be considered nonuniform DIF (Finch & French, 2014). Furthermore, with large samples it may achieve statistical significance without the differences being of practical significance. Consequently, it needs to be accompanied by a measure of DIF magnitude (Lai, Teresi, & Gershon,

2005). The most widely used criterion in this respect is that a difference in calibrations greater than .5 indicates that the item shows DIF and should be analyzed further and most likely eliminated. Lai et al. (2005) also proposed an intermediate category of DIF magnitude, corresponding to a difference in calibrations of between .35 and .5, this being adapted from the categories used for interpreting the Mantel Haenszel Delta DIF statistic.

Numerous studies (Cohen, Kim, & Subkoviak, 1991; DeMars, 2010; Finch & French, 2008; López, Stark, & Chernyshenko, 2009; Navas & Gómez-Benito, 2002; Pei & Li, 2010; Wang, 2004; Wang & Yeh, 2003, among many others) have examined the effect of various factors on the functioning of DIF detection techniques based on IRT. The most widely studied factors have been sample size, the magnitude and type of DIF, and anchoring procedures. However, much less attention has been paid to sample size ratio (i.e., the number of examinees in the reference group for each examinee in the focal group), a variable that could have differential effects on the process of parameter estimation, especially as regards the ability parameter, thus affecting the control of Type I error. Those studies that have examined this aspect indicate that sample size ratio may affect the Type I error rate; although the reported effect is not especially marked, it appears that the Type I error rate increases as the group sizes become more unbalanced. This variable has been studied for procedures such as the Mantel Haenszel statistic (Goodman, Willse, Allen, & Hlaric, 2011; Herrera & Gómez-Benito, 2008; Magis, Tuerlinckx, & De Boeck, 2015; Monahan & Ankenmann, 2010; Paek & Guo, 2011), logistic regression (Gómez-Benito, Hidalgo, & Padilla, 2009; Herrera & Gómez-Benito, 2008; Magis et al., 2015), and Lord's χ^2 coupled with the likelihood ratio (IRT-LR) (Lei, Chen, & Yu, 2006), for which the sample size ratio does not exceed 40 and commonly takes a value between 1 and 20. Given that methods based on IRT require the fulfilment of certain assumptions and model fit to the data, it is also important to examine DIF detection when the data are simulated using a three-parameter model—that is, in the case of model misfit.

As regards factors that affect power and Type I error rate, it is generally agreed that power is influenced by sample size, the percentage of items with DIF, and moderate item parameter values, whereas the Type I error rate is affected by variables such as impact, the percentage of items with DIF, and moderate item parameter values (especially in the difficulty parameter).

In light of the above, the main goal of the present study was to examine experimentally the effect of sample size ratio and model misfit on the power and Type I error rate of the Difficulty Parameter Differences procedure when implemented in Winsteps (Linacre, 2006). Based on the results we aim to establish guidelines regarding the functioning of this procedure, especially in those cases in which two groups differ substantially in size (sample size ratio ≥ 20). In this respect, we believe that the study is one of the first to examine experimentally variables such as sample size ratio, model misfit, and differences in the mean and variance of the ability distribution when

applying the Difficulty Parameter Differences procedure in Winsteps. A further, related objective is to provide new evidence about the robustness of the procedure under extreme conditions of model misfit and sample size ratio. The findings in this regard would be of considerable interest in relation to large scale applications of high-stake tests in educational contexts characterized by the presence of ethnic minorities. Indeed, the sample sizes and ratios that are simulated in this study were selected precisely to represent test situations in which certain ethnic groups are unequally represented in comparison with the general population. The three-parameter model (3PL c = .15) was chosen to represent model misfit as one would expect to encounter a certain degree of chance scoring in the context of educational testing.

Method

A Monte Carlo study was conducted in order to simulate the conditions of high-stake tests, specifically as regards test length, the value of item parameters and sample size. Thus, we considered a unidimensional test comprising 30 dichotomous items with responses from 130,000 examinees. The response matrices were generated using item parameter estimates obtained from databases of students' answers to a nationwide university entrance exam, the values of which are shown in Table 1. For each item with nonuniform DIF we simulated two with uniform DIF. The area between the item characteristic curves (ICCs) (Raju, 1988) was set at around .40, representing a moderate magnitude that can be detected by most procedures (Shepard, Camilli, & Williams, 1985).

Table 1. Item parameter estimates used to generate the response matrices.

Item no.	a	b	Item no.	a	b
1	2.20559	2.07487	16	0.75536	-1.0611
2	0.46125	1.46794	17	1.16234	2.08233
3	0.39115	0.5167	18	1.42727	2.56616
4	0.76307	2.40671	19	0.80862	-0.20613
5	0.43264	-0.9214	20	0.50056	1.76429
6	0.33913	2.92167	21	1.22742	2.08076
7	1.02945	-0.38977	22	0.79738	0.03815
8	0.51012	1.60746	23	0.48182	2.08688
9	1.29943	-0.95524	24	0.63603	0.76112
10	1.40483	-0.99862	25	0.36933	2.30862
11	1.08021	0.71909	26	0.5267	0.44556
12	0.75375	0.26373	27	0.9977	0.32505
13	0.6093	1.98956	28	0.58404	0.42478
14	1.55649	3.08462	29	1.60704	-1.55029
15	3.01816	1.85731	30	0.36103	1.4191

Note. The "a" and "b" columns correspond to the discrimination and difficulty parameters, respectively.

Variables

A total of five variables were manipulated in the simulation study: sample size ratio (r), model fit (m), impact (i), percentage of items with DIF (%DIF), and type of DIF (DIF). Table 2 shows the variables considered and the conditions manipulated.

The sample size ratio (r)—that is, the number of examinees in the reference group for each examinee in the focal group, was calculated as follows: $r = \frac{n_r}{n_f}$, where n_r is the number of examinees in the reference group and n_f is the number of examinees in the focal group.

Fit/misfit was defined on the basis of the model used to calculate the matrices for the probability of a correct item response. In the misfit condition these matrices were calculated using a three-parameter model (3PL-c) with a fixed guessing parameter of c = .15.

Impact (i) was defined as the difference between the groups (*r* and *f*) in the ability distribution parameters (μ and σ^2). In all the conditions we used a normal distribution with $\mu = 0$ and $\sigma^2 = 1$, *N(0,1)* for the reference group and manipulated the values of μ and σ^2 for the focal group.

Finally, we manipulated the percentage of items with DIF (%DIF) and the type of DIF (uniform or nonuniform).

Table 2. Description of the independent variables.

Variables and levels	Description of the levels
Sample size ratio (*r*), 5 levels	
r=20	n_f = 6190 ; n_r = 123,810
r=100	n_f = 1287; n_r n_r= 128,713
r=250	n_f = 518 ; n_r = 129,482
r=500	n_f = 259 ; n_r = 129,741
r=1000	n_f = 130 ; n_r = 129,870
Model fit (*m*), 2 levels	Model used to simulate data
m=Model fit	following a 1PL.
m=Model misfit	following a 3-parameter model with c = .15.
Impact (*i*), 4 levels	Differences in the ability distribution between groups
i=equal groups	Ability distribution of N(0,1) for both groups.
i=different means, (μ different)	Ability distribution of N(-1,1) for the focal group
i=different variances, (σ^2 different)	Ability distribution of N(0,1.5) for the focal group
i=different means and variances (μ and σ^2 different)	Ability distribution of N(-1,1.5) for the focal group
Percentage of items with DIF (*%DIF*), 3 levels	
0%	Equal parameter values across both groups for all items
10% items with DIF	3 items (3, 9 and 10) with different parameter values in the focal group
20% items with DIF	6 items (3, 9, 10, 21, 24 and 25) with different parameter values in the focal group
Type of DIF, 2 levels	
Uniform	Different difficulty for the focal group on items 3, 10, 21 and 25.
Non-uniform	Different discrimination for the focal group on items 9 and 24.

The dependent variables were the Type I error and power of the procedure for detecting DIF. The former was calculated according to the false positive (FP) rate in the condition of 0% DIF and the FP rate for non-DIF items in the 10% and 20% DIF conditions. The power of the procedure was estimated as the correct detection (CD) rate in the 10% and 20% DIF conditions. The design was fully crossed with 120 experimental conditions and 500 replications for each

condition. The number of replications was set following consideration of two sources of information. First, we followed the recommendation of Feinberg and Rubright (2016) for determining the minimum acceptable number of replications. Given that our parameter of interest is difficulty, since DIF will be well estimated provided that this parameter is well estimated, we calculated the standard deviation of these estimates for 50 replications (σ = .10) and established a 68% level of confidence in any parameter estimate being within ± two decimal points of the value that would be obtained if the entire population of replications were used (σ_M = .005). Thus, application of the formula proposed by Feinberg and Rubright (2016) gives the following:

$$n = \left(\frac{\hat{\sigma}}{\sigma_M}\right)^2 + 1 = \left(\frac{0.10}{0.005}\right)^2 + 1 = 401 \text{ replications}$$

The second source of information considered was the number of replications used in DIF studies published over the last five years. A review of the literature for DIF studies using simulated data identified a total of 72 studies in which the number of replications most commonly employed was 100 (33.3% of the studies), followed by <50 (20.8%), and then 500 and 1000 (15.3% of studies in each case). The remaining studies used 10000 replications (5.6% of studies), 200 (2.8%), and 300, 400, 1500, 2000, and 5000 replications (1.4% in each case). Based on this information we considered that 500 replications would enable us to obtain stable estimates.

Procedure

Data generation. We simulated a unidimensional test with 30 dichotomous items. In order to introduce DIF we modified the difficulty and discrimination parameters for the focal group, according to the type of DIF. The parameter values for DIF items are shown in Table 3.

Table 3. Parameters for the items for which DIF was simulated in the focal group.

Item	a	b	c	Area between the ICCs	Type of DIF
3	0.39115	0.999	0.15	0.410	Uniform
9	0.74231	-0.95524	0.15	0.400	Nonunif.
10	1.40483	-0.52	0.15	0.407	Uniform
21	1.22742	2.56054	0.15	0.408	Uniform
24	1.00591	0.76112	0.15	0.401	Nonunif.
25	0.36933	2.79	0.15	0.409	Uniform

Note. The "a", "b" and "c" columns correspond to the discrimination, difficulty and pseudo-guessing parameters, respectively.

The ability parameter was simulated following a normal distribution with different values of μ and σ^2 depending on the experimental condition of impact. Based on the vectors of the item and ability parameters we then used the 1PL or 3PL-c model to calculate the probability of an individual responding correctly to each test item.

Once the probability matrices had been obtained for all the experimental conditions, we generated the response matrices with values of *1* (correct) and *0* (incorrect), following a Bernoulli distribution with parameter $P = P(U_i|\theta)$. Each of the experimental conditions was replicated 500

times, generating 60,000 response matrices that were obtained using the R 2.6 software package (R Development Core Team, 2007).

Detection of DIF. The Winsteps 3.63 software package (Linacre, 2006) was used to detect DIF by means of the Difficulty Parameter Differences procedure, applying two criteria for the detection of DIF: the first was the *t* test of significance, while the second was the combination of the *t* test of significance and a measure of DIF magnitude. The latter was defined as a difference of .5 logits between the difficulty parameters of the reference and focal groups; this is in line with the proposals of Lai et al. (2005) and Linacre (2012), the latter stating that the size of DIF should be at least .5 logits for the DIF to be notable.

Data analysis. To assess the quality of the estimates we performed item difficulty recovery analyses, calculating Pearson correlation coefficients between the values of the simulated parameter and the estimates obtained through Winsteps.

The FP rate was calculated as the number of times DIF was detected in a non-DIF item across the 500 replications in each of the three conditions (0%, 10%, and 20% of items with DIF). The CD rate was calculated as the proportion of correctly detected items across the 500 replications of each experimental condition. In both cases we defined a significance level of .05, corrected via a Bonferroni adjustment due to the multiple comparisons conducted in our data analysis. Thus, for analysis of both the FP rate and the CD rate we applied the following criteria: a significance level of .00125 and a combination of a significance level of .00125 with the measure of DIF magnitude (a difference greater than or equal to .5 logits). Note that in applying the Bonferroni adjustment we took multiple comparisons into account, due to the interaction of the manipulated variables in the DIF test: sample size ratio, impact and model fit/misfit (5 x 4 x 2 = 40).

The effect of the manipulated variables was estimated by means of repeated measures analyses of variance for each of the conditions of DIF percentage, defining Type I error and power as dependent variables and sample size ratio, impact, and fit/misfit as independent variables. These analyses were complemented with descriptive analyses of the detection rates. All these analyses were performed using SPSS, v.18.

Results

The estimates obtained through Winsteps were largely accurate, since the mean correlations for the samples were above .90. In the model fit condition the correlations were all .99 regardless of the group size and impact that were simulated. In the model misfit condition the correlation for the reference group remained constant (r = .95), while that for the focal group ranged between .92 and .98.

Type I error

Significance test. In the 0% DIF condition the analyses of variance showed significant effects on the FP rate for the variables sample size ratio ($F = 35.83$, $p < .001$, $\eta^2 = .85$), impact ($F = 37.00$, $p < .001$, $\eta^2 = .80$), and model fit/misfit ($F = 65.17$, $p < .001$, $\eta^2 = .69$), as well as for their interactions. Applying Bradley's (1978) liberal criterion range of .025 to .075, mean FP rates were below .075 when the sample size ratio was r = 1000 or r = 500 ($\bar{x} = .032$; $\bar{x} = .066$; respectively), when the impact condition was i = equal groups ($\bar{x} = .051$) or i = different variances ($\bar{x} = .051$), and when the model fit the data ($\bar{x} = .002$). For the other variables the mean FP rate ranged between .11 and .25. As regards the interaction between variables, mean FP rates were below .075 for all conditions of sample size ratio and impact when the model fit. With model misfit the FP rate remained within an acceptable range (below .075) only when the groups were matched for ability distribution—for all conditions of sample size ratio—and when the groups differed in variance and r was equal to 250, 500, or 1000. For the other conditions the FP rate ranged between .11 and .87. Table 4 shows the mean FP rates under the conditions of 0%, 10%, and 20% of items with DIF.

Table 4. False positive rates associated with the different independent variables when applying the significance test as the criterion.

Impact	Sample size ratio				
	20	100	250	500	1000
			Model fit		
Equal groups	**.001**/.125/.236	**.001/.011/.021**	**.001/.005/.007**	**.001/.003/.004**	**.001/.003/.002**
μ different	**.001/.072**/.110	**.002/.009/.011**	**.002/.004/.006**	**.003/.004/.005**	**.002/.003/.004**
σ^2 different	**.002/.061**/.183	**.001/.006/.016**	**.001/.003/.006**	**.001/.002/.003**	**.001/.002/.003**
μ and σ^2 both different	**.014/.034/.073**	**.002/.004/.009**	**.001/.003/.004**	**.001/.003/.004**	**.001/.002/.002**
			Model misfit		
Equal groups	**.002/.029**/.083	**.002/.006/.014**	**.003/.004/.005**	**.003/.003/.005**	**.001/.002/.002**
μ different	.866/.847/.867	.683/.677/.685	.454/.457/.465	.276/.249/.266	.112/.092/.106
σ^2 different	.295/.313/.373	.112/.103/.130	**.044/.041/.050**	**.026/.024/.028**	**.025/.029/.029**
μ and σ^2 both different	.799/.770/.765	.578/.597/.611	.391/.387/.397	.218/.194/.206	.112/.082/.090

Note. The values in each cell express the false positive (FP) rates under the conditions of 0% / 10% / 20% of items with DIF, respectively. Values in boldface correspond to mean FP rates ≤ .075.

In the condition of 10% DIF, the analyses of variance showed that all the variables had a significant effect on the Type I error rate: sample size ratio ($F = 70.83$, $p < .001$, $\eta^2 = .93$), impact ($F = 30.51$, $p < .001$, $\eta^2 = .79$) and model fit/misfit ($F = 51.70$, $p < .001$, $\eta^2 = .67$). As regards the interactions between the variables, the FP rates behaved similarly to what was observed for the condition of 0% DIF: mean FP rates were below .075 in all the conditions of impact and sample size ratio when the model fit, except when i = equal groups and r = 20. In the case of model misfit, FP rates were only within the nominal range when i = equal groups, this for all conditions of sample size ratio, and when i = different variances for r = 250, r = 500, or r = 1,000.

In the condition of 20% of items with DIF, mean FP rates were similar to those obtained with 10% of items showing DIF, except for the condition of r = 20 and i = equal groups, i = different means, or i = different variances, in which cases the Type I error was above the nominal level for both model fit and misfit (the mean FP rates were between .083 and .867).

The combination of the significance test with the measure of DIF magnitude. For all three DIF conditions (0%, 10%, and 20% of items) and model fit, FP rates remained well below .075 for all conditions of sample size ratio and impact. With model misfit, FP rates were below .075 when the impact condition was equal groups or different values of σ^2, for all conditions of sample size ratio (see Table 5).

Table 5. False positive rates associated with the different independent variables when applying the combination of the significance test with the measure of DIF magnitude as the criterion.

Impact	Sample size ratio				
	20	100	250	500	1000
			Model fit		
Equal groups	**<.001/<.001/<.001**	**<.001/.001/.001**	**.001/.002/.003**	**.001/.003/.004**	**.001/.003/.002**
μ different	**<.001/.001/.001**	**.001/.003/.003**	**.001/.002/.004**	**.003/.004/.005**	**.002/.003/.004**
σ^2 different	**<.001/<.001/<.001**	**<.001/<.001/.001**	**.001/.002/.003**	**.001/.002/.003**	**.001/.002/.003**
μ and σ^2 both different	**<.001/<.001/<.001**	**<.001/.001/.003**	**.001/.002/.003**	**.001/.003/.004**	**.001/.002/.002**
			Model misfit		
Equal groups	**<.001/<.001/<.001**	**<.001/<.001/<.001**	**<.001/<.001/<.001**	**.002/.002/.003**	**.001/.002/.002**
μ different	.211/.193/.227	.217/.207/.230	.267/.253/.270	.253/.228/.247	.112/.092/.106
σ^2 different	**.033/.037/.040**	**.030/.029/.030**	**.026/.024/.024**	**.026/.023/.026**	**.025/.029/.029**
μ and σ^2 both different	.165/.156/.192	.193/.167/.194	.236/.214/.234	.207/.183/.196	.112/.082/.090

Note. The values in each cell express the false positive (FP) rates under the conditions of 0% / 10% / 20% of items with DIF, respectively. Values in boldface correspond to mean FP rates ≤ .075.

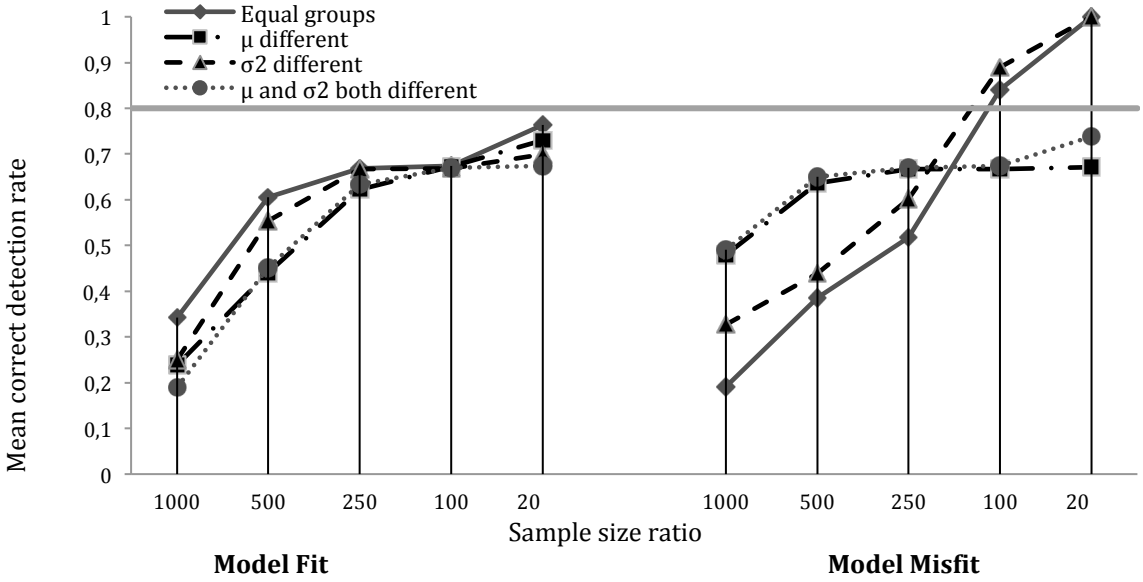
Power

Significance test. The analyses of variance for the 10% and 20% DIF conditions showed that significant effects were observed for sample size ratio ($F = 13.82$, $p = .050$, $\eta^2 = .87$; $F = 17.69$, $p = .003$, $\eta^2 = .78$, respectively). The power results are summarized graphically in Figure 1.

In the 10% of items with DIF condition and model fit, the CD rates were: (a) moderate (.62 to .76) for a sample size ratio of 20, 100, or 250, under all conditions of impact; (b) acceptable (.55 to .61) for a sample size ratio of 500 when the groups had either the same ability distribution or different σ^2 values. With model misfit, CD rates were: (a) desirable (above .80) for sample size ratios of 20 or 100 when the groups had either the same ability distribution or different σ^2 values; (b) moderate (between .60 and .79) when r = 500 and the impact conditions involved different μ or different values of both μ and σ^2 and when r = 250 for impact conditions involving a difference in the ability distribution. When r = 100 or r = 20, CD rates between .60 and 1 were obtained for all impact conditions.

In the 20% DIF condition and with model fit, CD rates were between .64 and .78 with sample size ratios of 100 when the groups had either the same ability distribution or different σ^2 and when r = 20 for all conditions of impact. With model misfit, CD rates were between .60 and 1 with sample size ratios of 100 when i = μ different and r = 20, regardless of the impact condition.

Condition: 10% of items with DIF



Condition: 20% of items with DIF

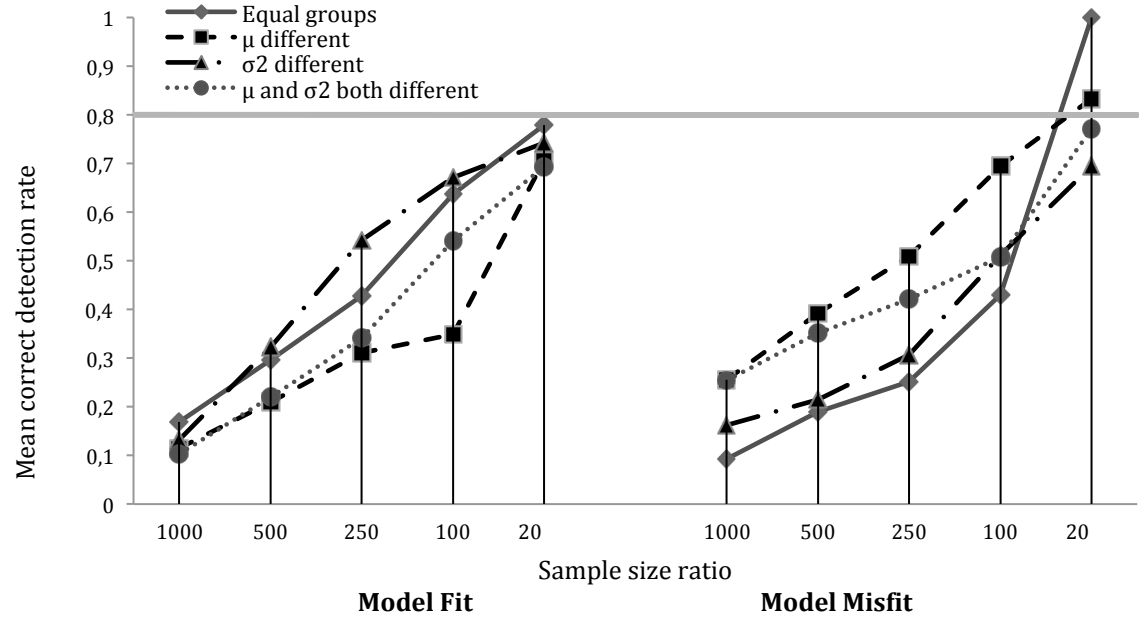


Figure 1. Mean correct detection (CD) rates with 10% and 20% of test items showing DIF and according to model fit, impact, and sample size ratio when the criterion for detecting DIF was the significance test. Note. The horizontal line indicates the desirable correct detection rate.

The combination of the significance test with the measure of DIF magnitude. In the condition of 10% of items with DIF, overall CD rates ranged between .19 and .67 for both model fit and model misfit. In the case of model fit, CD rates were: (a) acceptable with r = 500 and i = equal groups (.61) or i = σ^2 different (.55) and (b) moderate (.62 to .67) with sample size ratios of 250, 100, or 20 for all conditions of impact. With model misfit, CD rates were moderate (.61 to .67) when r = 500, 250, 100, or 20 and when i = μ different or i = μ and σ^2both different. Figure 2 shows the CD rates obtained when the criterion for detecting DIF was the combination of the significance test and the measure of DIF magnitude.

In the 20% DIF condition, CD rates were between .09 and .67. With model fit, rates were moderate (.62 to .67) with r = 20 for all impact conditions, and with r = 100 when the groups had either the same ability distribution or different σ^2. With model misfit, CD rates fell below acceptable levels (.09 to .40).

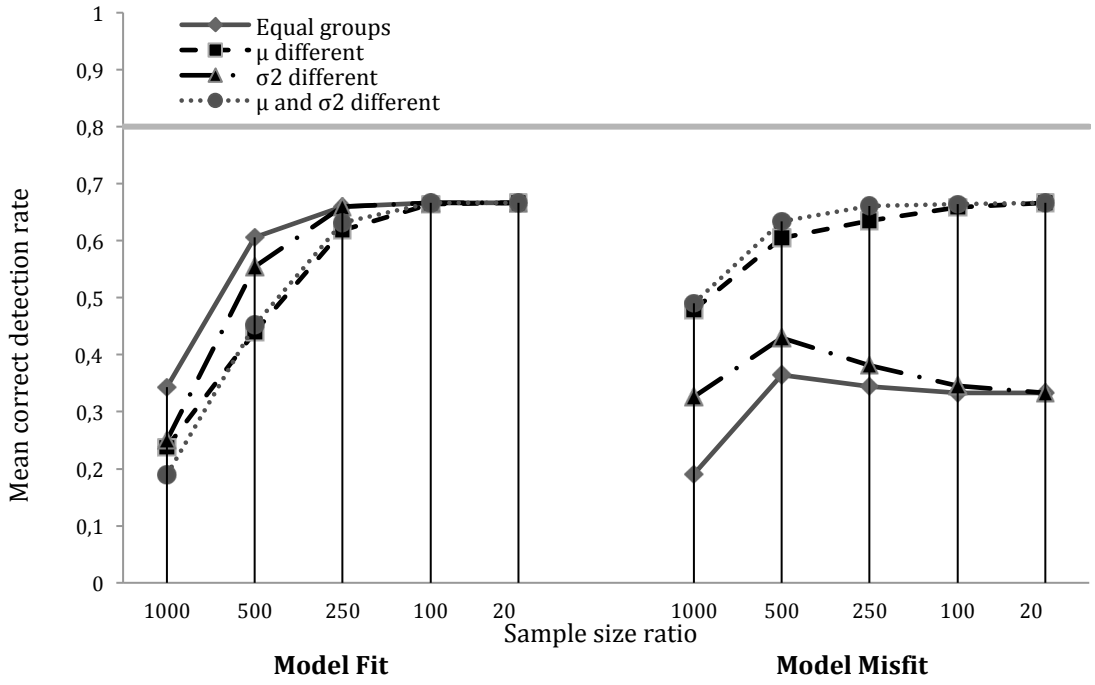
We also analyzed CD rates for uniform DIF and nonuniform DIF. As expected, detection of uniform DIF was adequate for the 10% DIF condition and model fit (.66 to 1) for all conditions of impact and sample size ratios, with the exception of r = 1000, while in the case of model misfit, rates were barely acceptable (around .50) for all the aforementioned conditions. For the condition of 20% DIF and model fit, rates ranged between .50 and 1 for all conditions of impact and sample size ratios of 100 and 20. Adequate rates (.51 to .81) were also obtained for a sample size ratio of 250 and all conditions of impact, except when the ability distribution showed differences in the mean.

With respect to the detection of nonuniform DIF, the procedure achieved CD rates above .80 under the condition of 10% DIF and model misfit and when the ability distribution showed differences in the mean or in both the mean and variance for all sample size ratios except r = 1000. Under the condition of 20% DIF and model misfit, CD rates fell to between .40 and .50 when the ability distribution showed differences in the mean or in both the mean and variance for all sample size ratios except r = 1000. Finally, for the conditions of 10% and 20% DIF and model fit, we considered that CD rates were not interpretable as they were below .004.

Discussion and Conclusions

In the context of large-scale applications of high-stake tests, studies of DIF and bias can help to ensure that the tests in question yield valid and equitable measures across population groups. This requires the availability of robust DIF detection procedures and knowledge of the conditions under which they function adequately. The general aim of this study was to contribute to our understanding of the conditions under which the Difficulty Parameter Differences procedure can be suitably applied, specifically by examining the effect of sample size ratio and model misfit on the power and Type I error rate of the procedure.

Condition: 10% of items with DIF



Condition: 20% of items with DIF

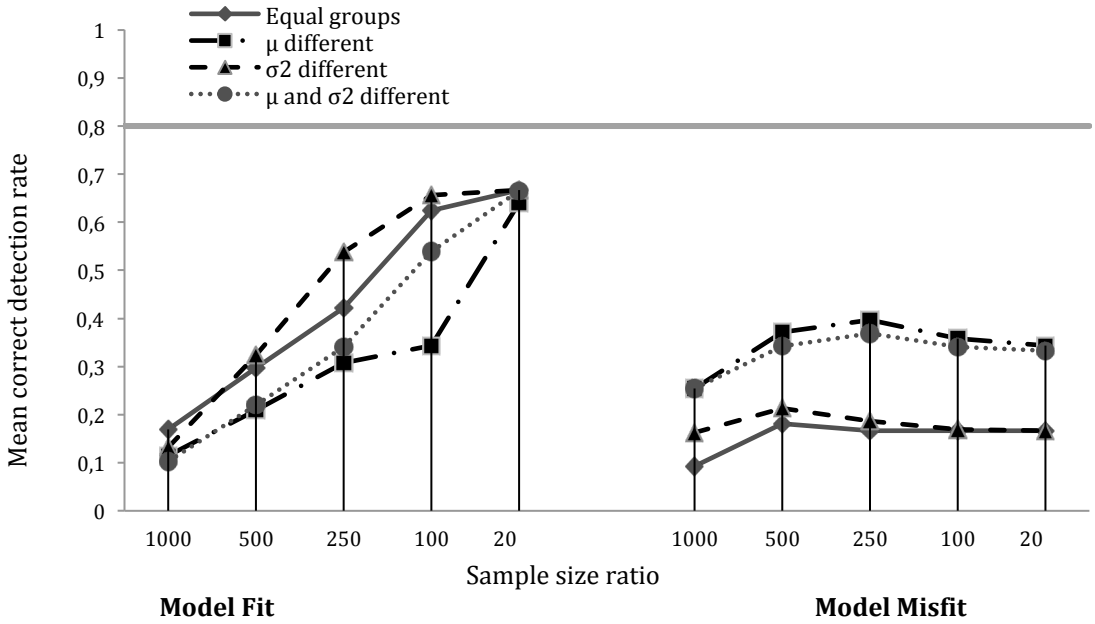


Figure 2. Mean correct detection (CD) rates with 10% and 20% of test items showing DIF and according to model fit, impact, and sample size ratio, when the criterion for detecting DIF was the combination of the significance test with the measure of DIF magnitude. Note. The horizontal line indicates the desirable correct detection rate.

One aspect that may produce negative effects in DIF studies is the precision of parameter estimates. Our results regarding the recovery and precision of item difficulty estimates were relatively adequate. This outcome is not surprising if one bears in mind that the item difficulty estimate is much more precise and robust than are the estimates of the discrimination and guessing parameters (Barnes & Wise, 1991; Harrison, 1986; Hulin, Lissak, & Drasgow, 1982). One factor that is known to affect the item parameter estimates is sample size, whereas the individual ability estimate is influenced by the number of items in the test (Barnes & Wise, 1991; Cohen, Kane, & Kim, 2001; Hulin et al., 1982). According to Barnes and Wise (1991), an adequate estimate of the item difficulty parameter requires a minimum of 200 individuals and 20 items, although the results of the present study suggest that item difficulty estimates can be relatively precise even with small samples (130 individuals). The results regarding the precision of parameter estimates in the condition of model misfit provide further evidence for the robustness of the estimates obtained using Winsteps.

With respect to the study objective, one interesting finding concerns the effect on FP and CD rates of sample size ratio, model fit/misfit, and their interaction with the different impact conditions tested. First, it is worth noting the good FP rates obtained under the condition of model fit, where none of the other manipulated variables had any considerable effect. These results suggest that the effect of model fit on the proportion of items incorrectly flagged as DIF items is the main variable and key issue to consider when applying the DIF detection procedure using Winsteps. This finding is similar to the basic assumption of the Rasch model, whereby a necessary condition is that the data fit the model.

Despite the compelling results for model fit, we also found some evidence of adequate FP rates under the condition of model misfit. Given that our aim was to provide information about the behavior of the Difficulty Parameter Differences procedure under different conditions, the results for this condition also provide evidence regarding the robustness of the procedure and, consequently, they were also taken into consideration.

The results show that the proportion of DIF items in the test (0%, 10%, or 20%) did not produce substantially different effects on the Type I error rate regardless of the criterion used to detect DIF. When the criterion for detecting DIF was the significance test, differential effects were observed for a sample size ratio of 20, for which FP rates were above .075 under the following conditions: (a) with model fit and 10% of items with DIF when the ability distribution was the same across groups; (b) with model fit and 20% of items with DIF when the ability distribution was the same across groups or showed a difference in the mean or a difference in variance; and (c) with model misfit and 20% of items with DIF when the ability distribution was the same across groups. In general, there does not appear to be a combined effect of sample size ratio and impact condition on Type I error rates when the model fits. Differential effects are, however, observed in the case of

model misfit: better control of Type I error rates was obtained when the ability distribution was the same across groups, for all sample size ratios, and when the ability distributions differed in variance for sample size ratios of 250, 500, and 1000. With respect to the effect of sample size ratio under the condition of model fit and no items with DIF, our results are similar to those reported for the Mantel-Haenzsel (MH) procedure by Paek and Guo (2011), who examined different conditions of sample size ratio and concluded that the ratio could be increased to 30 or 40 without raising Type I error rates above what was observed with smaller ratios.

In terms of the results obtained when comparing the two simulation models (fit and misfit), it could be argued that if the assumption of a constant discrimination parameter is not fulfilled, as occurs in the model misfit condition, this could be confused with DIF and suggest higher rates of Type I error. DeMars (2010) concluded that when the data have a nonzero asymptote, as is the case when c = .15, the discrimination parameter estimates will be different when there are large differences between the group means. Our results regarding the effect of impact when there is model misfit are similar to those reported by Monahan and Ankenmann (2005, 2010). In their 2005 study, in which they used a magnitude of change similar to that employed here, they tested different levels of impact on both the mean and variance of the ability distribution and found that FP rates were higher when both the mean and variance differed across groups, as compared with when only the variance differed. They obtained the same results (FP rates at the nominal level for equal groups but high for the conditions of impact, mainly when the mean or both the mean and the variance of the ability distribution differed) in their 2010 study, with conditions of impact similar to those used here. Pei and Li (2010) carried out a simulation study involving a 3PL model to assess the effect that differences in the variance of the ability distribution had on the DIF detection rates achieved by various techniques—namely, logistic regression, Mantel-Haenszel (MH), SIBTEST, and IRT-LR-DIF. They found that in the no-DIF condition the sample size ratio had no differential effect on FP rates and that a difference in variance across groups was the variable that had the most marked effect on DIF detection by logistic regression, the MH, and SIBTEST; the effect of differences in the mean ability distribution was minimal, probably because the difference used by these authors was less than that employed in the present study.

One result of interest is that obtained for conditions with DIF (10% and 20% of items) and model fit and for all the DIF conditions considered (0%, 10%, and 20%) and model misfit, since in both cases FP rates were higher with smaller sample size ratios. These results appear to contradict those obtained with other procedures such as the MH statistic, where FP rates are reported to be better controlled when group sizes are similar or equal (Guilera, Gómez-Benito, Hidalgo, & Sánchez-Meca, 2013; Paek & Guo, 2011). However, because our simulated study sample was very large and our interest was in examining the effect of sample size ratio, the smallest ratio we considered was 20; hence, under none of the conditions we examined were group sizes or sample

size ratios similar to those analyzed in other studies. One aspect that remains to be analyzed is the effect of group size when the test contains DIF items. In this respect, the results obtained by Paek and Guo (2011) regarding the accuracy of DIF detection when the size of the reference group is increased (1000, 2000, 3000) while the focal group size remains small (150, 200, or 250) suggest that as the size of the reference group increases (thus producing a larger sample size ratio) the accuracy of DIF detection also increases; hence, it is necessary with increasing focal group size to increase the size of the reference group to maintain or improve the accuracy of DIF detection. Our results show that FP rates were above the nominal level when the focal group size was greater than 6000 (a group size much greater than that used in the majority of studies) and the model fit the data, thereby suggesting that under conditions of DIF and a focal group size of this magnitude it would be necessary to increase the size of the reference group until obtaining a sample size ratio greater than 20 to achieve adequate control of Type I error. Further studies are required to elucidate this issue.

In the conditions with DIF items and model misfit there was an effect of the interaction between impact and sample size ratio: FP rates were higher in the presence of impact, principally when the difference concerned the mean or both the mean and the variance of the ability distribution and when the sample size ratio was between 20 and 100.

Another important finding of this study concerns the measure of the DIF magnitude we used in combination with the significance test. When applying this measure, FP rates were controlled when the model fit the data for all the conditions of sample size ratio, impact, and percentage of items with DIF. With model misfit, FP rates were affected by impact, most notably when the difference concerned the mean or both the mean and the variance of the ability distribution. It should be noted that FP rates were similar for all the conditions of DIF percentage and sample size ratios, indicating that this detection criterion is fairly robust to the presence of DIF and to differences in group sizes.

A further aspect that merits consideration concerns the application of Bonferroni adjustment in our analysis. Although a detailed discussion of the effect of this adjustment is beyond the scope of our objectives, we believe it is worth mentioning as the application of Bonferroni adjustment is not a standard practice in simulation studies about DIF detection procedures, despite the use of significance tests in multiple comparisons when several manipulated variables are included. As expected, FP rates decreased when the Bonferroni adjustment was applied, although perhaps the most interesting aspect is that, with the adjustment, the presence of DIF items does not appear to increase the proportion of incorrectly flagged items. For example, when the criterion for detecting DIF was the significance test under the condition of model fit and 10% DIF, the highest FP rate was .43 without the Bonferroni adjustment, falling to .13 when the adjustment was applied. Using the same detection criterion in the condition of 20% DIF, the FP

rate fell from .62 to .24. Under the same conditions but with model misfit, FP rates fell from .91 to .85 with 10% DIF, and from .95 to .87 with 20% DIF. However, the greatest decreases occurred when the groups did not differ in their ability distribution and the sample size ratio was 20: in this case, FP rates fell from .27 to .029 with 10% DIF and from .49 to .083 with 20% DIF. Readers who are interested in the results obtained prior to applying the Bonferroni adjustment should consult the supplementary material (SM1 and SM2).

Regarding the power of the procedure, this was most affected by sample size ratio: specifically, CD rates were higher with smaller sample size ratios regardless of the model fit and impact condition. However, power reached acceptable to moderate levels only when the sample size ratio was less than 250 under the condition of 10% of items with DIF and when the sample size ratio was less than 100 with 20% of DIF items, despite the fact that item difficulty estimates were relatively adequate even when the focal group size was small (n = 130). This would be an important result to bear in mind in the event of using the Difficulty Parameter Differences procedure to detect DIF in samples with widely differing group sizes, as occurs when educational tests are applied nationally to populations that include minority ethnic or cultural groups.

It is also noteworthy that CD rates were above .80 when the significance test was applied with model misfit under the following conditions: a) 10% of DIF, sample size ratios of 100 and 20, and no differences between the groups in the mean ability distribution and b) 20% of DIF, sample size ratio of 20, and no differences between the groups in the ability distribution or the differences are in the mean. These results may be related to the high CD rates (between .80 and 1) for nonuniform DIF under the condition of model misfit. It is worth noting here that Prieto-Marañón, Aguerri, Galibert, and Attorresi (2012) reported CD rates above .75 with simulated data and the 3PL c = .20 model using decision rules based on Mantel-Haenszel and Breslow-Day tests. However, this effect was not present when we applied the significance test with the measure of DIF magnitude. Whatever the case, further studies are needed to examine CD rates in the presence of model misfit and both uniform and nonuniform DIF, considering a range of values for the item parameters and DIF magnitude.

When the measure of DIF magnitude and the significance test were applied together, CD rates remained adequate (< .70). With model fit and 10% of items with DIF, the effect of impact was minimal with sample size ratios of 250, 100, or 20, whereas in the 20% DIF condition these two variables produced an interaction effect. With model misfit, impact had a marked effect on CD rates in both DIF conditions (10% and 20% of items). Variables that have been reported to affect CD rates include larger sample sizes, low percentages of items with DIF, and balanced DIF (Chen, Chen, & Shih, 2014; Fidalgo, Ferreres, & Muñiz, 2004). In their examination of the MH procedure, Paek and Guo (2011) found that even with small focal group sizes (n = 150 or 250), adequate CD rates were obtained when the reference group was considerably larger. The present

results do not show the same pattern, probably because Paek and Guo considered only one level of impact (difference in the variance of the ability distribution of .5), which was also smaller than what we simulated here, and because they considered a range of -1 to 1 for the item difficulty parameter.

It should also be noted that whereas detection rates for uniform DIF were acceptable under the 10% DIF condition with both model fit and model misfit, detection rates for nonuniform DIF were interpretable only with model misfit, in this case for both the 10% and 20% of DIF conditions.

We believe that the present study is one of the first to examine experimentally the effect of variables such as sample size ratio, model misfit, and differences in the mean and/or variance of the ability distribution on the power and Type I error rates of the Difficulty Parameter Differences procedure, applied in Winsteps with an extreme total sample size (n = 130,000). Although this procedure has been used in numerous empirical studies, the present results highlight a series of conditions under which the Difficulty Parameter Differences procedure may be less robust in terms of the balance it achieves between specificity and sensitivity when detecting DIF. This has important practical implications, not least with regard to the cost of test construction, since if the procedure fails to detect true DIF items or wrongly flags others as showing DIF when in fact they do not, this could lead to inadequate items being retained and useful ones being eliminated from tests. This is particularly relevant given that large-scale test administration is becoming increasingly common and that items and tests are published for use by others following their initial development. In that context it is essential that any DIF items are accurately detected so as to ensure that new items measuring the construct of interest are developed only when strictly necessary.

Given our findings regarding power and Type I error for the conditions simulated here, future studies should focus on evaluating the effect of purification procedures such as iterative linking and on developing metrics that offer a criterion for detecting DIF while achieving adequate control of Type I error rates and improved correct detection rates under extreme conditions. As noted by Gómez-Benito, Hidalgo, and Zumbo (2013), and by Hidalgo, Gómez-Benito, and Zumbo (2014), the combined use of a test statistic and a measure of effect size can reduce the rate of false positives, provided that the cutoffs or criteria used yield significant information in terms of DIF magnitude. In this respect, although the measure of DIF magnitude employed in the present study reduced the rate of false positives, this was accompanied by a loss of specificity for the detection procedure, which then achieved CD rates of between .67 and .09. Consequently, it would be necessary to assess other measures for controlling this aspect or apply a larger DIF magnitude (.64) that is able to detect moderate to large DIF, as suggested by Linacre (2012).

Future research should also examine the effects of other variables, including the difficulty and discrimination level of items with and without DIF, the size of the focal group, and other values for differences in the ability distribution. In a meta-analysis of research on the Mantel-Haenszel procedure, Guilera et al. (2013) concluded that the value of the simulated parameters (difficulty and discrimination) had a moderator effect on Type I error and power, with error increasing as the values of these parameters ceased to be moderate. This was particularly the case for the difficulty parameter. Li, Brooks, and Johanson (2012) similarly found that Type I error rates of the Mantel-Haenszel statistic and logistic regression were inflated when the discrimination parameter showed highly heterogeneous values for the test as a whole or when the value of this parameter was out from .6 and .9 for a given item. This could have affected our results when simulating data with the 3PL c = .15 model, since only one item fell within this interval.

Given the sample size ratios that were simulated in this study it remains unclear whether the observed effects are due to the ratio or to the size of the focal group. Although the literature contains certain observations in this respect, the total sample size and the sample size ratios considered in this study are large and quite different from those examined previously. It would therefore be important to evaluate the real effect of focal group size when using the Difficulty Parameter Differences procedure. It should also be noted that although the sample and group sizes simulated here may seem extreme compared with those of other simulation studies, the sample and group sizes are what one would expect to find in the context of large-scale, high-stakes test administrations, such as the university entrance exam data on which the present analysis is based.

Finally, the present study is one of only a few to provide information about the effect of differences in both the mean and variance of the ability distribution, with our results being similar to those reported by Monahan and Ankenmann (2005, 2010) and Pei and Li (2010). A task for future studies, however, would be to examine the effect of applying a difference of between .5 and 1 in the mean and variance of the ability distribution, together with the parameter values.

Practical Recommendations

First, it is important to note that the results for FP rates were better with model fit, since both the Type I error rate and the effect of the manipulated variables was minimal. This indicates that in practice it is highly advisable to ensure that the 1PL model fits the data. Following on from this condition, and bearing in mind the results for the power of the procedure and a comment made by one of the reviewers of this paper, users should also consider that an increase in DIF magnitude would improve statistical power, even though this aspect was not examined in the present study. For example, if we take the value of DIF magnitude recommended by Linacre (2012) for detecting moderate to large DIF (namely .64), it is likely that the power of the procedure will reach a level close to or above .80, thus achieving a better balance in terms of the Type I error rate and statistical

power under the condition of model fit when applying the significance test in combination with the measure of DIF magnitude as the criterion for detecting DIF.

Second, it should be noted that the effect of the impact variable is much better controlled when the combined criterion for detecting DIF is applied. In practice, users may have no knowledge about the presence or otherwise of impact in the attribute variable. Our main recommendation here is to apply the Difficulty Parameter Differences procedure with the combined criterion of the significance test and the measure of DIF magnitude, as this proved to be more robust to the possible effects of impact. At all events, it is important to bear in mind that although differences between groups in the attribute distribution was, here, an arbitrarily and experimentally manipulated variable about which applied users may have no a priori knowledge, users may be able to estimate the behavior of certain population groups with regard to certain variables, or make a posteriori calculations, in which case our results for this variable are applicable.

In terms of the practical recommendations that follow from the specific values of the variables we manipulated, we believe it is useful to highlight those conditions under which the procedure achieved adequate statistical power and control of Type I error rates. The following recommendations may serve as a guide for users, indicating those conditions in which the Difficulty Parameter Differences procedure is adequate.

The Difficulty Parameter Differences procedure with the significance test as criterion is appropriate when a test contains 10% of DIF items, if the groups have the same ability distribution (or at least the same mean), and when there are 100 examinees in the reference group for each one in the focal group. However, it should be noted that FP rates will be between .10 and .004, and the power of the procedure will be in the range .67 to .89.

If one is sure that the model fits the data, then the Difficulty Parameter Differences procedure may be used with confidence with sample size ratios of 500, provided the ability distribution is the same across groups, and also with ratios of 250, 100, or 20, since under these conditions FP rates are controlled for any difference between groups in the ability distribution. However, even when the model fits the data, researchers should bear in mind that CD rates are unlikely to exceed .76 when the sample size ratio is greater than 250.

In the event that the data fit a 3PL model with c = .15 and a test has 10% of DIF items, the Difficulty Parameter Differences procedure achieves adequate FP and CD rates when the groups do not differ in ability and when sample size ratios are between 20 and 100. In this case, FP rates are below .03 and CD rates range between .84 and 1. Practitioners may also consider a sample size ratio of 250 under the aforementioned condition or when the difference is in the variance of the ability, since in that case FP rates will be below .041, although caution should be exercised with respect to CD rates, which were found to be in the range of .52 to .60.

When 20% of a test's items show DIF, the procedure should only be used with model fit when the groups do not differ in ability or differ only in the variance of the ability distribution and with a sample size ratio of 100. In this case, FP rates are below .021 and CD rates will be between .64 and .67

If the 1PL model fits the data and the test has 10% of DIF items, then the combined application of the significance test and a measure of DIF magnitude enables Type I error rates to be controlled and adequate CD rates to be achieved (.62 to .67) for all the conditions of impact that were simulated in this study and with sample size ratios of 250, 100, or 20. If 20% of test items show DIF, practitioners may use the procedure with a sample size ratio of 20 under any condition of impact and with a ratio of 100 when the groups do not differ in ability or differ only in the variance of the ability distribution. However, if the data fit a 3PL model with c = .15 and the test has 10% or 20% of DIF items, we do not recommend using the Difficulty Parameter Differences procedure under any of the conditions considered in this study.

Acknowledgments

The authors would like to thank the members of the research group Methods and Instruments for Behavioral Sciences Research for their observations and suggestions over the course of this study and especially Victor Cervantes for generating the data simulation routines used. We are also grateful to Professor Mike Linacre for giving us free access to a Winsteps license, which was required to carry out in this research, and for his invaluable advice and rapid response to all our queries.

Funding

This study was partially funded by the Instituto para la Evaluación de la calidad de la Educación (ICFES), the Departamento Administrativo de Ciencia, Tecnología e Innovación, COLCIENCIAS [Project 255-2011], and the Agency for the Management of University and Research Grants of the Government of Catalonia [Grant No. 2014SGR1139].

ORCID

Ángela I. Berrío: https://orcid.org/0000-0003-2064-4594

Juana Gómez-Benito: https://orcid.org/0000-0002-4280-3106

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and Practice* (pp. 3-24). London, UK: Lawrence Erlbaum Associates.

Barnes, L. L., & Wise, S. L. (1991). The utility of modified one-parameter IRT model with small sample. *Applied Measurement in Education, 4*(2), 143-157. doi:10.1207 /s15324818ame0402_4

Bradley, J. (1978). Robustness. *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152. doi:10.1111/j.2044-8317.1978.tb00581.x

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences*. London, UK: Lawrence Erlbaum.

Chen, J. H., Chen, C. T., & Shih, C. L. (2014). Improving the control of Type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement*, *38*(1), 18-36. doi:10.1177/0146621613488643.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44. doi:10.1111/j.1745-3992.1998.tb00619.x

Cohen, A. S., Kim, S. H., & Subkoviak, M. J. (1991). Influence of prior distributions on detection of DIF. *Journal of Educational Measurement*, *28*(1), 49-59. doi:10.1111/j.1745-3984.1991.tb00343.x

Cohen, A. S., Kane, M. T., & Kim, S. H. (2001). The precision of simulation study results. *Applied Psychological Measurement*, *25*(2), 136-145. doi:10.1177/ 01466210122031966

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*(6), 961-972. doi:10.1177/ 0013164410366691

Draba, R. E. (1977, March). The identification and interpretation of item bias. *Mesa Memorandum*, 25 Retrieved from http://www.rasch.org/memo25.htm

Feinberg, R. A. & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36-49. doi:10.1111/emip.12111

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error Rates. *Journal of Experimental Education, 73*(1), 23-39. doi:10.3200/JEXE.71.1.23-40

Finch, H. W., & French, B. F. (2008). Anomalous Type I error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement*, *68*(5), 742-759. doi:10.1177/ 0013164407313370

Finch, H.W., & French, B. F. (2014). The impact of group pseudo-guessing parameter differences on the detection of uniform and nonuniform DIF. *Psychological Test and Assessment*

Modeling, *56*(1), 25-44. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2014_20140324/02_Finch.pdf

Gómez-Benito, J., Hidalgo, M. D., & Padilla, J. L. (2009). Efficacy of effect size measures in logistic regression. An application for detecting DIF. *Methodology*, *5*(1), 18-25. doi:10.1027/1614-2241.5.1.18

Gómez-Benito, J., Hidalgo, M. D., & Zumbo, B. D. (2013). Effectiveness of combining statistical tests and effect size when using logistic discriminant function regression to detect differential item functioning for polytomous items. *Educational and Psychological Measurement*, *73*(5), 875-897. doi:10.1177/ 0013164413492419

Goodman, J. T., Willse, J. T., Allen, N. L., & Hlaric, J. S. (2011). Identification of differential item functioning in assessment booklet designs with structurally missing data. *Educational and Psychological Measurement*, *71*(1), 80-94. doi:10.1177/0013164410387341

Guilera, G., Gómez-Benito, J., Hidalgo, M.D., & Sánchez-Meca, J. (2013). Type I error and statistical power of the Mantel Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods*, *18*(4), 553-571. doi: 10.1037/a0034306

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London, UK: Sage.

Harrison, D. A. (1986). Robutness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, *11*(2), 91-115. doi:10.3102/10769986011002091

Hauser, C., & Kingsbury, G. (2004). *Differential item functioning and differential test functioning in the Idaho Standards Achievement Test for Spring 2003*. Idaho: Northwest Evaluation Association.

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

Hidalgo, M. D., Gómez-Benito, J., & Zumbo, B. D. (2014). Binary logistic regression analysis for detecting differential item functioning: Effectiveness of R-squared and delta log odds ratio effect size measures. *Educational and Psychological Measurement*, 74. doi:10.1177/ 0013164414523618

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-146). Hillsdale, NJ: Lawrence Erlbaum.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Rasch Measurement*, *6*(3), 249-260. doi:10.1177/014662168200600301

Lai, J. S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of defferential item functioning (DIF) for small sample sizes. *Evaluation and the Health Professions*, *28*(3), 283-294. doi:10.1177/0163278705278276

Li, Y., Brooks, G., Johanson, G. A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, *72*(5), 847-861. doi: 10.1177/0013164411432333.

Linacre, J. M., & Wright, B. D. (1987). *Item bias: Mantel Haenszel and the Rasch model*. Chicago, IL: Psychometric Laboratory, University of Chicago.

Linacre, J. M. (2006). *Winsteps-Ministeps: Rasch model computer programs.* [Computer Software]. Chicago, USA.

Lei, P. W., Chen, S. Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptative testing enviroment. *Journal of Educational Measurement*, *43*(3), 245-264. doi: 10.1111/j.1745-3984.2006.00015.x

López, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251-265. doi:10.1177/0146621608321760

Magis, D., Tuerlinckx, F. & De Boeck, P. (2015). Detection of differential item functioning using the lasso aproach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111-135. doi: 10.3102/1076998614559747

Monahan, P. O., & Ankenmann, R. D. (2005). Effect of unequal variances in proficiency distributions on Type I error of the Mantel-Haenszel chi-square test for differential item functioning. *Journal of Educational Measurement*, *42*(2), 101-131. doi:10.1111/j.1745-3984.2005.00006

Monahan, P. O., & Ankenmann, R. D. (2010). Alternative matching scores to control types I error of the Mantel-Haenszel procedure for DIF in dichotomously scored items conforming to 3PL IRT and nonparametric 4PBCB models. *Applied Psychological Measurement*, *34*(3), 193-210. doi:10.1177/0146621609359283

Navas, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment*, *18*(1), 9-15. doi:10.1027//1015-5759.18.1.9

Paek, I., & Guo, H. W. (2011). Accuracy of DIF estimates and power in unbalanced designs using the Mantel-Haenszel DIF detection procedure. *Applied Psychological Measurement*, *35*, 518-535. doi:10.1177/0146621611420559

Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likehood ratio for DIF detection. *Applied Psychological Measurement*, *34*(6), 453-456. doi:10.1177/0146621610367789

Prieto-Marañón, P., Aguerri, M. E., Galibert, M. S., & Attorresi, H. F. (2012). Detection of differential item functioning using decision rules based on Mantel-Haenszel procedure and Breslow-Day Tests. *Methodology, 8*(2), 63–70. doi:10.1027/1614-2241/a000038

Raju, N. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495-502. doi:10.1007/BF02294403

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen, Denmark: Danish Institute for Educational Research.

R Development Core Team (2007). R Software (Versión 2.6). [Computer Software]. Retrieved from: http://www.r-project.org

Schulz, E. M., Perlman, C., Rise, W. K., & Wright, B. D. (1992). An empirical comparison of Rasch and Mantel Haenszel procedures for assessing differential item functioning. In G. Engelhard Jr. & M. Wilson (Eds.), *Objective measurement*: *Theory into practice* (pp. 65-82). Greenwood.

Shepard, L., Camilli, G. & Wiliams, D. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*(2), 77-105. doi:10.1111/j.1745-3984.1985.tb01050.x

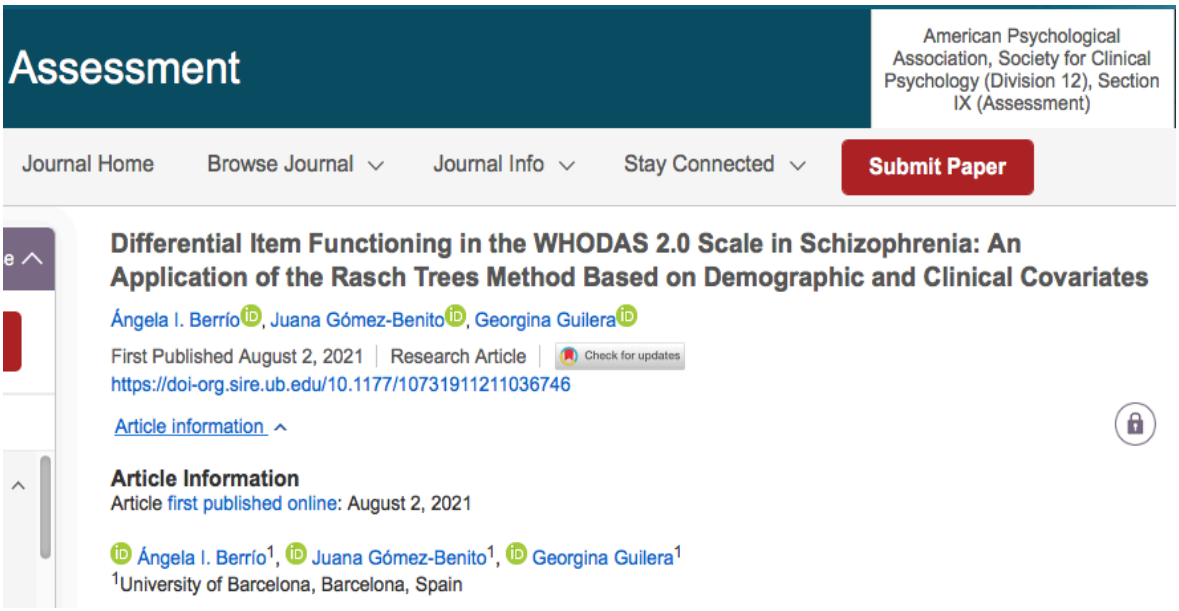
Wang, W.C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*(3), 221-261. doi:10.3200/JEXE.72.3.221-261

Wang, W. C., & Yeh Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479-498. doi:10.1177/0146621603259902

Wright, B., Mead, R., & Draba, R. (1976, October). Detecting and correcting test item bias with a logistic response model. *Mesa Memorandum*, 22. Retrieved from http://www.reasch.org/memo22.htm

3.3. Estudio 3: Análisis con datos empíricos

3.3.1. Datos identificativos



Título: Differential Item Functioning in the WHODAS 2.0 Scale in Schizophrenia: An Application of the Rasch Trees Method Based on Demographic and Clinical Covariates.

Autoras: Ángela I. Berrío, Juana Gómez-Benito & Georgina Guilera

Año: 2021

Revista: Assessment

Métrica según Journal Citation Reports para el año 2020:

- Factor de Impacto: 4,667
- Rango por factor de impacto: 23 de 131, Cuartil: Q1, Categoría: Psychology, Clinical (SSCI).
- Rango por Journal Citation Indicator (JCI): 14 de 172, Cuartil: Q1 (D1), Categoría: Psychology, Clinical (SSCI).

Material Suplementario: se puede consultar en el anexo 3.

DOI: https://doi.org/10.1177/10731911211036746

3.3.2. Resumen

Debido a que en el primer estudio uno de los seis métodos estudiados con datos simulados más recientemente fue el método basado en el modelo Rasch y que entre los procedimientos que se incluyeron en esta categoría los referidos a árboles de Rasch fueron las dos propuestas más recientes, se decidió ilustrar la aplicación de dichos procedimientos para analizar la presencia de DIF en la escala WHODAS 2.0 en personas diagnosticadas con esquizofrenia. Esta decisión se fundamentó en las siguientes razones: en primer lugar, no se encontraron estudios que analizaran el funcionamiento diferencial en la escala WHODAS con personas con esquizofrenia y aunque el grupo GEIMAC ya había aportado información previa sobre las propiedades psicométricas de la escala (Galindo-Garre et al., 2015; Guilera et al., 2012), el DIF no se había analizado. En segundo lugar, la escala WHODAS ha sido de las más empleadas para evaluar la discapacidad y funcionalidad, y en personas con esquizofrenia se han identificado algunas diferencias en el puntaje obtenido en dicha escala entre subgrupos de pacientes de acuerdo con variables de carácter demográfico y clínico. En tercer lugar, se consideró particularmente pertinente la aplicación de estos procedimientos ya que no requieren una definición previa de las personas que conformarán el grupo de referencia y el focal, lo cual resulta una ventaja especialmente cuando se trata de variables de tipo continuo. En este estudio las variables clínicas emplearon los puntajes en dos escalas ampliamente aplicadas en personas con esquizofrenia y por tanto estas variables fueron de tipo cuantitativo continuo. En quinto lugar, se consideró que la aplicación de estos procedimientos en una escala utilizada en el ámbito de la salud, teniendo en cuenta las ventajas que tienen respecto de los procedimientos clásicos, permitiría aportar evidencias de su aplicabilidad. Finalmente, como todo procedimiento basado en el modelo Rasch éstos podrían resultar apropiados en el ámbito de la salud porque su requerimiento en cuanto a tamaño de muestra es menor en comparación con otros clásicos, se pueden aplicar a ítems en escala Likert y en cuanto al número de ítems tampoco se requiere de tests muy extensos.

Además de analizar la presencia de DIF en la escala WHODAS 2.0 en una muestra de personas con esquizofrenia, con este estudio se pretendió ilustrar la aplicación de procedimientos que no requieren la especificación a priori de los grupos focal y referencia e identificar las similitudes e información complementaria que dichos procedimientos pueden proporcionar al clínico interesado en analizar DIF.

Para cumplir los objetivos del estudio se analizaron las respuestas de 352 personas con esquizofrenia a la escala WHODAS 2.0, de ellas 280 respondieron los 36 ítems que conforman la escala y fueron incluídas en este estudio. La detección de DIF se realizó con los procedimientos PCM-IFT y Tree-PCM que requirieron la exclusión de las respuestas de aquellas personas que respondieron con la misma categoría de respuesta a todos los ítems en un dominio, y aplicaron un ajuste de Bonferroni para controlar los falsos positivos que podrían presentarse por las múltiples comparaciones en la detección de DIF con varias covariables.

Los resultados mostraron que sólo el procedimiento PCM-IFT detectó DIF en un único ítem (ítem 5: actividades sexuales, del dominio 4: Relaciones con otras personas). El DIF fue detectado por la variable edad con un punto de corte a la edad de 41 años indicando que las personas con esquizofrenia mayores de 41 años reportaron mayor dificultad al realizar actividades sexuales que aquellos con 41 años o menos cuando sus niveles de discapacidad respecto al dominio de relaciones con otras personas están igualados.

Por su parte, Tree-PCM no detectó DIF. Sin embargo, los resultados para el dominio 4: Relaciones con otras personas indicaron que relacionarse con personas que no conoce y actividades sexuales (ítems 1 y 5, respectivamente) fueron las actividades más problemáticas de realizar para las personas con esquizofrenia.

Además, los dos procedimientos de detección de DIF en el dominio 4 encontraron que en el ítem 5 se presentó un menor valor para el umbral 2 respecto del umbral 1, lo cual podría indicar que muy pocas personas respondieron a la categoría de

respuesta 2 (leve dificultad), que no hay un valor del atributo en el que sea más probable responder con dicha categoría de respuesta o que personas que tienen un nivel de atributo alto (alta discapacidad respecto de las relaciones con otras personas) indicaron una dificultad leve en las actividades sexuales.

3.3.3. Versión postprint

Differential Item Functioning in the WHODAS 2.0 Scale in Schizophrenia: An Application of the Rasch Trees Method Based on Demographic and Clinical Covariates

Ángela I. Berrío[1,2], Juana Gómez-Benito[1,2], and Georgina Guilera[1,2]

[1] Quantitative Psychology Unit, Department of Social Psychology and Quantitative Psychology, Faculty of Psychology, University of Barcelona, Barcelona, Spain
[2] Group on Measurement Invariance and Analysis of Change (GEIMAC), Institute of Neurosciences, University of Barcelona, Barcelona, Spain.
Corresponding Author: Ángela I. Berrío, Quantitative Psychology Unit, Department of Social Psychology and Quantitative Psychology, Faculty of Psychology, University of Barcelona, Passeig Vall d'Hebron 171, 08035 Barcelona, Spain. E-mail address: aiberriob@ub.edu

Abstract

Identifying disability score differences in people with schizophrenia according to sociodemographic and clinical variables can help design better rehabilitation or care programs, but in order to compare the scores, it is necessary to confirm the measurement invariance. This study analyses differential item functioning (DIF) in the WHODAS 2.0 (WHO Disability Assessment Schedule) by applying two procedures based on Rasch trees (TREE-PCM and PCM-IFT). A total of 352 patients with schizophrenia spectrum disorder aged between 18 and 55 years took part. Sociodemographic (gender, age, marital status, and education) and clinical (depressive symptomatology, and presence of positive and negative symptoms) covariates were analysed in each of the WHODAS 2.0 domains. The TREE-PCM did not detect DIF, while with PCM-IFT an item with DIF was detected for the age variable. Although the findings suggest that only one item presents DIF, this refers to important issues when assessing functioning in patients with schizophrenia and should be reviewed.

Keywords: DIF, WHODAS 2.0, Schizophrenia, Rasch trees method, TREE-PCM, PCM-IFT.

Since the World Health Organisation (WHO) has acknowledged that disability has become as important as mortality due to people living longer and its corresponding impact on increasing chronic diseases, disability assessment has taken on particular relevance in both the assessment and the impact of the disease itself and the effectiveness and performance of the health system (Üstün et al., 2010). The WHO Disability Assessment Schedule (WHODAS 2.0) is an instrument developed to evaluate disability across cultures and illness, based on a comprehensive set of International Classification of Functioning, Disability and Health (ICF) items (Üstün et al., 2010). The utility and psychometric properties of WHODAS 2.0 in people with schizophrenia have already been analysed (Galindo-Garre et al., 2015; Guilera et al., 2012), demonstrating that this

instrument boasts suitable evidence of validity and, in general, satisfactory psychometric properties. According to the study by Guilera et al. (2012), each of the WHODAS 2.0 domains presented an internal consistency in the acceptable to excellent range (Cronbach's α between 0.70 to 0.91), and in terms of its factorial structure, the six-domain model was the best fit compared with the single second-order factor model and an alternative model with two second-order factors that reflects the two dimensions of disability in terms of the ICF model (activity limitations and participation). Galindo-Garre et al. (2015) provided additional evidence on the unidimensionality of each of the WHODAS 2.0 domains, the appropriate discrimination of items at different levels of disability, and the increase in the latent trait when the total score increases.

In addition, some studies have found differences between subgroups of people with schizophrenia in the WHODAS 2.0 scores by variables such as depression (Akinsulore et al., 2015; Dan et al., 2011; Guilera et al., 2012; McKibbin et al., 2004; Sjonnense et al., 2016; Strassnig et al., 2018), symptomatology (Akinsulore et al., 2015; R. Chen et al., 2019; Ertugrul & Ulug, 2004; Guilera et al., 2012; Park et al., 2019; Strassnig et al., 2018), occupational status (Adegbaju et al., 2013; R. Chen et al., 2019; Guilera et al., 2012; Olagunju et al., 2016; Strassnig et al., 2018), smoking (Aguocha et al., 2018), gender (Adegbaju et al., 2013), years of education and residence (R. Chen et al., 2019), age (Adegbaju et al., 2013; R. Chen et al., 2019; McKibbin et al., 2004; Olagunju et al., 2016), marital status (Adegbaju et al., 2013; Akinsulore et al., 2015), ethnicity and number of children (Olagunju et al., 2016). In order to confirm that the differences indicated in the WHODAS scores are valid and are not due to subgroup measuring problems, an invariance analysis must be performed. This analysis essentially consists of comparing psychometric characteristics both at test level and for the items between subgroups of people defined by the aforementioned variables (depression, symptomatology, occupational status, etc.). To confirm invariance, the psychometric characteristics must be maintained between the subgroups, otherwise invariance is not confirmed and the comparison of the WHODAS scores between these subgroups is not reliable. One strategy for drawing conclusions about the lack of invariance at item level is called differential item functioning (DIF).

Since the presence of DIF indicates that the probability of a correct response among equally able test-takers is different for subgroups that can be defined by gender, race, age or other variables (Strobl et al., 2015), the DIF analysis provides evidence about the test's validity. In other words, DIF indicates that the probability of endorsement of one response category among people or patients who have the same level of disability is different for subgroups of people defined by covariates such as gender, age, and so on.

In mental health, in view of the possible implications for the patient of an appropriate evaluation both at diagnosis and in the course of the disease and in the rehabilitation stage, DIF takes on particular relevance. For example, if it is asserted that people with schizophrenia and with

depressive symptomatology report more disability compared with those who do not have these symptoms, many mental health programmes would establish this variable as fundamental to the patient's recovery. However, if DIF exists for this covariate, the comparison of the WHODAS scores between people with schizophrenia with or without depressive symptomatology loses its validity and indicates that two people with the same level of disability, one with depressive symptoms and the other without, have a different likelihood of reporting a major or minor problem when carrying out the activities asked about in the WHODAS items. For example, a person with depressive symptomatology will endorse a response category which indicates a greater problem, whereas the opposite happens with a person without depressive symptomatology. If the foregoing is fulfilled, the differences indicated by various authors, in regard to the depression variable in the WHODAS are due to other aspects and not to the presence of depressive symptomatology. Therefore, this covariate could lose explanatory power in this specific case.

Despite the relevance of studying DIF in scales such as WHODAS, only one previous study tested DIF in a population-based sample of myocardial infarction patients for age, gender, education, marital status, presence of comorbidities, overall health status, and smoking status, in which no DIF was detected for any of the variables (Kirchberger et al., 2014). However, to our knowledge, no DIF study has been performed for this scale in schizophrenia that makes it possible to confirm, or not confirm, the invariance of the items in the variables for which other authors have found differences. In addition, the DIF study for WHODAS 2.0 domains is especially relevant in a population with schizophrenia due to the significant floor effect in several domains that Guilera et al. (2012) detected.

In terms of the results regarding invariance of the WHODAS scale, some studies have been carried out especially in the young or adult population. Generally speaking, the results provide evidence of invariance in the WHODAS 12-item version. Kimber et al. (2015) found measurement invariance in the general young population, while Tompke et al. (2020) found partial measurement invariance in all domains except in the cognition and participation domains between young people with and without physical or mental conditions. As regards the adult population, the results show partial invariance between adults with a mental disorder and their carers in all domains except cognition (Zhou et al., 2020).

A variety of methods have been proposed to detect DIF, in their classic form, most of these methods require subgroup prespecification that will be analysed for the presence of DIF, which could be particularly problematic where continuous variables are concerned. These include the Mantel Haenszel test (Holland & Thayer, 1988), Logistic Regression procedures (Swaminathan & Rogers, 1990), Lord chi-squared test (Lord, 1980), difference in the difficulty parameter (Linacre & Wright, 1987), IRT-Likelihood Ratio (Thissen et al., 1993) and many more. The most studied variables in DIF detection have been gender, race and ethnicity, all with previously defined

subgroups. Another variable which is fairly frequently found in studies on DIF is age. As this is a continuous variable, it requires predetermination of the groups to be analysed, defining the cutoff point for this variable using criteria established by the investigator, which may not necessarily comply with the behaviour of the variable in question.

Some disadvantages have been mooted for the classic methods. First, when this concerns continuous variables, the cutoff point for dividing the subgroups is established in advance by the investigator, and although this choice is based on some established criteria, it may generate a loss of information in terms of the presence of DIF (Strobl et al., 2015) since all possible cutoff points are not tested. Second, when DIF is tested, the assumption of non-DIF for the other items does not hold in all situations and is therefore not a suitable interpretation (Magis et al., 2010; Tutz & Berger, 2016). Third, the detection of DIF in an item does not mean that all the examinees that belong to a subgroup are consistently at an advantage or disadvantage (Y.F. Chen & Jiao, 2014; Cohen & Bolt, 2005).

To solve these drawbacks, an intermediate approach based on the Rasch model was recently developed, the Rasch Trees Method. It uses manifest covariates and all the possible subgroups that these variables permit are evaluated. This maintains a suitable degree of interpretability and a sufficient set of DIF indicators is explored (Strobl et al., 2015) without the need to specify the subgroups of examinees a priori.

Rasch Trees Method

This method uses recursive partitioning techniques to automatically identify the subgroups of examinees or items that present DIF. If the results show that the data has partitions, subgroups or nodes, DIF will exist. If, on the other hand, it is kept as a single set, the data is fully adjusted to the Rasch model and is therefore invariable, that is, there is no DIF.

The Rasch Trees Method recognises two procedures. Strobl et al. (2015) propose that when dichotomous items are applied, subgroups or regions of the covariates in which the measurement invariance is not maintained are identified based on a set of covariates that are recursively partitioned (Rasch-TREE). Subsequently, Komboz et al. (2018) published an extension of their procedure to evaluate polytomous items (TREE-PCM). It should be noted that this procedure does not automatically identify the items responsible for the DIF but rather offers more general detection at a covariate level, in other words it identifies the covariate level that can cause differential functioning in a scale, although the parameter value of the items in each of these groups can be obtained.

In order to carry out this procedure, the parameter of the items is estimated using the entire sample, the stability of these parameters is evaluated with respect to each of the covariates of interest and, if significant instability is detected, the sample is divided by the covariate that presents the strongest instability at the cutoff point that leads to the greatest improvement in the fit of the

model. These steps are repeated consecutively with the resulting subsamples until there is no longer significant instability or the subsample is too small (Strolb et al., 2010). The partitions that are generated by the presence of DIF are called nodes. For example, if we find DIF for the gender variable, our sample will have one partition and two nodes (one for men and the other for women).

An important aspect to bear in mind is that this procedure applies a Bonferroni correction regarding the p value taken as a criterion to generate the partitions, controlling the false positive rate due to the multiple comparisons in the DIF detection in several covariates.

The other procedure was proposed by Tutz and Berger (2016) and is called the Item-Focused Trees (IFT). In it, DIF is detected at item level, identifying those items responsible for the noninvariance. Bollmann et al. (2018) presented the application of the procedure in polytomous items (PCM-IFT). This method also uses recursive partitioning for DIF detection focused on the items that produce it. For each item, the total data matrix is used as a starting point, and the different covariates and all their corresponding split-points, that is, the cutoff points that can generate a division of the data matrix, are successively tested.

This method tests the null hypothesis of no differences between the item parameter on a set of variables. The DIF occurs in the variable with the greatest value in the statistical test, when a split-point presents differences in the item parameter and then nodes that represent the covariate partitions are generated. For this purpose, all possible split-points are taken into account, permutating the values of the covariate *n* times (*n* = number of permutations) and calculating the hypothesis test with a p value = α /m, where m = number of covariates under study (p value is adapted in the following steps, according to the number of covariates that remain under study—e.g., when the procedure takes the second variable p value = α /m - 1). For each resulting node, the values of the item's parameter are presented and it should be kept in mind that the structure of the trees may vary between the test items, contrary to what happens with the TREE-PCM procedure. It can clearly be seen that although the two Rasch Trees procedures approach DIF analysis from a different perspective, the information they provide is complementary and it is interesting to apply them simultaneously. While it is true that the TREE-PCM procedure focuses on the detection of differential functioning at a more general level relating to the covariate in the test, which could be similar to the detection of differential test functioning, the PCM-IFT performs the detection of differential functioning directly at item level, analysing the performance of each item in depth.

It should be taken into account that the PCM-IFT procedure tests the null hypothesis that represents the equality of the difficulty parameter of an item between the reference and focal groups. To accept the null hypothesis, the item must obtain the same value of the parameter in the comparison groups. Therefore, this procedure can answer questions such as: which item(s) present DIF by gender, age, or other variables of interest? What subgroup of examinees—defined by variables such as age, anxiety level, gender, and so on—are seen affected by the presence of DIF in

a given item? Note that the focus is directly on the performance of the item and the variables analysed.

For its part, the TREE-PCM procedure tests the null hypothesis that represents the equality of the difficulty parameter of all the items of a test as a whole, which is why it is stated that this procedure is aimed at evaluating differential functioning at the level of the test. To accept the null hypothesis, all the test items must be free of DIF, but to reject it, there must be sufficient instability between the parameters of the items as a whole. In this case, questions may be answered such as: Does the W test show differential functioning with respect to variables such as gender, age, or others? In what subgroups of examinees—defined by variables such as age, anxiety level, gender, and so on—is the W test inadequate due to the presence of differential functioning? Note that the focus of attention is on the test as a whole and the variables that are analysed.

Regarding the efficacy of the Rasch Trees Method against other more commonly used procedures, on the one hand Strobl et al. (2015) and Tutz and Berger (2016) have studied the behaviour of the Rasch-TREE and IFT procedures, respectively, for dichotomous items. On the other, Komboz et al. (2018) and Bollmann et al. (2018) have analysed the behaviour of the TREE-PCM and IFT-PCM procedures, respectively, in polytomous items. They have provided studies with simulated data which as a whole indicate that the procedures based on Rasch Trees present Type I error and power rates which are comparable with those obtained using other classical procedures (Mantel Haenszel, Logistic Regression and Likelihood Ratio Test). However, Type I error rates tend to be lower when the Rasch Trees Method is applied, especially when DIF is generated by the interaction of more than one covariate, and power rates tend to be notably higher in the Rasch Trees Method when the cutoff point which generates the reference group and focus group in the covariates of interest is not known.

Taking the foregoing into account, the purpose of this study was twofold; on the one hand to perform, for the first time ever, a DIF analysis of WHODAS 2.0 with a sample of people with schizophrenia, with a special emphasis on clinical variables such as depressive symptomatology and the presence of positive and negative symptoms, as a source of DIF, without neglecting demographic variables such as gender, age, marital status, and education; and on the other hand, to apply new procedures to detect DIF without the a priori specification of subgroups and to identify the similarities, differences, and additional information between these types of procedures. The choice of the clinical covariates for analysing the presence of DIF is one of the strengths of this article, since they have not been considered in previous studies and it has been demonstrated that these are important variables in practice for clinicians and for patients, both in the diagnostic process and in the treatment and recovery processes. Furthermore, sociodemographic variables have been considered to supplement the DIF analysis.

Method

Participants

A total of 352 persons with schizophrenia treated in various centres and hospitals throughout Spain met this study's inclusion criteria and gave their written consent. Participants were included if they were aged between 18 and 55 years, met the diagnostic criteria for schizophrenia spectrum disorder established by the DSM-IV-TR (American Psychiatric Association, 2000), assessed by clinicians with experience in mental health, and were able to read and write in Spanish. Persons who were participating in a clinical trial or who were experiencing an episode of major depression or another medical or neurological condition or psychiatric disorder that required treatment were excluded.

The mean age of participants was 36.70 years (SD = 8.27) and an average period of 11.72 (SD = 8.46) years had elapsed since the diagnosis. The majority of the participants were male (69.3%), unemployed (67.6%), had reached secondary school (47.9%), were single (83.8%) and were living with their original family (68.2%). All participants were Caucasian. Table 1 shows the participants' socio-demographic characteristics in greater detail. A total of 280 patients with schizophrenia answered all the questions related to the covariates of interest in this study.

Table 1. Sociodemographic characteristics of the participants.

Variable	n	%	Variable	n	%
Educational level			Occupational status		
Functional illiterate	7	2.0	Student	15	4.3
Primary school	121	34.5	Wage earner	74	21.0
Secondary school	168	47.9	Self-employed	13	3.7
University	55	15.7	Non-paid work	3	0.9
			Disability	136	38.6
			Unemployed	53	15.1
			Retired	29	8.2
			Housewife	9	2.6
			Other	20	5.7
Marital status			Living arrangement		
Single	294	83.8	Alone	37	10.5
Married	25	7.1	Original family	240	68.2
Living with partner	11	3.1	Own family	50	14.2
Separated or divorced	19	5.4	Friends	5	1.4
Widow or widower	2	0.6	Sheltered housing	12	3.4
			Other (psychiatric rehabilitation unit, residence for the elderly)	8	2.3

Instruments

The WHODAS 2.0 (Vázquez-Barquero et al., 2006; World Health Organization, 2000) is a generic instrument for measuring disability based on the ICF framework. To assess the level of disability and everyday functioning, the WHODAS 2.0 covers six domains: D1: Cognition–Understanding and Communicating; D2: Mobility–Moving and Getting Around; D3: Self-Care–Attending to one's Hygiene, Dressing, Eating, and Staying Alone; D4: Getting Along–Interacting

with Other People; D5: Life Activities–Domestic Responsibilities, Leisure, Work, and School; and D6: Participation–Joining in Community Activities, Participating in Society. In each domain, participants had to answer questions about how much difficulty they have had to do something during the last 30 days on a 5-point scale ranging from 1 (*none*) to 5 (*extreme/cannot do*). The higher the score the person obtains, the greater their disability. All participants were assessed by means of the 36-item version of the WHODAS 2.0, although those not working or studying answered 32 items since domain D5 of WHODAS 2.0 has four items which should only be answered by people who work or study.

The Hamilton Rating Scale for Depression (HAM-D; Hamilton, 1960; Ramos-Brieva & Cordero Villafáfila, 1986) is one of the most widely used scales for measuring depression and continues to be one of the most commonly used for patients with schizophrenia. It was used as a clinical variable indicative of depressive symptomatology. The HAM-D consists of 17 items related to depressive symptoms experienced during the past week on a scale of 5 (*from absent to very severe*) or 3 (*from absent to definite*) response categories. The total score may vary from 0 to 52; a score of less than 8 indicates no depressive symptomatology, while a score higher than 22 indicates severe depressive symptomatology. In patients with schizophrenia this scale has been widely used and has reported adequate sensitivity (≥ 0.76) and specificity (≥ 0.88), in addition to showing some effectiveness in distinguishing between patients with mild, moderate and severe depression (Müller et al., 2006). However, uncertainties regarding its functioning have been established with respect to patients in the symptomatic phase, concluding that the scale may be of value when used to evaluate patients with schizophrenia in the symptom remission phase (Grover, et al., 2017). In terms of the reliability of the test, Bagby et al. (2004) found that out of 13 studies, 10 reported a Cronbach's $\alpha \geq 0.70$ and 3 studies between 0.46 and 0.88, and that in general, the majority of the HAM-D psychometric properties are adequately met (internal consistency, interrater coefficient, and criterion, convergent and discriminant validity). Similar findings were obtained by Lako, et al. (2012), who, in a systematic review reported an internal consistency in the range of 0.73 to 0.77, a test—retest reliability of between 0.65 and 0.80, an interrater reliability of 0.93 to 0.95, a concurrent validity of 0.80 and 0.74 with the MADRS (Montgomery—Asberg Depression Rating Scale) and the CDSS (Calgary Depression Scale for Schizophrenia) respectively, a specificity of 0.83, and sensitivity of 0.79. Henceforth this variable is labelled as 'depression'.

The severity of the positive and negative symptoms was quantified by the Positive and Negative Syndrome Scale (PANSS; Kay et al., 1987; Peralta-Martín & Cuesta-Zorita, 1994) by total PANSS score. The PANSS consists of 30 items and severity is evaluated on a 7-point scale ranging from 1 (*absent*) to 7 (*extreme*). A score below 59 is considered to be mild symptoms, while a score of 116 or more is severe. A recent study by Barrios et al. (2019) indicated that the

PANSS was the most widely used instrument in empirical research on functioning in schizophrenia. As far as its psychometric properties are concerned, it can be said that they are adequate and that there are various studies both of the adaptation of the instrument to other languages and of the analysis of its items, its validity and reliability. In general, the analysis of items based on the item response theory concluded that the items are very good or good at measuring the overall severity of the disease (Santor et al., 2007). The internal consistency of the scale has been reported to be between 0.62 and 0.92 (Peralta-Martín & Cuesta-Zorita, 1994) and between 0.80 and 0.88 (Ivanova et al., 2018; Linden et al., 2007), and concurrent validity with CGI was in the moderate to high range (0.41-0.61 and 0.61-0.80; Ivanova et al., 2018), while criterion validity with SANS (Scale for the Assessment of Negative Symptoms) was found to be between 0.70 and 0.81 (Peralta-Martín & Cuesta-Zorita, 1994). Henceforth this variable is labelled as 'symptomatology'.

All participants gave their written consent to participate and the study was approved by the Ethics Committee of the University of Barcelona.

Statistical analysis

The DIF analysis was applied to variables such as gender, age, marital status, education, depression, and symptomatology for each WHODAS 2.0 domain through two R software packages (R Development Core Team, 2010): the Psychotree (Zeileis et al., 2018), and the DIFtree (Berger, 2019). The six-domain model was used for this analysis because, as was previously mentioned in the Guilera et al. (2012) study, it was the one which showed the best fit.

The Psychotree package applies TREE-PCM, which detects the values of the covariates in which the biggest difference in the difficulty of the items is found and divides the sample by this value (Strobl, et al., 2010). It is important to bear in mind that if the result is a single node, one joint Rasch model is fit and it will be able to describe the entire data set appropriately. If, in contrast, the Rasch tree presents at least one split, DIF is present and the Rasch model is unfit. Furthermore, due to the number of covariates included in the analysis, which means multiple comparisons and their effect on the increase of the false-positive rate, this procedure applies Bonferroni's adjustment for p value, which in this case was 0.008 (0.05/6 covariates).

The DIFtree package identifies the items responsible for the split of covariates and therefore provides information about items with DIF. According to Tutz and Berger (2016), the procedure detects the covariate that presents the greatest differences between the parameters of the item analysed and uses a permutation test to obtain a decision on DIF controlling for a given significance level. In this study 1,200 permutations were used according to Berger's recommendation and a correction was applied to the level of significance that aimed to control the likelihood of a false detection of DIF or of one variable as responsible for DIF. This correction starts with a p value of 0.008 (0.05/6 covariates studied), but it is amended each time the analysis

of a covariate ends. In a second step, when the analysis of a covariate has been exhausted, the p value will be 0.01 (0.05/5 covariates studied) and so on.

In order to obtain a database to which all of the people assessed could contribute or provide discriminative measurement information, it was necessary to exclude those people that answered all items with the same response category in each WHODAS 2.0 domain. This exclusion criterion is justified insofar as it allows obtaining a set of data with a certain level of variability in the responses of the participants that makes it possible to estimate the measures of the items under the Rasch model and particularly under the Partial Credit Model (Bollmann et al, 2018; Komboz et al, 2018; Linacre, 2021; Zeileis et al, 2018). On the contrary, keeping in the analysis people who have answered all the questions of the same domain with the same category, causes problems of fitting the data to the model. A summary of the algorithm applied by each method for the detection of DIF can be found in the online Supplementary Material 1.

Results

The percentage of participants who were omitted from the DIF analysis because they answered all the questions of a domain (D1-D6) with the same category were in their order: 20.9%, 64.5%, 69.3%, 16.1%, 34.6%, 13.7%.

Of the 6 domains in WHODAS, the only one in which DIF was found was D4 (Getting Along – Interacting with Other People). As shown in Figure 1, the PCM-IFT detected DIF for the age variable solely for item 5—Sexual activities, showing a split at an age of 41 years, that is, for patients with schizophrenia who respond to item 5 there is a different probability of endorsement between those who are 41 years or younger and those over 41. Furthermore, at the end of Figure 1 (in the resulting nodes), the parameter threshold value for each group can be seen (Bollmann et al., 2018). When analysing the graphical results of PCM-IFT it is important to take into account two aspects. First, in the age group ≤ 41 years, you can see that the value of the second threshold is slightly lower than the first one, which would indicate that a lower latent trait is required to pass the second threshold than for the first. Second, when comparing the two end nodes, you can see that the four threshold parameters are higher for the 41 years or younger group versus the over 41 years group. Therefore, it could be stated that patients with schizophrenia aged older than 41 years reported greater problems in terms of functionality with respect to sexual activities and, on the other hand, that in order to report problems in sexual activities they need a lower level of general disability than those aged 41 years or less.

The TREE-PCM procedure did not detect DIF under any covariate in the D4 (Getting Along–Interacting with Other People) domain and, therefore, the results only show one resulting node. Figure 2 presents the plot region for each of the items in the domain (on the X axis) with the estimated parameter threshold values (on the Y axis) for the dataset analysed. In general, it can be

seen that item 3—Getting along with people who are close to you—was reported by patients with schizophrenia as the activity least problematic to carry out, while item 1—Dealing with people you do not know—and item 5—Sexual activities—were the ones in which participants reported as the activities most problematic to carry out. Additionally, in item 5, the dashed lines indicated reversed threshold 1 and 2, which could be indicating that these categories are not modal within some range of the ability assessed by this item and therefore there is no ability value that is more likely to be answered with the second category (mild). Ordered thresholds and ordered category parameters, in WHODAS context, indicate that the higher the response category chosen, the greater the level of disability. In other words, those who select the first response category are expected to have a lower disability level, whereas those who select a higher response category will have a higher disability level. In other words reverse thresholds can occur when there is a very low frequency of response in a category or when a low category is answered by people with a higher level of ability[1], but this does not affect the order of the measurement scale (Adams et al., 2012; Linacre, 1999b). Taking into account the above, what this result in item 5 could be indicating is that category 2 has not been widely chosen by patients with schizophrenia, or that people with less disability than expected reported having mild to moderate levels of difficulty in sexual activities than the people who reported that they had no or mild difficulties in this respect.

Discussion

In general, the results from this study provide evidence supporting the idea of invariance in WHODAS and its validity for assessing patients with schizophrenia. This study demonstrated the presence of DIF in just one item of the 36 included in WHODAS 2.0 for people with schizophrenia. DIF was detected for age but only for the item which examines difficulty in regard to sexual activities, suggesting that the differences indicated in the WHODAS for groups of people according to age should be analysed cautiously in reference to this item but not for the whole scale.

With regard to the results by items, in a previous study, Kirchberger et al. (2014) did not detect DIF in any item, probably because, unlike this study, all the variables used by these investigators were sociodemographic, the age variable was divided into two groups with a cutoff point of 65 years, because it was performed with a population with a medical condition and, finally, because an abridged 12-item version of the WHODAS was applied that did not include the item detected with DIF in this study. This seems to indicate that if the item with DIF is eliminated or if a version of the WHODAS that does not contain this item is applied, measuring disability in people with schizophrenia may be improved. In this regard, it is important to note that some authors have highlighted the relevance of sexual activities (item 5, Getting Along domain) when illness severity in schizophrenia is assessed by healthcare professionals (Sedano-Capdevilla et al., 2018), and deem sexual functions, dressing, and informal relationships, to be relevant aspects in schizophrenia

(Nuño et al., 2018). However, from the factors mentioned above, which are important in the evaluation of patients with schizophrenia, item 5: Sexual activities was found to have DIF for the age variable, therefore it would be advisable to review its content.

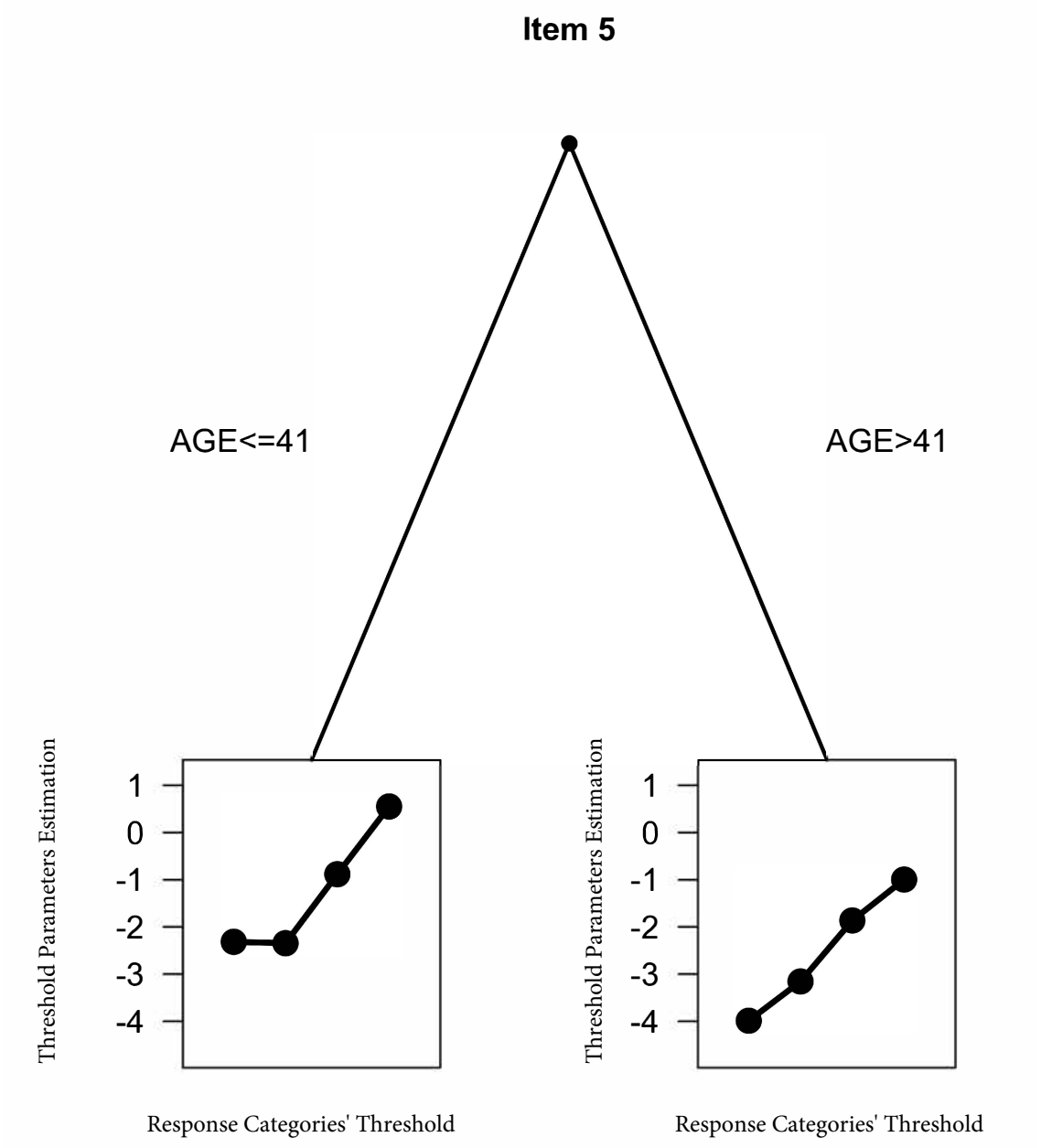


Figure 1. Detection of DIF using the Rasch Trees PCM-IFT procedure in item 5 of the Getting along-interacting with other people domain.
Note: The two branches of the tree indicate that DIF exist for the age variable. The data are subdivided into two splits: those aged 41 years and under, and those over 41. At the end of these branches are two boxes containing line graphs. These boxes indicate the threshold estimation on Partial Credit Model between one response category and the other.

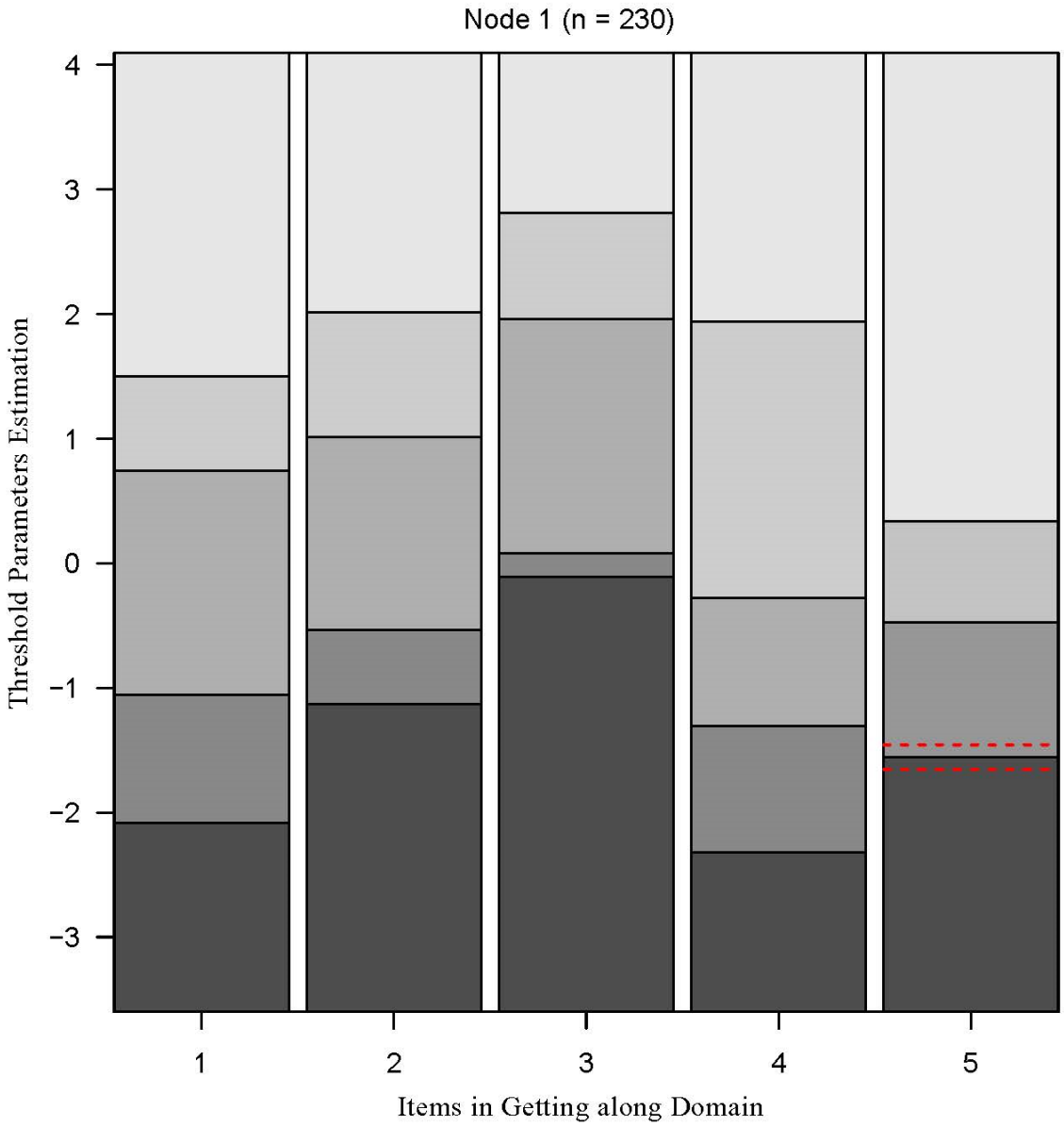


Figure 2. Thresholds estimation between response categories by Partial Credit Model for the Getting along- interacting with other people domain using the TREE-PCM procedure.

The 41-year cutoff point for dividing the groups to be compared is another important aspect in terms of the results obtained. For all the continuous covariates included, no criterion or cutoff point was predetermined before the DIF analysis. If it was necessary to establish a prior criterion, the normal procedure is to use the median as a cutoff point. In our data the median was an age of 37 years and applying this criterion would probably have omitted the detection of DIF or hidden the differences between the thresholds obtained, as the results obtained by Strobl et al. (2015) and Komboz et al. (2018) suggested.

With regard to the differences in the detection of DIF between methods, first of all, TREE-PCM did not detect DIF in any WHODAS 2.0 domain, whereas PCM-IFT detected DIF for age in

item 5 (sexual activities) of the Getting Along domain. These differences may be due to the fact that TREE-PCM detects global DIF at domain level, whereas PCM-IFT is specific for one item, and that PCM-IFT is more sensitive when only one or two items have DIF, whereas TREE-PCM is more sensitive when there are multiple items with minor differences in their parameters (Bollmann et al., 2018).

Furthermore, and regarding item 5 of the getting along domain, TREE-PCM and PCM-IFT found reversed threshold 1 and 2. This information could be interesting for analysing the subjacent response patron and its related with DIF presence. However, it should be clarified that the presence of reverse threshold does not necessarily indicate a problem with the measurement or the adjustment of the data to the Rasch model (Linacre, 1999a). In relation to this result, Wetzel and Carstensen (2014), using data from the revised NEO personality inventory and applying a simulated study, they found that response categories could differentiate between participants with different ability levels despite the presence of reversed thresholds.

In conclusion, this study found that the only covariate that induced DIF, was age for a single item and as a result it is considered to provide sufficient evidence supporting the validity of the WHODAS for measuring disability in people with schizophrenia. The DIF methods could be routinely applied to analyse DIF since they provide complementary information: TREE-PCM report the total DIF induced by the covariates in each domain, that is, the total DIF in the domain, indicating that one or several items of this domain may present DIF; PCM-IFT detect covariates that induced DIF in each item, providing more detailed information at item level.

The results obtained in this study point to more evidence in favour of the validity of the WHODAS 2.0 when people with schizophrenia are evaluated and seem indicate that there is not sufficient evidence to consider that differences found by other authors regarding the covariates analysed here were artificial. However, certain caution should be exercised with the results obtained in the D2 domain (Mobility–Moving and Getting Around) and in the D3 domain (Self-Care–Attending to one's Hygiene, Dressing, Eating and Staying alone), since in them a large percentage of the participants answered all their items with the same response category (mostly it was the lowest category), which does not allow the adequate estimation of the item's parameter and therefore they should have been omitted from the analysis.

However, this study presents some limitations that may affect the level of generalisation of the results. First, the instrument used was the Spanish version of the WHODAS. Therefore, until cross-cultural studies are conducted, these results cannot be extended for the instrument in other languages and cultures. Second, the results of this study should not be considered a comparison of the effectiveness of the DIF detection methods since this is not a controlled experimental study. Instead, this study has been designed to illustrate the application of the Rasch Trees method and analyse the differences and additional information that may be provided when DIF is evaluated in

the WHODAS 2.0. Third, the lack of variability in the participants' responses affected the application of Rasch Trees procedures in domains D2 and D3, for which it was necessary to delete from the database those participants who responded with the same category to all the items in a specific domain, thereby losing potentially significant information about the type of response from schizophrenic patients in the WHODAS 2.0. Fourth, in terms of the clinical variables evaluated, it may have been better to evaluate the depressive symptomatology using an instrument such as the Calgary Depression Scale for Schizophrenia (CDSS- Addington et al., 1990), since it has been found to be one of the instruments that best differentiates between depressive symptoms and schizophrenic symptomatology (Lako et al., 2012).

Future research could study the behaviour of different criteria in order to establish the cutoff point in classical methods, compare the DIF methods to these cutoff points, use a sample of people with schizophrenia with greater variability in the PANSS and HAM-D scores, use other indicators for symptomatology and depression, consider the DIF analysis separately based on positive and negative symptomatology, and analyse DIF through other variables of interest such as occupational status, smoking and others. Regarding the reverse threshold in different items of the scale and the absence of variability in the responses to the items of a domain, we consider that it would be relevant to carry out a study on the distribution of patients with schizophrenia in the response categories and the order in the parameter estimated for each of them, as well as the possibility of collapsing possible problematic categories.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Spain's Ministry of Economy and Competitiveness [Grant PSI2015-67984-R], and by Agency for the Management of University and Research Grants of the Government of Catalonia [Grant 2017SGR1681].

ORCID iDs

Ángela I. Berrío: https://orcid.org/0000-0003-2064-4594

Juana Gómez-Benito: https://orcid.org/0000-0002-4280-3106

Georgina Guilera: https://orcid.org/0000-0002-4941-2511

Note

1. At the request of one of the reviewers, and although it is not the focus of the article, it should be noted that there are other items that presented reverse threshold by TREE-PCM. In Mobility–Moving and Getting Around domain item 3, in Self-Care–Attending to one's Hygiene, Dressing, Eating, and Staying Alone domain items 1 to 4, in Life Activities-Domestic Responsibilities, Leisure, Work, and School domain items 6 and 7, and in Participation–Joining in Community Activities, Participating in Society domain items 2 and 5.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*(4), 547-573. https://doi.org/10.1177/0013164411432166

Adegbaju, D. A., Olagunju, A. T., & Uwakwe, R. (2013). A comparative analysis of disability in individuals with bipolar affective disorder and schizophrenia in a sub-Saharan African mental health hospital: Towards evidence-guided rehabilitation intervention. *Social Psychiatry and Psychiatric Epidemiology, 48*(9), 1405-1415. https://doi.org/10.1007/s00127-013-0654-6

Addington, D., Addington, J., & Schissel, B. (1990). A depression rating scale for schizophrenics. *Schizophrenia Research, 3*(4), 247-251. https://doi.org/10.1016/0920-9964(90)90005-r

Aguocha, C., Uwakwe, R., Olose, E., Amadi, K., Onyeama, G., & Duru, C. (2018). Clinical implication of smoking among patients with schizophrenia at a tertiary institution in south east Nigeria. *African Health Science, 18*(1), 102-110. https://doi.org/10.4314/ahs.v18i1.14

Akinsulore, A., Mapayi, B. M., Aloba, O. O., Oloniniyi, L., Fatoye, F. O., & Makanjoula, R. O. A. (2015). Disability assessment as an outcome measure: A comparative study of Nigerian outpatients with schizophrenia and healthy control. *Annals of General Psychiatry, 14*, Article 40. https://doi.org/10.1186/s12991-015-0079-6

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders—Four Edition- Tet Revision (DSM-IV-TR)*. American Psychiatric Association.

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton Depression Rating Scale: Has the gold standard become a lead weight? *American Journal of Psychiatry, 161*(12), 2163-2177. https://doi.org/10.1176/appi.ajp.161.12.2163

Barrios, M., Guilera, G., Hidalgo, M. D., Cheung, E. C. F., Chan, R. C. K., & Gómez-Benito, J. (2019). The most commonly used instruments in research on functioning in schizophrenia. What are they measuring? *European Psychologist*. https://doi.org/10.1027/1016-9040/a000386

Berger, M. (2019). *DIFtree: Item focused trees for the identification of items in differential item functioning* (R package version 3.1.4) [Computer Software]. http://CRAN.R-project.org/package=DIFtree

Bollmann, S., Berger, M., & Tutz, G. (2018). Item-focused trees for the detection of differential item functioning in partial credit models. *Educational and Psychological Measurement, 78*(5), 781-804. https://doi.org/10.1177/0013164417722179

Chen, R., Llou, T.-H., Miao, N.-F., Chang, K.-H., Yen, C.-F., Liao, H.-F., Chi, W. C., & Chou, K.-R.et al. (2019). Using World Health Organization disability assessment Schedule 2.0 in people with schizophrenia: A 4-year follow-up. *European Archives of psychiatry and Clinical Neuroscience*, *270*(3), 301-310. https://doi.org/10.1007/s00406-019-01000-5

Chen, Y. F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of PISA 2009 reading assessment. *Educational Assessment*, *19*(2), 77-96. https://doi.org/10.1080/10627197.2014.903650

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*(2), 133-148. https://doi.org/10.1111/j.1745-3984.2005.00007

Dan, A., Kumar, S., Avasthi, A., & Grover, S. (2011). A comparative study on quality of life of patients of schizophrenia with and without depression. *Psychiatry Research*, *189*(2), 185-189. https://doi.org/10.1016/j.psychres.2011.02.017

Ertugrul, A., & Ulug, B. (2004). Perception of stigma among patients with schizophrenia. *Social Psychiatry and Psychiatric Epidemiology*, *39*(1), 73-77. https://doi.org/10.1007/s00127-004-0697-9

Galindo-Garre, F., Hidalgo, M. D., Guilera, G., Pino, O., Rojo, J. E., & Gómez-Benito, J. (2015). Modeling the World Health Organization disability assessment schedule II using non-parametric item response models. *International Journal of Methods in Psychiatric Research*, *24*(1), 1-10. https://doi.org/10.1002/mpr.1462

Grover, S., Sahoo, S., Dua, D., Chakrabarti, S., & Avasthi, A. (2017). Scales for assessment of depression in schizophrenia: Factor analysis of Calgary Depression Rating Scale and Hamilton Depression Rating Scale. *Psychiatry Research*, *252*(June), 333-339. https://doi.org/10.1016 /j.psychres.2017.03.018

Guilera, G., Gómez-Benito, J., Pino, O., Rojo, J. E., Cuesta, M. J., Martínez-Arán, A., Safont, G., Tabarés-Seisdedos, R., Vieta, E., Bernardo, M., Crespo-Facorro, B., Franco, M., & Rejas, J. (2012). Utility of the World Health Organization disability assessment schedule II in schizophrenia. *Schizophrenia Research*, *138*(2-3), 240-247. https://doi.org/10.1016/j.schres.2012.03.031

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology Neurosurgery & Psychiatry*, *23*(1), 56-62. https://doi.org/10.1136/jnnp.23.1.56

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds). *Test validity* (pp.129-145). Lawrence Erlbaum Associates.

Ivanova, E., Khan, A., Liharska, L., Reznik, A., Kuzmin, S., Kushnir, O., Agarkov, A., Bokhan, N., Pogorelova, T., Khomenko, O., Chernysheva, K., Morozova, M., Rupchev, G., Lepilkina, T., Ozornina, N., Govorin, N., Malakhova, A., Hmara, N., … Opler, L. A. (2018). Validation of the Russian version of the Positive and Negative Syndrome Scale (PANSS-Ru) and normative data. *Innovations in Clinical Neuroscience*, *15*(9-10), 32-48.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, *13*(2), 261-276. https://doi.org/10.1093/schbul/13.2.261

Kimber, M., Rehm, J., & Ferro, M. A. (2015). Measurement invariance of the WHODAS 2.0 in a population-based sample of youth. *PLOS ONE*, *10*(11), e0142385. https://doi.org/10.1371 /journal.pone.0142385

Kirchberger, I., Braitmayer, K., Coenen, M., Oberhauser, C., & Meisinger, C. (2014). Feasibility and psychometric properties of the German 12-item WHO disability assessment schedule (WHODAS 2.0) in a population-based sample of patients with myocardial infraction from the MONICA/KORA myocardial infraction registry. *Population Health Metrics*, *12*(1), Article 27. https://doi.org/10.1186/s12963-014-0027-8

Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch model. *Educational and Psychological Measurement*, *78*(1), 128-166. https://doi.org/10.1177 /0013164416664394

Lako, I. M., Bruggeman, R., Knegtering, H., Wiersma, D., Shoevers, R. A., Slooff, C. J., & Taxis, K. (2012). A systematic review of instruments to measure depressive symptoms in patients with schizophrenia. *Journal of Affective Disorders*, *140*(2012), 38-47. https://doi.org/10.1016 /j.jad.2011.10.014

Linacre, J. M. (1999a). Category disordering (disordered categories) vs. threshold disordering (disordered thresholds). *Rasch Measurement Transactions*, *13*(1), 675. https://www.rasch.org/rmt /rmt131a.htm

Linacre, J. M. (1999b). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103-122.

Linacre, J. M. (2021). *Winsteps ® Rasch-Model Computer Programs: User's guide*. Winsteps.com.

Linacre, J. M., & Wright, B. D. (1987). *Item bias: Mantel Haenszel and the Rasch model*. Psychometric Laboratory, Department of Education, University of Chicago. https://www.rasch.org/memo39.pdf

Linden, M., Scheel, T., & Rettig, K. (2007). Validation of the factorial structure of the Positive and Negative Syndrome Scale in use by untrained psychiatrists in routine care. *International Journal of Psychiatry in Clinical Practice*, *11*(1), 53-60. https://doi.org/10.1080/13651500600884419

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

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

McKibbin, C., Patterson, T., & Jeste, D. V. (2004). Assessing disability in older patients with schizophrenia: Results from the WHODAS-II. *Journal of Nervous and Mental Disease*, *192*(6), 405-413. https://doi.org/10.1097/01.nmd.0000130133.32276.83

Müller, M. J., Müller, K. M., & Fellgiebel, A. (2006). Detection of depression in acute schizophrenia: Sensitivity and specificity of 2 standard observer rating scales. *Canadian Journal of Psychiatry*, *51*(6), 387-392. https://doi.org/10.1177/070674370605100609

Nuño, L., Barrios, M., Rojo, E., Gómez-Benito, J., & Guilera, G. (2018). Validation of the ICF core sets for schizophrenia from the perspective of psychiatrists: An international Delphi study. *Journal of Psychiatric Research*, *103*, 134-141. https://doi.org/10.1016/j.jpsychires.2018.05.012

Olagunju, A. T., Adegbaju, D. A., & Uwakwe, R. (2016). Disability among attendees with schizophrenia in a Nigerian hospital: Further evidence for integrated rehabilitative treatment designs. *Mental Illness*, *8*(2), 6647. https://doi.org/10.4081/mi.2016.6647

Park, K., Lee, D.-K., Lee. H., Kim, C.-E., & Ryu, S. (2019). Functional disabilities evaluated using World Health Organization disability assessment schedule 2.0 in patients with chronic schizophrenia and its related factors. *Journal of Korean Neuropsychiatry Association*, *58*(1), 47-54. https://doi.org /10.4306/jknpa.2019.58.1.47

Peralta-Martín, V., & Cuesta-Zorita, M. J. (1994). Validación de la escala de los síndromes positivos y negativos (PANSS) en una muestra de esquizofrénicos [The validation on the Positive and Negative Syndrome Scale (PANSS) in a sample of schizophrenics]. *Actas Luso Españolas de Neurología Psiquiatría y Ciencias Afines*, *22*(4), 171–177.

R Development Core Team (2010). *R: A language and environment for statistical computing*. http://www.R-project.org/

Ramos-Brieva, J., & Cordero Villafáfila, A. (1986). Validación de la versión castellana de la escala Hamilton para la depresión [The validation of the Spanish version of the Hamilton Rating

Scale for Depression]. *Actas Luso Españolas de Neurología Psiquiatría y Ciencias Afines*, *14*(4), 324-334.

Santor, D. A., Ascher-Svanum, H., Lidenmayer, J.-P., & Obenchain, R. L. (2007). Item response analysis of the Positive and Negative Syndrome Scale. *BMC Psychiatry*, *7*, Article 66. https://doi.org/ 10.1186/1471-244X-7-66

Sedano-Capdevila, A., Barrigón, M. L., Delgado-Gomez, D., Barahona, I., Aroca, F., Peñuelas-Calvo, I., Miguelez-Fernandez, M., Rodríguez-Jover, A., Amodeo-Escribano, S., González-Granado, M., & Baca-García, E. (2018). WHODAS 2.0 as a measure of severity of illness: Results of a FLDA Analysis. *Computational and Mathematical Methods in Medicine*, *2018*, Article 7353624. https://doi.org/10.1155/2018/7353624

Sjonnense, K., Bulloch, A. G. M., Williams, J., Lavorato, D., & Patten, S. B. (2016). Characterization of disability in Canadians with mental disorders using an abbreviated version of a DSM-5 emerging measure: The 12-item WHO disability assessment Schedule (WHODAS) 2.0. *Canadian Journal of Psychiatry*, *61*(4), 227-235. https://doi.org/10.1177/0706743716632514

Strassnig, M., Kotov, R., Fochtmann, L., Kalin, M., Bromet, E. J., & Harvey, P. D. (2018). Associations of independent living and labor force participation with impairment indicators in schizophrenia and bipolar disorder at 20-year follow-up. *Schizophrenia Research, 197*(July), 150-155. https://doi.org/10.1016/j.schres.2018.02.009

Strobl, C., Kopf, J., & Zeileis, A. (2010). *A new method for detecting differential item functioning in the Rasch model* (Technical Report Number 92), Department of Statistics, University of Munich. https://www.econstor.eu/bistream/10419/73503/1/74479868.pdf

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289-316. https://doi.org/10.1007/S11336-013-9388-3

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb0075.x

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Lawrence Erlbaum Associates.

Tompke, B. K., Tang, J., Oltean, I. I., Buchan, M. C., Reaume, S. V., & Ferro, M. A. (2020). Measurement invariance of the WHODAS 2.0 across youth with and without physical or mental conditions. *Assessment*, *27*(7), 1490-1501. https://doi.org/10.1177/1073191118816435

Tutz, G., & Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika*, *81*(3), 727-750. https://doi.org/10.1007/s11336-015-9488-3

Üstün, T. B., Kostanjsek, N., Chatterji, S., & Rehm, J. (2010). *Measuring Health and Disability for WHO Disability Assessment Schedule: WHODAS 2.0*. WHO Press.

Vázquez-Baquero, J. L., Herrera Castañedo, S., Vázquez Bourgón, E., & Gaite Pindado, L. (2006). *Cuestionario para la evaluación de la discapacidad de la Organización Mundial de la Salud: Versión española del World Health Organization disability assessment schedule II, WHO-DAS II* [The World Health Organization questionnaire in evaluating disabilities: Spanish version of the World Health Organization disability assessment schedule II, WHO-DAS II]. Subdirección General de Información Administrativa y Publicaciones Ministerio de Trabajo y Asuntos Sociales [General Subdirectorate for Administrative Information and Publications, Ministry of Work and Social Affairs]. https://biadmin.cibersam.es/Intranet/Ficheros/GetFichero.aspx?FileName=449 _6f9f03b4 -ffa5-4327-a4e5-4e5ed73c2c27.pdf

Wetzel, E., & Carstensen, C. H. (2014). Reversed threshold in partial credit models: A reason for collapsing categories? *Assessment*, *21*(6), 765-774. https://doi.org/10.1177/1073191114530775

World Health Organization. (2000). *World Health Organization: Disability assessment schedule II (WHO-DAS II)*. WHO Press.

Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., & Kopf, J. (2018). *Psychotree: Recursive partitioning based on psychometric models* (R package version 0.15-2) [Computer Software]. http://CRAN.R-project.org/package=psychotree

Zhou, W., Liu, Q., Yu, Y., Xiao, S., Chen, L., Khoshnood, K., & Zheng, S. (2020). Proxy reliability of the 12-item World Health Organization disability assessment schedule II among adult patients with mental disorders. *Quality of Life Research*, *29*(8), 2219-2229. https://doi.org/10.1007/s11136-020-02474-w

4. Discusión

Esta tesis ha cumplido con su objetivo principal al proporcionar información respecto a las tendencias y nuevos desarrollos en métodos de detección de DIF mediante una síntesis de la literatura, un estudio con datos simulados y otro con datos empíricos. En conjunto, los tres estudios indican las tendencias en el análisis de DIF, señalan algunos vacíos en el estudio de la eficacia de los procedimientos y sugieren las ventajas de promover el uso de técnicas que permiten un análisis de DIF sin especificar previamente la conformación de los grupos de referencia y focal.

El análisis y síntesis de la producción científica en métodos de detección de DIF en estudios con datos simulados, además de describir las variables principales y comunes en los estudios de interés, señalar tendencias y presentar los métodos más frecuentemente estudiados y aquellos más recientes, proporcionó una primera aproximación de la estructura u organización interna de la investigación en este ámbito de estudio. Esto último permitió visualizar los tópicos de mayor interés para los investigadores y en su interior algunos conceptos de reciente desarrollo al analizar su aparición temporal, lo que proporcionó información para comprender el enfoque que va tomando la investigación, las disciplinas involucradas o que influencian su desarrollo y los temas en torno a los cuales ha girado el interés de los investigadores.

En cuanto al análisis de DIF y el tamaño de muestras, algunos investigadores han concluido que un tamaño de grupo mínimo de 200 participantes resulta adecuado para la detección de DIF (Clauser & Mazor, 1998; Zumbo, 1999), esto pudo motivar a que los investigadores tomaran este tamaño como un criterio y por ende, la mayor frecuencia de estudios con datos simulados que definieron tamaños de grupo de 500 y 1,000 obtenida en el primer estudio. Sin embargo,

también se ha reconocido que bajo el modelo Rasch las estimaciones del parámetro de dificultad son adecuadas y estables en muestras de entre 30 y 250 participantes (Linacre, 1994) y más recientemente, Belzak (2020) recomendó para el análisis de DIF la aplicación de métodos más parsimoniosos (regresión logística o basado en el modelo de un parámetro) en lugar de otros más complejos (ejemplo, basados en el modelo de dos parámetros) cuando los tamaños de grupo son tan pequeños como 50 o 100. Esto último sustenta los resultados obtenidos en el primer estudio respecto al método basado en el modelo Rasch, puesto que no es de extrañar que sea uno de los métodos más estudiados y al mismo tiempo, un método que con sus desarrollos más recientes continúa siendo el foco de interés de los investigadores, en especial si se tiene en cuenta la tendencia de la investigación actual orientada hacia tamaños de muestra pequeños. Además, corrobora los resultados obtenidos en el estudio 2 cuando el grupo focal estaba conformado por 130 examinados (razón de tamaños 1,000) y permite tener una cierta confianza de los resultados obtenidos en el estudio 3 con un tamaño de muestra total de 280 participantes o menos.

Cabe destacar que en esta tesis se analizó DIF bajo el modelo Rasch tanto con tamaños de grupos pequeños (estudios 2 y 3), como otros muy grandes propios de lo esperado en las pruebas de aplicación masiva o a gran escala (estudio 2). Los resultados evidenciaron la robustez de los procedimientos basados en el modelo Rasch en ámbitos de aplicación masiva con tamaños de muestra de 130,000 examinados y razones de tamaños entre 20 y 1,000 o en otros con muestras más modestas (en torno a 280 o menos). Avalando lo mencionado en los párrafos anteriores.

Se demostró la robustez del modelo Rasch para detectar DIF frente al efecto de la razón de tamaños, especialmente cuando los datos se ajustan a un modelo 1PL en la condición 0% ítems con DIF. Por tanto, tal como concluyen Paek y Guo (2011), la razón de tamaños puede incrementarse a 30 o 40 sin afectar las tasas de falsos positivos obtenidas con razones de tamaño más pequeñas, o incluso a partir de la

evidencia aportada, incrementar las razones de tamaño hasta 1,000 no presenta un efecto relevante en las tasas de falsos positivos. Este resultado es relevante si se tiene en cuenta que las últimas propuestas para la detección de DIF bajo el modelo Rasch promueven que no se definan los grupos previo al análisis y por tanto, no es posible determinar de antemano la razón de tamaños entre ellos, tal como ocurre en el análisis de DIF del estudio 3.

En relación con lo anterior, un aspecto que permanece bajo discusión es el posible efecto de los tamaños de los grupos por separado, independientemente de las razones de tamaño. A este respecto Paek y Guo (2011) sugirieron que mantener constante el tamaño del grupo focal e incrementar el grupo de referencia (obteniendo razones de tamaño mayores) produce una mayor adecuación en la detección de DIF. En este orden de ideas, las tasas de falsos positivos por encima del nivel nominal cuando el grupo focal esta formado por más de 6,000 examinados, cuando la razón de tamaños es 20, podrían disminuirse si se incrementa el tamaño del grupo de referencia hasta obtener una razón de tamaños por encima de 20.

Otro hallazgo que aportó evidencia de la robustez de los procedimientos basados en modelo Rasch fue el control adecuado de las tasas de falsos positivos cuando los datos no se ajustan al modelo o cuando se simulan bajo un modelo 3PL. Esta evidencia sorprende si se tiene en cuenta que cuando el parámetro de discriminación constante no se mantiene, las estimaciones del parámetro de dificultad en los grupos se afectan dando lugar a la detección de DIF (DeMars & Jurich, 2015) y que cuando los datos no tienen una asíntota cero (como ocurre cuando c = 0.15), las estimaciones del parámetro de discriminación varían, especialmente cuando hay diferencias de medias entre los grupos (DeMars, 2010), lo que incrementa la posibilidad de detectar DIF erróneamente.

Con relación a esto último es preciso mencionar que en el ámbito empírico una manera, pero no la única, de asegurar un cierto ajuste de los datos al modelo

Rasch es revisar previamente los patrones de respuesta y eliminar aquellas en las que un examinado responda a todas las preguntas con la misma categoría de respuesta, esto aunque asegura un mejor ajuste de los datos al modelo porque se extrae información poco variable que no aporta a la medición, no asegura un ajuste al 100% y por ello cobran relevancia los hallazgos del estudio 2 respecto a la robustez del modelo Rasch frente a datos que no se ajustan.

Respecto a la potencia del modelo Rasch para detectar DIF, se obtuvieron tasas de detección correcta en niveles aceptables a moderados con razones de tamaño 250 o menos que no se ven afectadas por el modelo de simulación de los datos o las diferencias en la distribución del nivel de atributo. Además, se alcanzaron tasas altas de detección de DIF nouniforme con datos simulados bajo un modelo 3PL, similares a las reportadas por Prieto-Marañón et al. (2012) con datos simulados bajo el modelo 3PL c = 0.20 usando reglas de decisión basadas en MH y test Breslow-Day. Otro resultado que sorprendió y que constituye una evidencia a favor de la robustez fue que la detección de DIF uniforme presenta tasas aceptables de detección correcta incluso cuando los datos se ajustan a un modelo 3PL. Por lo anterior, en el estudio con datos empíricos que identificó un ítem con DIF uniforme, es esperable que se haya detectado adecuadamente el ítem incluso si había impacto entre los grupos o si los datos no estaban ajustados al modelo Rasch.

En general, las variables que afectan las tasas de detección correcta incluyen tamaños de muestra grandes, bajo porcentaje de ítems con DIF y DIF balanceado (Chen et al., 2014; Fidalgo et al., 2004). A este respecto, la evidencia aportada sugiere que la razón de tamaños también afecta las tasas de detección correcta y que bajo algunas condiciones esta variable puede presentar una interacción con las diferencias en la distribución del nivel de atributo dependiendo del criterio empleado para la detección de DIF (la prueba de significación o la combinación de la prueba de significación con una medida de magnitud de DIF). Además, el efecto del porcentaje de ítems con DIF no presenta una clara relación con las tasas de

detección correcta. A este respecto, los resultados del estudio 3 parecen estar en la dirección de lo demostrado por Bollmann et al. (2018), cuyo estudio concluyó que PCM-IFT es más sensible cuando sólo uno o dos ítems presentan DIF, mientras que Tree-PCM es más sensible cuando hay múltiples ítems con diferencias en el parámetro. Por tanto PCM-IFT resulta menos sensible al efecto del % de ítems con DIF al contrario de Tree-PCM.

Otro aspecto que pudo haber afectado los resultados tanto de las tasas de falsos positivos como de las tasas de detección correcta, especialmente cuando el criterio de detección de DIF es la prueba de significación, son los valores de los parámetros de dificultad y discriminación. Al respecto Guilera et al. (2013) concluyeron que dichas variables tiene un efecto moderador sobre la potencia y el error tipo I, mientras que Li et al. (2012) encontraron que valores del parámetro de discriminación fuera del intervalo 0.6 y 0.9 producen una inflación del error tipo I. Teniendo en cuenta los resultados bajo un modelo 3PL, el valor del parámetro de discriminación pudo incrementar su efecto sobre las tasas de falsos positivos puesto que la mayoría de ítems tenían valores de discriminación por fuera de dicho intervalo. Sin embargo, como esta variable no fue el centro del análisis, se requiere un estudio más a profundidad a futuro.

Por otro lado, una aportación relevante de la tesis constituye la ventaja demostrada que tienen los procedimientos basados en árboles de Rasch en el análisis de DIF con variables continuas puesto que la detección de DIF se hizo por la variable edad con un punto de corte en 41 años. Si se hubiese aplicado el criterio a priori más empleado por los investigadores (la mediana) el punto de corte para esta variable hubiera sido 37 años. Adoptar este tipo de criterios previo al análisis de datos puede resultar en la no detección de DIF para este ítem o no descubrir diferencias entre los umbrales de una categoría de respuesta a otra, tal como sugirieron los resultados de los estudios de Strobl et al. (2015) y Komboz et al. (2018).

Respecto a los umbrales entre categorías de respuesta, los dos procedimientos basados en árboles de Rasch señalaron la presencia de umbrales inversos entre el 1 y el 2 para el ítem identificado con DIF por PCM-IFT. Aunque este hallazgo no fue exclusivo para dicho ítem, se considera oportuno comentar que tener umbrales inversos no necesariamente indica un problema de medición o de desajuste al modelo Rasch (Linacre, 1999), aunque sí puede señalar la necesidad de hacer un análisis más detallado respecto al patrón de respuestas de las personas con esquizofrenia o indicar la necesidad de revisar cómo las personas con este diagnóstico están interpretando la pregunta o las opciones de respuesta.

En cuanto a la presencia de DIF en la escala WHODAS, la detección se realizó en un ítem que se refiere a las dificultades respecto de la actividad sexual, tema que ha sido señalado por diversos autores como relevante al medir las problemáticas que afrontan las personas con esquizofrenia (Sedano-Capdevilla et al., 2018; Nuño et al., 2018), por lo que la recomendación está más orientada a revisar el ítem en lugar de eliminarlo de la escala.

Los resultados obtenidos indican que el método de detección de DIF basado en el modelo Rasch ha sido uno de los que mayor interés ha suscitado entre los investigadores. Esta popularidad puede incrementarse si se tiene en cuenta que esta tesis aporta evidencia a favor de su robustez no sólo frente a tamaños de grupos pequeños, sino también frente a condiciones de datos que no ajustan al modelo, razón de tamaños o presencia de impacto. Esto es especialmente interesante si se tiene en cuenta que al aplicar los procedimientos basados en árboles de Rasch no se puede tener información previa sobre el tamaño de los grupos o las diferencias en la distribución del nivel de atributo entre los grupos. Además, los árboles de Rasch han demostrado que son aplicables en test cortos, pueden proporcionar información complementaria que enriquece el análisis de DIF y muestran una gran ventaja cuando se incorporan variables continuas como fuente de DIF.

En definitiva, esta tesis aporta evidencia que promueve la discusión respecto a aspectos metodológicos propios de la investigación en métodos de detección de DIF, no sólo para el diseño de estudios centrados en determinar el funcionamiento de dichos métodos a partir de datos simulados, sino también respecto a condiciones particulares de evaluación que pueden afectar la detección de DIF cuando se aplica a datos empíricos.

5. Conclusiones e implicaciones prácticas

En conjunto, los resultados obtenidos proporcionan evidencia de que los métodos de detección de DIF son un campo de estudio fructífero que, por su evolución durante las últimas 4 décadas, proporcionan una gran variedad de procedimientos que permiten evaluar la presencia o no de DIF. La evolución de estos métodos ha estado ligada a la mejora de su funcionamiento y a las demandas planteadas desde la práctica para poder contar con métodos que abarquen determinadas características de la evaluación. Así, se aporta evidencia tanto experimental como empírica sobre la robustez del método basado en el modelo Rasch en dos ámbitos de estudio diferentes en cuanto al tamaño de muestra, la longitud del test, y tipo de ítems, pero muy similares en cuanto a las consecuencias que puede tener para la persona la estimación inadecuada de su nivel de habilidad. Respecto a los procedimientos de detección de DIF basados en el modelo Rasch que fueron empleados en esta tesis, los estudios demostraron que éstos pueden emplearse tanto con muestras muy grandes de examinados, como con muestras más modestas. Por tanto, resultan hábiles para detectar DIF tanto en pruebas de aplicación masiva como en pruebas aplicadas en el ámbito clínico, donde los subgrupos de comparación pueden ser pequeños. También, se aporta evidencia respecto a la utilidad de emplear procedimientos de detección de DIF que no requieren el establecimiento de un criterio a priori para dividir los grupos de referencia y focal cuando la variable de interés es cuantitativa, en un análisis conjunto de covariables en distinta escala de medida.

Cabe destacar que los hallazgos de esta tesis son de gran utilidad tanto para los estudiantes interesados en la detección de DIF, como para los usuarios de dichos métodos y para los investigadores interesados en el funcionamiento de éstos. El primer estudio integró y consolidó la información respecto al diseño de estudios con datos simulados sobre métodos de detección de DIF, lo que claramente puede

tomarse como una guía para los investigadores o estudiantes interesados en hacer estudios con datos simulados. Por ejemplo, la discusión sobre el número de réplicas y su evolución, los tamaños de muestra, razón de tamaños, presencia de impacto, y longitud del test, entre otras, ha señalado los vacíos y tendencias recientes en el diseño de este tipo de estudios y puede servir para orientar futuras investigaciones. Las recomendaciones generales que se desprenden de la discusión de los resultados pueden resumirse en los siguientes puntos:

1. Para la definición del número de réplicas se recomienda el uso de la fórmula propuesta por Feinberg y Rubright (2016), en lugar de centrarse en aspectos como el tiempo de cómputo o el valor más frecuente que se encuentre en la literatura.
2. Al nombrar el procedimiento de detección de DIF que se estudiará, es recomendable siempre hacer mención del método matriz del que procede. Por ejemplo, en el primer estudio se propuso una clasificación en 22 métodos generales que contenían diferentes procedimientos o estadísticos para detectar o refinar la detección de DIF. Esta clasificación podría emplearse con el objetivo de unificar la denominación de los procedimientos y facilitar así la identificación de los métodos que se pretenden valorar.
3. Respecto a las condiciones experimentales, las variables y valores más comunes han sido: tamaños de muestras de 1,000 examinados, razón de tamaños de muestras en las que el tamaño del grupo de referencia corresponde a 2 veces el tamaño del grupo focal, ítems dicotómicos, y la ausencia de impacto. Por lo que evaluar el efecto de estas variables con valores por encima o debajo de éstos podría constituir un aporte a las tendencias actuales, así como el análisis de DIF con datos multinivel.
4. La aplicación de procedimientos que no requieren la definición a priori de los grupos de referencia y focal, y los procedimientos basados en el modelo de diagnóstico cognitivo, requieren mayor investigación.

Otro de los aportes relevantes se refiere a los hallazgos del procedimiento diferencia del parámetro de dificultad para ser aplicado, en el ámbito educativo, en pruebas nacionales de altas consecuencias. Para el usuario o investigador interesado en aplicar este procedimiento en condiciones similares a las simuladas, se deben tener en cuenta los siguientes aspectos:

1. El uso del criterio combinado del test de significación y la magnitud de DIF produce un mayor control del efecto de variables como las diferencias en la distribución del nivel de atributo y razón de tamaños, que cuando se usa el criterio del test de significación únicamente. Además, el uso del ajuste de Bonferroni puede ser especialmente relevante cuando el test contiene un 10% o 20% de ítems con DIF.
2. De preferencia, este procedimiento funciona adecuadamente cuando los datos se ajustan a un modelo de un parámetro y en general puede emplearse cualquier criterio de detección de DIF. Sin embargo, para asegurar un uso adecuado se recomienda que el usuario tenga en cuenta las siguientes indicaciones:
    a) El procedimiento puede presentar tasas de falsos positivos (FP < 0.13) y de detección correcta (DC ≥ 0.60) en niveles aceptables cuando se emplea la prueba de significación como criterio de detección de DIF con 10% de ítems con DIF si la razón de tamaños es 500 y los grupos no difieren en su distribución del nivel de atributo, o si la razón es 250 o menos en cualquier condición de diferencia en la distribución del nivel de atributo.
    b) Se podrán obtener tasas de falsos positivos adecuadas (FP < 0.02) y tasas de detección correcta aceptables (DC ≥ 0.64) usando como criterio de detección de DIF la prueba de significación cuando existe un 20% de ítems con DIF, la razón de tamaños es 100 y los grupos no difieren en la distribución del nivel de atributo o difieren únicamente en la varianza.
    c) Al emplear como criterio de detección de DIF la prueba de significación combinada con la magnitud de DIF se presentarán tasas de falsos positivos en niveles adecuados y tasas de detección correcta en niveles

aceptables si hay un 10% de ítems con DIF y la razón de tamaños es 250 o menos en cualquier condición de diferencia en la distribución del nivel de atributo, y si hay un 20% de ítems con DIF cuando la razón es 20 en cualquier condición de diferencia en la distribución del nivel de atributo o cuando la razón es 100 y la distribución del nivel de atributo no difiere entre los grupos o difiere únicamente en la varianza.

3. Cuando el usuario tenga razones para pensar en la presencia de un alto porcentaje de respuestas al azar, y por tanto que los datos se ajustan a un modelo 3PL, no se recomienda el uso del criterio combinado bajo ninguna de las condiciones simuladas. Por tanto, se deberá aplicar el procedimiento usando sólo el criterio del test de significación cuando hay 10% ítems con DIF, no hay diferencias en la distribución del nivel de atributo o las diferencias están en la varianza de la distribución y la razón de tamaños es 100. En este caso las tasas de falsos positivos serán aceptables y las tasas de detección correcta serán muy adecuadas. En algún caso también puede considerarse una razón de tamaños 250 en las mismas condiciones de diferencias en la distribución del nivel de atributo, pero se debe tener en cuenta que las tasas de detección correcta disminuyen a un intervalo entre 0.52 y 0.60, mientras que las tasas de falsos positivos serán adecuadas.
4. Si el usuario desconoce el modelo logístico al cual se ajustan los datos, se recomienda emplear como criterio de detección de DIF la prueba de significación. Pero ha de tener en cuenta que si el test contiene 10% ítems con DIF este procedimiento es susceptible de ser aplicado cuando los grupos tienen la misma distribución del nivel de atributo o al menos las diferencias en dicha distribución no se encuentra en la media, y los grupos se configuran de tal manera que hay 100 examinados en el grupo de referencia por cada uno del grupo focal. En este caso se esperarán tasas de falsos positivos y de detección correcta en niveles aceptables (por debajo de 0.10 y sobre 0.67, respectivamente).

Por otro lado, la aplicación de los procedimientos basados en árboles de Rasch para la detección de DIF proporciona información novedosa respecto al análisis de DIF en la escala WHODAS 2.0 aplicada a personas con esquizofrenia, tomando como variables de interés algunas de corte clínico y otras demográficas. Para el usuario de la escala WHODAS es importante recalcar que:

1. La única variable, de todas las estudiadas, que indujo DIF fue la edad para un único ítem. Por ello se ha considerado que la evidencia proporcionada concluye a favor de la validez de la escala WHODAS al valorar la discapacidad en personas con esquizofrenia.
2. La evidencia proporcionada parece indicar que las diferencias encontradas por otros autores en variables como depresión, sintomatología y estado civil entre otros, no se pueden considerar como artificiales.
3. No se recomienda eliminar el ítem con DIF puesto que éste valora las dificultades con las actividades sexuales y este tema ha sido reconocido como un aspecto de interés entre las problemáticas que afrontan las personas con esquizofrenia. Por tanto se deberá mantener teniendo en cuenta que para la variable edad se ha identificado DIF.
4. Además, se debe tener cierta precaución con los resultados obtenidos en los dominios sobre movilidad y autocuidado debido a que en ellos un gran porcentaje de participantes respondieron a todos los ítems del dominio con la misma categoría de respuesta. Dicha categoría, en la mayoría de los casos se refería a la no existencia de problemas. Esto podría estar indicando algunas deficiencias de la escala al evaluar funcionalidad en personas con esquizofrenia, o un estilo de respuesta específico de esta población ante determinadas temáticas.

Finalmente, para los interesados en realizar análisis de DIF mediante los procedimientos de árboles de Rasch se debe tener en cuenta que los resultados obtenidos parecen indicar que:

1. Los procedimientos Tree-PCM y PCM-IFT se pueden aplicar en pruebas cortas, con ítems politómicos y analizando variables de diversa índole sin

necesidad de establecer puntos de corte a priori para las variables contínuas. Ya que el análisis de DIF se realizó para cada dominio del WHODAS, se puede concluir que estos procedimientos pueden aplicarse en tests cortos formados por cuatro, cinco u ocho ítems.

2. Los procedimientos Tree-PCM y PCM-IFT proporcionan información complementaria y permiten además de la detección de DIF, valorar información referente a los umbrales entre las categorías de respuesta, lo que podría orientar de alguna manera los estudios de sesgo o sugerir patrones de respuesta. Además, por su énfasis en el test en su conjunto y en los ítems, el uso combinado de estos procedimientos resulta de gran interés para valorar el efecto del DIF sobre la medición.

En definitiva esta tesis aporta información relevante y novedosa, para el investigador y el usuario de métodos de detección de DIF, que permite contar con un panorama general de las condiciones experimentales más estudiadas y aquellas poco tratadas, además permite orientar las condiciones ideales o aceptables de aplicación de los métodos de detección de DIF basados en el modelo Rasch en ámbitos propios de la evaluación educativa y promueve la aplicación de procedimientos de desarrollo reciente que no requieren la definición de grupos a priori específicamente en el caso de la evaluación en el ámbito de la salud, aunque no con ello se ha querido sugerir que este tipo de procedimientos sean de uso restringido en este ámbito de evaluación.

6. Fortalezas, limitaciones y futuras direcciones

La principal fortaleza de esta tesis ha sido la aplicación de tres aproximaciones metodológicas para enriquecer la información sobre el campo de estudio en métodos de detección de DIF y sobre algunos de los procedimientos más empleados o de reciente desarrollo. Así se ha aportado información de diversa índole y con distinto énfasis que en conjunto ha permitido proporcionar una visión general de las tendencias en cuanto a los métodos de detección de DIF.

Relacionado con lo anterior, la rigurosidad en el tratamiento o generación de los datos es un aporte relevante que incrementa la fortaleza de la tesis y que puede verse reflejado en los siguientes aspectos: (a) realizar una revisión sistemática siguiendo la guía PRISMA es en sí mismo un criterio de calidad, (b) diseñar un estudio con datos simulados en el cual el criterio para la elección de las variables y sus valores se basa únicamente en la construcción de bases de datos lo más cercanas a la realidad y no el tiempo de procesamiento de datos, así como el establecimiento del número de réplicas a partir de criterios definidos, constituyen un buen argumento para juzgar su calidad y (c) el abordaje empírico con la participación de un número importante de personas con esquizofrenia (N = 280) como también, el análisis de DIF con procedimientos que demuestran una ventaja en cuanto al análisis conjunto de covariables y que no requieren de la definición a priori de los grupos de referencia y focal; aportan en conjunto evidencia de la calidad de los estudios.

Otra de las fortalezas de esta tesis lo constituye el hecho de que para nuestro conocimiento esta ha sido la primera vez que se ha analizado toda la literatura científica producida respecto a los métodos de detección de DIF mediante datos simulados, lo que ha permitido hacer un análisis de todo el campo de estudio y detectar algunos vacíos que se abordaron mediante los estudios experimental y

empírico. Así mismo, ha sido la primera vez que se ha analizado el funcionamiento del procedimiento diferencia del parámetro de dificultad bajo condiciones de tamaño de muestra grande, razón de tamaños entre 20 y 1,000, diferencias en la distribución del nivel de atributo entre grupos y datos simulados bajo un modelo 1PL o 3PL, comparando dos criterios de detección de DIF: el test de significación y la combinación del test de significación con una medida de magnitud del DIF. Además, el análisis de DIF en la escala WHODAS en personas con esquizofrenia ha sido otro aporte original más aún si se tiene en cuenta que los procedimientos empleados para dicho análisis son de reciente desarrollo y aportan ventajas ya mencionadas en el texto frente a los procedimientos clásicos.

Sin embargo, hay que señalar algunas limitaciones de los estudios que componen esta tesis: En el primer estudio se reconoce que al realizar una revisión sistematica se comparten limitaciones propias de este tipo de estudios ya que la información fue obtenida de bases de datos específicas, y aunque se procuró que éstas fueran variadas y abarcaran los diferentes ámbitos de estudio que han promovido el desarrollo de los métodos de detección de DIF, es posible que se hayan perdido algunas publicaciones potencialmente relevantes.

Por otro lado, el nivel de generalización de los resultados de los estudios sobre el procedimiento diferencia del parámetro de dificultad y árboles de Rasch, está limitado por las condiciones particulares que se simularon y la versión en español que se empleó del WHODAS 2.0. En el primer caso, se debe tener en cuenta que el objetivo era hacer una simulación de las pruebas a gran escala con altas consecuencias en educación y se consideró que los valores empleados cubren muchas de las pruebas de este tipo, incluso de las pruebas internacionales como puede ser PISA (Por sus siglas en inglés: Programme for International Student Assessment), o TIMSS (Por sus siglas en ingles: Third International Mathematics and Science Study). En el caso del WHODAS, los resultados no pueden extenderse a versiones de la escala para otras culturas o idiomas hasta que no se hayan practicado estudios transculturales.

Finalmente, si bien es cierto que esta tesis ha proporcionado una visión global de las tendencias y desarrollos recientes en métodos de detección de DIF, los estudios han indicado algunos aspectos que será necesario abordar en futuras investigaciones. En primer lugar, convendría realizar una revisión sistemática con estudios de tipo empírico en el ámbito clínico o educativo y comparar las tendencias y desarrollos con los resultados obtenidos en esta tesis. En segundo lugar, hacer un análisis más profundo del efecto que pueden tener diferentes valores de la discriminación y dificultad de los ítems, esclarecer el papel que juega el tamaño del grupo focal en la detección de DIF con el procedimiento diferencia del parámetro de dificultad, y probar si el funcionamiento del procedimiento mejora al aumentar el criterio de magnitud de DIF de 0,5 a 0,64 o al emplear una métrica similar a la métrica delta del MH. En este último aspecto ya se viene trabajando y se ha redactado un documento de trabajo con resultados preliminares sobre una métrica propuesta. En tercer lugar, convendría analizar experimentalmente el funcionamiento de los procedimientos árboles de Rasch en presencia de datos simulados bajo un modelo 3PL y compararlo con el procedimiento diferencia del parámetro de dificultad. Así mismo probar diferentes tamaños de tests y covariables en diversas escalas de medida. En cuarto lugar y para aquellos interesados en la escala WHODAS, se sugiere realizar otros estudios de DIF con árboles de Rasch empleando otras medidas para las covariables depresión y sintomatología, y analizar la presencia de sintomatología positiva o negativa diferencialmente como fuente de DIF. Otro aspecto que queda pendiente por analizar hace referencia a los patrones de respuesta de las personas con esquizofrenia en el WHODAS 2.0 y la distribución de éstas en las opciones de respuesta de los ítems, especialmente en los dominios de movilidad y autocuidado.

Por todo ello, esta tesis aporta resultados de gran interés que, a pesar de las debilidades comentadas, tiene fortalezas metodológicas que permiten asegurar un impacto relevante no sólo en el ámbito científico-investigativo, sino también en el

ámbito aplicado. Esto último es más palpable si se tiene en cuenta que se trata de una tesis centrada en una temática netamente metodológica, pero cuyos estudios se han diseñado de tal manera que sus resultados tengan un impacto respecto a la mejora de la calidad de la evaluación de altas consecuencias tanto en el ámbito de salud como en el educativo. Además, los resultados han señalado algunos vacíos vigentes que han permitido mantener una continuidad en esta línea de investigación y también integrar el trabajo que vienen realizando tanto el grupo de investigación de la Universidad Nacional de Colombia, como el grupo de investigación de la Universidad de Barcelona, fortaleciendo así el trabajo colaborativo.

7. Referencias

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Ali, U. S., Chang, H.-H., & Anderson, C. J. (2015). *Location indices for ordinal polytomous items based on item response theory* (Research Report No. RR-15-20). Educational Testing Service. http://dx.doi.org/10.1002/ets2.12065

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Psychological Association.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, *37*(3), 387-416. https://doi.org/10.3102/1076998611411913

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detection differential item functioning. *Journal of Educational Measurement*, *36*(4), 277-300.

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias: a simulation study. *Advances in Statistical Analysis*, *94*, 117-127. https://doi.org/10.1007/s10182-010-0126-1

Belzak, W. C. M. (2020). Testing differential item functioning in small samples. *Multivariate Behavioral Research*, *55*(5), 722-747. https://doi.org/10.1080/00273171.2019.1671162

Bernstein, I., Samuels, E., Woo, A., & Hagge, S. L. (2013). Assessing DIF among small samples with separate calibration t and mantel-haenszel x^2 statistics in the Rasch model. *Journal of Applied Measurement*, *14*(4), 389-399.

Bollmann, S., Berger, M., & Tutz, G. (2018). Item-focused trees for the detection of differential item functioning in partial credit models. *Educational and Psychological Measurement*, *78*(5), 781-804. https://doi.org/10.1177/0013164417722179

Bolt, M. D. (2002). A monte carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*(2), 113-141. https://doi.org/10.1207/S15324818AME1502_01

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? En P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.397-417). Routledge. https://doi.org/10.4324/9780203357811

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publications.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement Issues and Practice*, *17*(1), 31-44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational measurement*, *31*(1), 67-78. https://doi.org/10.1111/j.1745-3984.1994.tb00435.x

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*(3), 333-353. https://doi.org/10.1111/j.1745-3984.1996.tb00496.x

Chen, J.-H, Chen, C.-T, & Shih, C.-L. (2014). Improving the control of type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement*, *38*(1), 18-36. https://doi.org/10.1177/0146621613488643

Chen, Y.-F, & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: En example of the

PISA 2009 reading assessment. *Educational Assessment*, *19*(2), 77-96. https://doi.org/10.1080/10627197.2014.903650

Cheong, Y. F. (2006). Analysis of school contet effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*, *6*(1), 57-79. https://doi.org/10.1207/s15327574ijt0601_4

Dabra, R. E. (1977, marzo). The identification and interpretation of item bias. *Mesa Memorandum*, *25*. http://www.rasch.org/memo25.htm

De Ayala, R. J. (2009). *The theory and practice of item response theory*. (D. A. Kenny, Ed.). Guilford Publications.

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*(6), 961-972. https://doi.org/10.1177/0013164410366691

DeMars, C. E., & Jurich, D. P. (2015). The interaction of ability differences and guessing when modeling differential item functioning with the Rasch model: conventional and tailored calibration. *Educational and Psychological Measurement*, *75*(4), 610-633. https://doi.org/10.1177/0013164414554082

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. En P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.35-66). Routledge. https://doi.org/10.4324/9780203357811

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. (1951). *Intelligence and Cultural Differences*. University of Chicago Press.

Feinberg, R. A., & Rubright, J. D. (2016), Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36-49. https://doi.org/10.1111/emip.12111

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Liberal and Conservative Differential Item Functioning Detection Using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II Error Rates. *The Journal of Experimental Education, 73*(1), 23-39. https://doi.org/10.3200/JEXE.71.1.23-40

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (1999). Aplicación de una etapa, dos etapas e iterativamente de lo estadísticos Mantel-Haenszel. *Psicológica*, *20*(3), 227-242.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*(3), 315-332. https://doi.org/10.1111/j.1745-3984.1996.tb00495.x

Galindo-Garre, F., Hidalgo, M. D., Guilera, G., Pino, O., Rojo, J. E., & Gómez-Benito, J. (2015). Modeling the World Health Organization disability assessment schedule II using non-parametric item response models. *International Journal of Methods in Psychiatric Research*, *24*(1), 1-10. https://doi.org/10.1002/mpr.1462

Gattamorta, K. A., Penfield, R. D., & Myers, N. D. (2012). Modeling item-level and step-level invariance effects in polytomous items using the partial credit model. *International Journal of Testing*, *12*(3), 252-272. https://doi.org/10.1080/15305058.2011.630546

Gómez-Benito, J., Hidalgo, M. D., & Padilla, J. L. (2009). Efficacy of effect size measures in logistic regression. An application for detecting DIF. *Methodology*, *5*(1), 18-25. https://doi.org/10.1027/1614-2241.5.1.18

Guilera, G., Gómez-Benito, J., Hidalgo, M. D., & Sánchez-Meca, J. (2007). Un meta-análisis del procedimiento Mantel-Haenszel en la detección del DIF en ítems dicotómicos. *Anuario de Psicología*, *38*(3), 431-442. https://raco.cat/index.php/AnuarioPsicologia/article/view/76574

Guilera, G., Gómez-Benito, J., Hidalgo, M. D., & Sánchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods*, 18(4), 553-571. https://doi.org/10.1037/a0034306

Guilera, G., Gómez-Benito, J., Pino, O., Rojo, J. E., Cuesta, M. J., Martínez-Arán, A., et al. (2012). Utility of the World Health Organization disability assessment schedule II in schizophrenia. *Schizophrenia Research*, *138*, 240-247. https://doi.org/10.1016/j.schres.2012.03.031

Hauser, C., & Kingsbury, G. (2004). *Differential item functioning and differential test functioning in the "Idaho standards achievement tests" for spring 2003*. Northwest Evaluation Association. https://files.eric.ed.gov/fulltext/ED491248.pdf

Herrera, A. N., Gómez-Benito, J., & Hidalgo, M. D. (2005). Detección de sesgo en los items mediante análisis de tablas de contingencia. *Avances en medición*, *3*, 29-52.

Hessen, D. J. (2003). *Differential item functioning: Types of DIF and observed score based detection methods* [Disertación doctoral, University of Amsterdam]. UvA-DARE (digital Academic Repository). https://hdl.handle.net/11245/1.250210

Hidalgo, M. D., & Gómez-Benito, J. (2010). Differential item functioning. En P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3a ed., vol. 4, pp. 36-44). Elsevier-Science & Technology.

Hidalgo, M. D. & Gómez-Benito, J. (2006). Nonuniform DIF detection using discriminant logistic analysis and multinomial logistic regression: A comparison for polytomous items. *Quality & Quantity*, *40*, 805-823. https://doi.org/10.1007/s11135-005-3964-2

Hidalgo, M. D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment*, *19*(1), 1-11. https://doi.org/10.1027//1015-5759.19.1.1

Hidalgo, M. D., Gómez-Benito, J., & Padilla, J. L. (2005). Regresión logística: alternativas de análisis en la detección del funcionamiento diferencial del ítem. *Psicothema*, *17*(3), 509-515. https://www.redalyc.org/articulo.oa?id=72717324

Hidalgo, M. D., & López-Pina, J. E. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and mantel-haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903-915. https://doi.org/10.1177/0013164403261769

Hidalgo-Montesinos, M. D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment*, *19*(1), 1-11. https://doi.org/10.1027/1015-5759.19.1.1

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Erlbaum.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Report No. RR-85-43). Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00128.x

Huang, H.-Y. (2014). Effects of the common scales setting in the assessment of differential item functioning. *Psychological Reports: Measures & Statistics*, *114*(1), 104-125. https://doi.org/10.2466/03.PR0.114k11w0

Kankaras, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence or ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, *40*(2), 279-310. https://doi.org/10.1177/0049124111405301

Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch model. *Educational and Psychological Measurement*, *78*(1), 128-166. https://doi.org/10.1177/0013164416664394

Kopf, J. (2013). *Model-based recursive partitioning meets item response theory: New statistical methods for the detection of differential item functioning and appropriate anchor selection* [Disertación doctoral, Ludwig-Maimilians-Universität München]. Elektronische Hochschulschriften. https://edoc.ub.uni-muenchen.de/16434/

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selections strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22-56. https://doi.org/10.1177/0013164414529792

Lai, J.-S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & The Health Professions*, *28*(3), 283-294. https://doi.org/10.1177/0163278705278276

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647-677. https://doi.org/10.1007/BF02294041

Li, Y., Brooks, G., Johanson, G. A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, *72(5)*, 847-861. https://doi.org/10.1177/0013164411432333

Linacre, J. M. (2021). *Winsteps ® Rasch measurement computer program User's Guide*. Version 5.1.0. Winsteps.com.

Linacre, J. M. (2012). *A user's guide to Winsteps & ministeps Rasch-Model computer programs*. Winsteps.com.

Linacre, J. M. (1999). Category disordering (disordered categories) vs. threshold disordering (disordered thresholds). *Rasch Measurement Transactions*, *13*(1), 675. https://www.rasch.org/rmt /rmt131a.htm

Linacre, J. M. (1994). Sample Size and Item Calibration [or Person Measure] Stability. *Rasch Measurement Transactions*, *7*(4), 328. https://www.rasch.org/rmt/rmt74m.htm

Linacre, J. M., & Wright, B. D. (1987, febrero). Item Bias: Mantel-Haenszel and the Rasch Model. *Mesa Memorandum*, *39*. https://www.rasch.org/memo39.pdf.

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnosis models. *Frontiers in Psychology*, *10*, 1137. https://doi.org/10.3389/fpsyg.2019.01137

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Magidson, J. & Vermunt, J. K. (2001). Latent Class Factor and Cluster Models, Biplots and Related Graphical Displays. *Sociological Methodology*, *31*, 223-264. https://doi.org/10.1111/0081-1750.00096

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

Magis, D., & Facon, B. (2014). deltaPlotR: An R package for differential item functioning analysis with Angoff's delta plot. *Journal of Statistical Software*, *59*, 1-19. https://doi.org/10.18637/jss.v059.c01

Magis, D., & Facon, B. (2012). Angoff's delta method revisited: Improving DIF detection under small samples. *British Journal of Mathematical & Statistical Psychology*, *62*(2), 302-321. https://doi.org/10.1111/j.2044-8317.2011.02025.x

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). *A review of recent developments in differential item functioning* (Research Report No. RR-08-43). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02129.x

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543. https://doi.org/10.1007/BF02294825

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012-1027. https://doi.org/10.1037/0003-066X.35.11.1012

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on attitude survey. *Journal of Educational Measurement*, *41*(4), 331-344.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461-473. https://doi.org/10.1007/s11336-007-9039-7

Millsap, R. E., & Everson, H. T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297-334. https://doi.org/10.1177/014662169301700401

Millsap, R. E., & Meredith, W. (1992). Inferential Conditions in the Statistical Detection of Measurement Bias. *Applied Psychological Measurement*, *16*(4), 389-402. https://doi.org/10.1177/014662169201600411

Moher, D., Liberati, A., Tetzlaff, J., Altman, D., and the PRISMA Group. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*, *8*, 336-341. https://doi.org/10.1016/j.ijsu.2010.02.007

Nuño, L., Barrios, M., Rojo, E., Gómez-Benito, J., & Guilera, G. (2018). Validation of the ICF core sets for schizophrenia from the perspective of psychiatrists: An international Delphi study. *Journal of Psychiatric Research*, *103*, 134-141. https://doi.org/10.1016/j.jpsychires.2018.05.012

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(2), 107-124. https://doi.org/10.1080/10705519809540095

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning Quantitative applications in the social science* (2ª ed., Vol. 161). Sage publications.

Padilla, J. L., & Benítez, I. (2017). A rationale for and demonstration of the use of DIF and mixed methods. En B. D. Zumbo & A. M. Hubley (Eds.). *Understanding and investigating response processes in validation research* (pp. 193-209). Springer International Publishing. https://doi.org/10.1007/978-3-319-56129-5_11

Paek, I., & Guo, H. W. (2011). Accuracy of DIF estimates and power in unbalanced designs using the Mantel-Haenszel DIF detection procedure. *Applied Psychological Measurement*, *35*, 518-535. https://doi.org/10.1177/0146621611420559

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maimum likelihood estimation context and its comparison with mantel-haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*(6), 1023-1046. https://doi.org/10.1177/0013164411400734

Penfield, R. D. (2010) Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement*, *47*(2), 129-149.

Penfield, R. D., Alvarez, K., & Lee, O. (2009) Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education*, *22*(1), 61-78. https://doi.org/10.1080/08957340802558367

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, *40*(4), 353-370. https://doi.org/10.1111/j.1745-3984.2003.tb01151.x

Potenza, M. T., & Dorans, N. J. (1995). DIF Assessment for Polytomously Scored Items: A Framework for Classification and Evaluation. *Applied Psychological Measurement*, *19*(1), 23–37. https://doi.org/10.1177/014662169501900104

Prieto-Marañón, P., Aguerri, M. E., Galibert, M. S., & Attorresi, H. F. (2012). Detection of differential item functioning using decision rules based on Mantel-Haenszel procedure and Breslow-Day Tests. *Methodology*, *8*(2), 63–70. https://doi.org/10.1027/1614-2241/a000038

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and mantel-haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116. https://doi.org/10.1177/014662169301700201

Rouquette, A., Hardouin, J.-B., Vanhaesebrouck, A., Sébille, V., & Coste, J. (2019). Differential item functioning (DIF) in composite health measurement scale: recommendations for characterizing DIF with meaningful consequences within the Rasch model framework. *Plos One*, *14*(4), e0215073. https://doi.org/10.1371/journal.pone.0215073

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*, 1–97. https://doi.org/10.1007/BF03372160

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1996). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing different item functioning. En G. Engelhard, Jr., & M. Wilson (Eds.). *Objective Measurement: Theory into practice*, (Vol. 3, pp. 65-82). Ablex Publishing Company.

Sedano-Capdevila, A., Barrigón, M. L., Delgado-Gomez, D., Barahona, I., Aroca, F., Peñuelas-Calvo, I., Miguelez-Fernandez, M., Rodríguez-Jover, A., Amodeo-Escribano, S., González-Granado, M., & Baca-García, E. (2018). WHODAS 2.0 as a measure of severity of illness: Results of a FLDA Analysis. *Computational and Mathematical Methods in Medicine*, *2018*, Article 7353624. https://doi.org/10.1155/2018/7353624

Seol, H. (1999). Detecting differential item functioning with five standardized item-fit índices in the Rasch model. *Journal of Outcome Measurement*, *3*(3), 233-247.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test

bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194. https://doi.org/10.1007/BF02294572

Sinharay, S., Dorans, N. J., Grant, M. C., & Blew, E. O. (2009). Using past data to enhance small sample DIF estimation: A bayesian approach. *Journal of Educational and Behavioral Statistical*, *34*(1), 74-96. https://doi.org/10.3102/1076998607309021

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187. http://dx.doi.org/10.1080/13803611.2013.767621

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289-316. https://doi.org/10.1007/s11336-013-9388-3

Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology science quarterly*, *50*(4), 538-612.

Tutz, G., & Berger, M. (2016). Item-focused trees for the identification of items in differential item functioning. *Psychometrika*, *81*(3), 727-750. https://doi.org/10.1007/s11336-015-9488-3

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70. https://doi.org/10.1177/109442810031002

Van Eck, N. J., & Waltman, L. (2018). *VOSviewer Visualizing scientific landscapes* (versión 1.6.9) [Software]. Universiteit Leiden & CWTS: Meaningful Metrics. https://www.vosviewer.com

Wang, W.-C. (2004). Effect of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Eperimental Education*, *72*(3), 221-261. https://doi.org/10.3200/JEXE.72.3.221-261

Wang, W.-C., & Su, Y.-H. (2004). Factors influencing the mantel and generalized mantel-haenzel methods for the assessment of differential item functioning in

polytomous items. *Applied Psychological Measurement*, *28*(6), 450-480. https://doi.org/10.1177/0146621604269792

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for posytomous items. *Applied Psychological Measurement*, *30*(1), 22-42. https://doi.org/10.1177/0146621605279867

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense. https://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf

Zwick, R., Thayer, D., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical bayes approach. *Journal of Educational and Behavioral Statistics*, *25*(2), 225-247. https://doi.org/10.2307/1165333

8. Anexos

8.1. Anexo 1: Material suplementario del Estudio 1: Revisión sistemática

Supplementary material 1

List of keywords grouped by cluster.

Cluster	Terms	Frequency	Strength	Mean year
Red	DIF	161	449	2008.7
	IRT	57	188	2009
	IRT-LR-DIF	13	49	2009.8
	MIMIC	8	23	2011.6
	Confirmatory factor analysis	6	21	2010.2
	Likelihood ratio test	5	15	2009
	Scale purification	3	11	2010
	Matching variable	2	8	2010.5
	Missing data	2	8	2010
	Multiple imputation	2	8	2013.5
	Empirical histogram	2	6	2009.5
	Measurement equivalence	2	4	2007
	Ordinal logistic regression	2	3	2010
	Model selection strategies	1	2	2012
	Two-Way DIF	1	2	2012
Green	Logistic regression	22	85	2008.2
	Monte Carlo simulation	7	36	2011.7
	Log-Linear model	2	22	2003
	Signed area	2	22	2003
	SOS1	2	22	2003
	SOS3	2	22	2003
	Unsigned area	2	22	2003
	Non-Directional DIF	2	15	2000
	Comparative survey research	1	13	2002
	Cross-Cultural survey research	1	13	2002
	Two-Stage bias detection	1	13	2002
	Logit model	1	5	2002
	Restricted factor analysis	1	5	2002
Blue	Testlets	6	13	2011.3
	Simulation	4	10	2012
	Linking	3	16	2014.3
	Generalized graded unfolding model	3	14	2012.7
	Unfolding	2	13	2001.5
	Em Algorithm	2	9	2002.5
	Equating	2	6	2015
	POLYSIBTEST	2	6	2011
	Latent trait theory	1	6	1992
	Marginal maximum likelihood	1	6	1992
	Non-Monotone trace lines	1	6	1992
	Single-Peaked preference functions	1	6	1992
	Multiple-Group bifactor model	1	3	2013

Cluster	Terms	Frequency	Strength	Mean year
Yellow	Measurement invariance	19	65	2011.6
	Type I error	18	61	2006.6
	Graded response model	9	32	2005.4
	Power	8	28	2007.1
	Partial credit model	3	13	2008.7
	Measurement bias	3	10	2005.3
	Test purification	3	10	2004.3
	Model-Based recursive partitioning	2	9	2016.5
	Average signed area	2	8	2003.5
	Factor analysis	2	5	2011
	Latent moderated structures	2	5	2011
	Rating scale model	1	5	2018
	Tree	1	5	2018
Purple	Polytomous items	17	44	2007.7
	Bias	4	19	2004
	DIF detection	4	9	2006.3
	Generalized Mantel-Haenszel	3	8	2009.3
	Unidirectional DIF	2	14	1996
	Crossing DIF	2	13	2001.5
	Multidimensional IRT	2	10	2006.5
	Cross-Cultural assessment	2	9	2008.5
	Logistic regression procedure crossing	1	8	1996
	Randomization test	1	8	1996
	Multigroup comparisons	1	6	2007
Turquoise	SIBTEST	21	84	2005.4
	Item response function	3	16	1997.7
	Multidimensionality	3	9	2005
	Regression correction	2	10	1997.5
	Kernel smoothing	2	8	2003
	Local DIF	1	6	1996
	Log odds ratio	1	6	1999
	M-H D DIF parameter statistic	1	6	1999
	MULTISIB	1	6	1997
	Smoothed SIBTEST	1	6	1996
	Testgraf	1	6	1996
Orange	Rasch model	18	59	2011
	Effect size	8	19	2011.1
	Two-parameter logistic model	2	22	2000
	Generalized partial credit model	2	13	2003
	Model fit	2	13	2006
	Recursive partitioning	2	4	2016
	Lagrange multiplier test	1	9	1998
	Nominal response model	1	9	1998
	OPLM	1	9	1998
	Rao's efficient score test	1	9	1998
Brown	Empirical Bayes	4	13	2007
	Loss function	4	13	2004.5
	Small sample	4	11	2009.5
	CAT	3	9	2006.3
	Penalized maximum likelihood	3	7	2016
	Bayesian analysis	2	6	2011

Cluster	Terms	Frequency	Strength	Mean year
	MCMC	2	6	2012
	Educational assessment	2	5	2013
	Law School Admission Test	1	6	2002
Pink	DFIT	8	41	2007.8
	DTF	7	38	2003
	Nonuniform DIF	5	26	2003.8
	Uniform DIF	4	24	2001.8
	Lord's chi-2	3	14	1994
	Three-parameter logistic model	3	14	2009.3
	Area measures	2	12	1994
	Compensatory DIF	1	8	1995
	Non-Compensatory DIF	1	8	1995
Salmon Pink	Mantel-Haenszel	39	124	2005.4
	Simulation study	3	14	2009.7
	Sample size	3	11	2007.3
	Ability distribution	1	7	2015
	International assessment	1	7	2015
	Thick matching	1	7	2015
	Thin matching	1	7	2015
	Type II error	1	5	2004
Olive Green	Item bias	19	87	2003.3
	Test bias	3	12	2001.3
	Ordinal items	3	5	2009
	Statistics	3	4	2009.7
	Research methodology	2	4	2005.5
	Bias/DIF	1	8	1993
	Standardization	1	8	1993
	Valid subtest	1	8	1993

Supplementary material 2

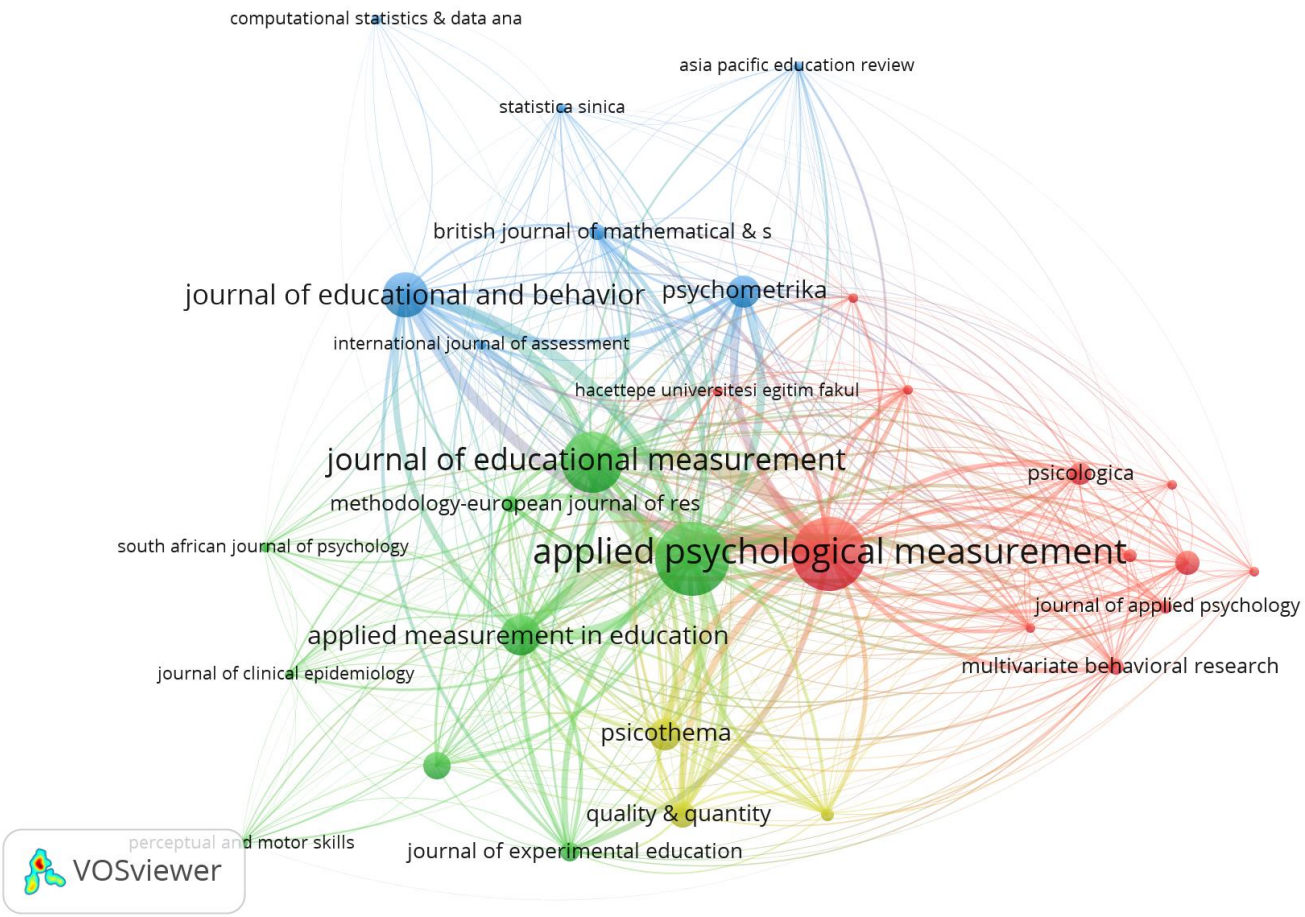


Fig. 1s Distance-based map of journals, constructed using the bibliographic coupling approach.

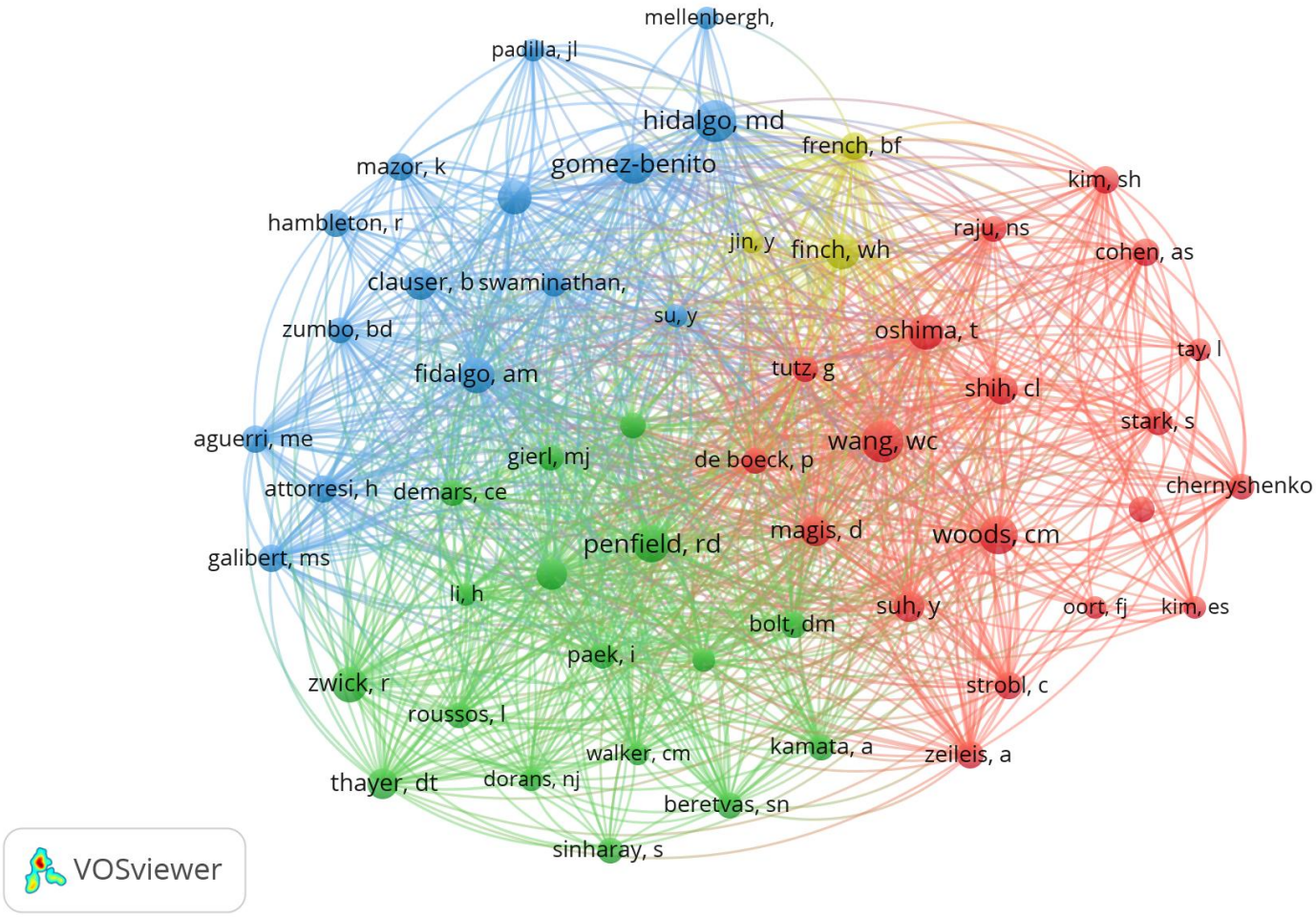


Fig. 2s Distance-based map of authors collaboration, constructed using the bibliographic coupling approach.

8.2. Anexo 2: Material suplementario del Estudio 2: Análisis con datos simulados

Supplementary material 1: SM1
False positive rates associated with the different independent variables when applying the significance test as the criterion.

Impact	Sample size ratio				
	20	100	250	500	1000
			Model fit		
Equal groups	**.05**/.43/.62	**.05**/.15/.21	**.05**/.09/.11	**.05**/**.07**/.08	**.04**/**.06**/**.07**
μ different	**.06**/.35/.42	**.05**/.12/.14	**.05**/.08/.09	**.05**/**.06**/**.07**	**.04**/**.05**/**.06**
σ^2 different	**.06**/.32/.60	**.05**/.11/.19	**.05**/.08/.11	**.05**/**.06**/**.07**	**.04**/**.05**/**.06**
μ and σ^2 both different	.11/.24/.38	**.06**/.09/.12	**.05**/**.07**/.08	**.05**/**.06**/**.07**	**.04**/**.05**/**.05**
			Model misfit		
Equal groups	**.05**/.27/.49	**.05**/.10/.16	**.06**/.08/.10	**.06**/**.07**/.08	**.05**/**.05**/**.06**
μ different	.88/.91/.95	.83/.81/.83	.70/.70/.71	.55/.53/.55	.35/.35/.33
σ^2 different	.46/.54/.64	.27/.28/.33	.17/.17/.20	.14/.14/.15	.15/.14/.15
μ and σ^2 both different	.85/.86/.87	.75/.74/.74	.62/.63/.63	.46/.46/.48	.30/.27/.28

Note. The values in each cell express the false positive (FP) rates under the conditions of 0% / 10% / 20% of items with DIF, respectively. Values in bold correspond to mean FP rates ≤ .07.

Supplementary material 2: SM2

False positive rates associated with the different independent variables when applying the combination of the significance test with the measure of DIF magnitude as the criterion.

Impact	Sample size ratio				
	20	100	250	500	1000
			Model fit		
Equal groups	**<.01/<.01/<.01**	**<.01/<.01/<.01**	**<.01/.01/.01**	**.02/.03/.04**	**.04/.05/.06**
μ different	**<.01/<.01/<.01**	**.01/.02/.02**	**.02/.03/.03**	**.03/.04/.04**	**.04/.05/.05**
σ^2 different	**<.01/<.01/<.01**	**<.01/<.01/<.01**	**<.01/.01/.01**	**.02/.03/.03**	**.04/.05/.06**
μ and σ^2 both different	**<.01/<.01/<.01**	**<.01/<.01/.01**	**.02/.03/.03**	**.03/.04/.04**	**.04/.05/.05**
			Model misfit		
Equal groups	**<.01/<.01/<.01**	**<.01/<.01/<.01**	**<.01/<.01/<.01**	**<.01/<.01/<.01**	**.03/.02/.02**
μ different	.21/.19/.23	.22/.21/.23	.27/.25/.27	.28/.26/.28	.27/.26/.26
σ^2 different	**.03/.04/.04**	**.03/.03/.03**	**.03/.02/.02**	**.04/.04/.04**	.11/.10/.11
μ and σ^2 both different	.16/.16/.19	.19/.17/.19	.24/.21/.23	.24/.22/.24	.24/.21/.22

Note. The values in each cell express the false positive (FP) rates under the conditions of 0% / 10% / 20% of items with DIF, respectively. Values in bold correspond to mean FP rates ≤ .07.

8.3. Anexo 3: Material suplementario del Estudio 3: Análisis con datos empíricos

Supplementary material 1:
Summary of the algorithms followed by the DIF detection methods.

TREE-PCM by Psychotree

1. The parameter of the items is estimated using the common conditional maximum likelihood approach for the entire sample.
2. It is evaluated whether the parameter of the items differs throughout any one of the covariates.
3. The sample is split along the covariate that shows the biggest difference and two separate Rasch models are estimated at the cutoff point that leads to the best fit of the model using the Lagrange-multiplier statistic.
4. This process (steps 1 to 3) is repeated iteratively until a stopping criterion is reached. There are two kinds of stopping criteria: if there are no more significant parameter differences or if the sub-sample is too small.

PCM-IFT by DIFtree

1. It is adjusted for all the items to the Rasch model, examining all the covariates and possible splits for all of them. Item parameter is estimated through joint maximum likelihood and splits are established by examining the null hypothesis of equal parameter of item by subgroups by the Likelihood Ratio (LR) test.
2. The model that has the best fit or minimal deviance is selected. Of all the possible item, covariate and split combinations, the best split-point found for one item and one covariate is the one that produces the smallest p value.
3. It is decided whether or not to perform splitting based on the dependence of item parameter and the selected covariate. For a particular item and covariate, the split-point with maximum LR test should be selected by permutation test. If it is significant, the split point is confirmed and model parameters are estimated, if not the procedure stops and no DIF is detected. Another criterion to define the split-point is the minimal sample size in each node.

UNIVERSITAT DE BARCELONA