UNIVERSIDAD DE MURCIA

FACULTAD DE PSICOLOGÍA

Funcionamiento diferencial del ítem:

Una aproximación a la regresión logística

Dña. Mª Dolores López Martínez

2017

UNIVERSIDAD DE MURCIA

FACULTAD DE PSICOLOGÍA

Tesis Doctoral:

Funcionamiento diferencial del ítem: Una aproximación a la regresión logística

Dirigida por:

Dra. Dña. Mª Dolores Hidalgo Montesinos

Catedrática del Departamento de Psicología Básica y Metodología

de la Universidad de Murcia

Dra. Dña. Juana Gómez Benito

Catedrática del Departamento de Psicología Social y Psicología Cuantitativa

de la Universidad de Barcelona

Dra. Dña. Georgina Guilera Ferré

Profesora Agregada del Departamento de Psicología Social y Psicología Cuantitativa

de la Universidad de Barcelona

Presentada por:

Dña. Mª Dolores López Martínez

2017

En la actualidad, la humanidad lucha por conseguir una sociedad más justa y equitativa. Una lucha, donde lo que prima es alcanzar la igualdad de oportunidades, para todos los seres humanos. Por lo tanto, bajo este esfuerzo de todos y todas, no hay otra opción, que trabajar duro por conseguir que la medición y evaluación psicológica, educativa y social, sean lo más equitativas y justas posibles por y para todos y todas.

A mi familia

Agradecimientos

Me gustaría expresar mi más sincero agradecimiento a todas las personas que han colaborado, directa o indirectamente, para que esta tesis se haya llevado a cabo.

Especialmente a mis directoras, Dra. Hidalgo, Dra. Gómez y Dra. Guilera, por la orientación, el seguimiento y la supervisión continúa que he recibido de ellas. Pero, sobre todo, por la motivación y el apoyo que me han dado, por su entusiasmo, su paciencia, su dedicación y su entrega. Durante estos años, me han dado su confianza, me han apoyado y animado incondicionalmente, han sido flexibles, constantes y comprometidas conmigo, a pesar de mis circunstancias personales y laborales. En esta tesis, me llevo parte de su sabiduría y experiencia, porque, con ellas sabía que esto sólo podía salir bien.

Especialmente, quiero agradecer el apoyo humano y personal de mi directora, en la Universidad de Murcia, Lola Hidalgo. Ella ha sido mi mentora, mi amiga, mi psicóloga y mi apoyo. Lola ha confiado en mí, sin titubear, ha sido comprensiva y me ha animado constantemente. Ella me ha ido mostrando la luz en este largo camino y me ha transmitido la pasión por la Psicometría.

De igual modo, quiero agradecer a mis padres y a mi hermana, que siempre han estado, y están, ahí, junto a mí, formando parte y apoyándome en lo personal y en lo profesional. Porque, como dijo, y dice, mi padre “sin trabajo duro y si esfuerzo no se alcanzan tus sueños” (cuando, a sus treinta años, hacía la maleta para mudarse de Jaén a Madrid).

Gracias a mis buenos amigos y amigas. Gracias a Fernando que, en este último año, ha sido imprescindible en mí día a día.

A todas ellas y a todos ellos, gracias.

Índice

Índice de tablas

Índice de figuras

INTRODUCCIÓN

La investigación psicométrica realiza grandes esfuerzos por asegurar la equidad de las mediciones y evaluaciones. Es a lo largo de siglo XX cuando cobra mayor intensidad este esfuerzo y se desarrolla, más intensamente, la investigación en este campo. Es también durante este mismo siglo, en el que los tests pasan a ser el instrumento de medida estandarizado más empleado en las ciencias sociales y de la salud.

Ya a comienzos de siglo, se ponen de manifiesto diferencias en el funcionamiento de algunos tests. Esta preocupación hace que se dedique cierta atención al estudio de los tests, y a su validez. Durante los años 60 ya se habla de tests sesgados, refiriéndose a aquellos que tienden a favorecer, en sus resultados, a un grupo de examinados frente a otro (habitualmente a mayorías frente a minorías).

El concepto de validez hace referencia al grado en el que el test mide lo que pretende medir o sirve para el propósito por el que ha sido construido. Este concepto ha ido reafirmando su protagonismo, dentro de la psicometría, a lo largo del tiempo. En 1985 los

Standards of Educational and Psychological Testing (APA, AERA, NCME, 1985, p. 8), indican que "La validez es la consideración más importante en la evaluación de un test. El concepto se refiere a la adecuación, significado y utilidad de las inferencias específicas hechas con las puntuaciones de los tests".

El concepto de sesgo también ha ido cambiando, en 1988, Holland y Thayer sugieren cambiar el término sesgo por el término funcionamiento diferencial del ítem (DIF), ya que el término sesgo mostraba una connotación negativa equiparable con injusticia, parcialidad e inequidad contra los grupos minoritarios o menos favorecidos. Así, una definición de DIF la dan Clauser y Mazor (1998), indicando que este se da cuando sujetos con el mismo nivel de habilidad, pero de grupos diferentes, tienen diferentes probabilidades de acertar un ítem.

El creciente interés despertado por la validez de los cuestionarios y tests no cesa. La penúltima edición de los Standards for Educational and Psychological Testing (AERA, APA y NCME, 1999, p. 9) reafirma la importancia de la validez, al valorarla como "… la consideración más importante en la elaboración y evaluación de los tests". Este interés continúa hasta la actualidad, la última edición de los *Standards for Educational and Psychological Testing* de 2014, continúa dedicando una especial atención a la necesidad de garantizar el uso correcto de los tests.

La comprobación de la equidad de los instrumentos de medición se constituye como una cuestión obligada, dada su importancia y dada la existencia de herramientas que la evalúan. Esta comprobación acaba siendo una cuestión básica para decidir el uso o no de los instrumentos, en un contexto aplicado. Por ello, es importante exigir a los cuestionarios y tests que sean fiables (precisos) y válidos.

El estudio del DIF también se ve impulsado por la necesidad de asegurar la equivalencia entre varias versiones de un mismo test, debido al progresivo desarrollo de los estudios transculturales y transidiomáticos. Dado el mundo globalizado en el que vivimos, cada vez es más frecuente encontrar estudios que buscan comparar a diferentes poblaciones o culturas o que demandan el instrumento para hacerlo. La adaptación de tests y cuestionarios, para estos fines, se ha convertido en un trabajo cotidiano para los psicómetras.

Actualmente, presenta gran relevancia el estudio de la eficiencia y efectividad de las técnicas de detección del DIF, con el fin de evaluar la adecuación de éstas (potencia y error) en sus evaluaciones y los diferentes contextos en los que se recomienda aplicar una u otra.

En la siguiente tesis se presentan tres trabajos que estudian el funcionamiento de la regresión logística (RL) como técnica de análisis y detección del DIF.

En primer lugar, se presenta una revisión bibliométrica con el fin de analizar la producción científica relacionada con el uso de la RL para detectar el DIF, con el objetivo de ofrecer una visión general de la actividad de investigación en este campo y caracterizar sus aspectos más importantes y su evolución durante la última década del siglo XX y principios del siglo XXI.

En segundo lugar, se presenta una revisión sistemática de la literatura que plantea una revisión de los trabajos de simulación encontrados, con el fin de obtener un mapa de los estudios realizados hasta el momento sobre el funcionamiento de la RL, bajo distintas condiciones, en la detección de ítems con DIF.

En tercer y último lugar, se presenta un estudio de simulación, con el fin de comparar la eficacia para detectar ítems con DIF de la regresión logística discriminante (DLR), técnica que emplea la puntuación observada del test como criterio de equiparación, y de IRTLRDIF, una técnica basada en la teoría de respuesta a los ítems, en tests cortos con ítems politómicos.

Esta tesis pretende aportar algo más de luz en el estudio de la RL como técnica de detección de DIF, facilitando recomendaciones y pautas para un óptimo uso de este procedimiento en la detección del funcionamiento diferencial de los ítems.

CAPÍTULO 1

MARCO TEÓRICO: DIF Y PROCEDIMIENTOS DE ANÁLISIS

1.1. Desarrollo histórico

A partir del siglo XX, comienza la preocupación por el estudio del sesgo en los tests, y por la justicia en las mediciones en Psicología y Educación.

A comienzos de este siglo, Binet y Simon (1908), en la primera revisión que realizan de su escala de inteligencia, muestran preocupación por un posible sesgo en algunos de los ítems de la escala; éstos observan similitud de resultados entre niños y niñas de ambientes semejantes. Los niños y niñas de familias acomodadas solían presentar una edad mental superior a la edad mental de los que pertenecían a los suburbios, es decir, se daba un peor rendimiento en niños y niñas de estatus socioeconómico más bajo. Ambos autores, concluyeron que ésto podía deberse a los efectos del entrenamiento cultural, en lugar de a diferencias reales en el constructo medido. Por lo que, en la última de las revisiones que hicieron de su escala, Binet y Simon (1911) suprimieron algunas de las pruebas de la edición

de 1908, especialmente aquéllas donde se había comprobado que se reflejaba información adquirida en la escuela o conocimientos rutinarios.

También Stern (1914), quién acuñó el término de "cociente intelectual", al estudiar las diferencias relacionadas con la clase social en Alemania, encontró que los tests podrían favorecer las diferencias de una clase social sobre otra.

Pero no es hasta 1951 cuando Eells, Davis, Havighurst, Herrick y Tyler, en un estudio realizado en la Universidad de Chicago, ponen de manifiesto las disparidades del funcionamiento en algunos tests de inteligencia en función del grupo donde se aplicaran. Este trabajo puede considerarse como el pionero en el estudio del sesgo de los ítems, como muchos expertos en el tema reconocen (Camilli y Shepard, 1994).

Es en este momento cuando se da comienzo a la mayor preocupación por la justicia en los tests, y la comunidad científica comienza a emplear el concepto de "sesgo".

En el desarrollo histórico del sesgo, Jensen (1969) fue también importante. Éste defendía que la inteligencia era heredada, por lo que las diferencias observadas en las pruebas entre grupos raciales se explicaban genéticamente. La posición de Jensen, partidario de una explicación genética, era contraria a la de otros. En contra se hallaban los partidarios de explicar estas diferencias mediante determinantes ambientales y sociales. Estos últimos atribuían las diferencias entre grupos al sesgo de las pruebas, como Kagan (1975) que tras analizar el tipo de preguntas e ítem de tests de C.I. estandarizados, concluye que éstos son un instrumento seriamente sesgado, que casi garantiza puntajes más altos a los niños blancos de clase media que a los niños de otros grupos.

La polémica generada alrededor del trabajo de Jensen (1969) puso de manifiesto la necesidad de evaluar hasta qué punto las diferencias observadas en las pruebas se debían a las características reales de los grupos o a errores generados por el instrumento.

A todo este interés por la búsqueda de la justicia en los tests, se le añaden algunas sentencias judiciales importantes por discriminación en la selección de personal, como la de Griggs, en 1971, coincidiendo con el movimiento de los derechos civiles en Estados Unidos, donde se reivindicaba igualdad de derechos y oportunidades para los grupos más desfavorecidos, y el empleo de test como fuente de información para la toma de decisiones; es fácil de entender cómo surge el mayor interés, entre los investigadores, por explicar estas disparidades y por obtener "tests libres de cultura".

Jensen (1980) vuelve a poner de manifiesto un posible sesgo cultural de los tests. En su libro "Bias in Mental Testing" afirma que ni las pruebas verbales de inteligencia, ni las no verbales, están sesgadas de manera significativa en contra de los niños nacidos en Estados Unidos, pero pertenecientes a grupos minoritarios. Dándose, así, mayor controversia acerca de la parcialidad de los tests respecto a determinados grupos.

Aunque, tal y como indica Muñiz (1998), las publicaciones psicométricas especializadas de los 50's y los 60's y la edición de 1966 de los *Standards for Educational and Psychological Test and Manuals* habían ignorado por completo este tema; si miramos hacia atrás, como dicen, Osterlind y Everson (2009), podemos ver cómo a finales de los setenta comienza el desarrollo de métodos estadísticos para la identificación de los ítems potencialmente sesgados. Apropiándose la comunidad psicométrica de la discusión que, hasta

ese momento, se había mantenido en las esferas legal, política, social y de la teoría psicológica. La comunidad psicométrica comienza a establecer criterios objetivos para los análisis de los tests, proponiéndose las primeras técnicas analíticas. Sin embargo, no es hasta mediados de los años ochenta, cuando surge el marco estadístico general y viable para el análisis del DIF a gran escala, introducidos por Paul Holland y sus colegas del *Educational Testing Service* (Holland, 1985, Holland y Thayer, 1988).

En este período queda consolidado el interés por la validez de los tests y cuestionarios. Los esfuerzos se centran en poder evaluar y asegurar que los instrumentos de medida que se emplean son válidos, analizando los ítems que son variables a través de las distintas muestras, pertenecientes a la misma población. Este fenómeno se conoce como funcionamiento diferencial del ítem (DIF), y se produce cuando los subgrupos de los examinados tienen los mismos niveles en el rasgo pero difieren en sus probabilidades de dar una respuesta correcta (Roussos y Stout, 1996).

1.2. Sesgo, impacto y DIF

El concepto de *sesgo* se comienza a emplear por la comunidad científica en los años 60. Cleary (1968, p. 115) ofrece una definición de "Test sesgado":

> Una test está sesgado para los miembros de un subgrupo de la población si, en la predicción de un criterio, para el cual se diseñó la prueba, se hacen errores consistentes de predicción no nulos para los miembros del subgrupo. En otras palabras, la prueba está sesgada si la puntuación de criterio pronosticada a partir de la línea de regresión común es consistentemente demasiado alta demasiado o baja para los miembros del subgrupo.

Tras la publicación del trabajo de Jensen (1969), comienza la polémica alrededor de las diferencias observadas en las pruebas, y cabía aclarar si éstas se debían a las características reales de los grupos o a errores generados por el instrumento. Este debate generó un nuevo conflicto semántico: ¿sesgo cultural o propiedades psicométricas distintas?

En este debate, el concepto de sesgo se asocia a cualquier diferencia sistemática entre grupos diferentes y, en consecuencia, el término sesgo tiene una connotación negativa equiparable con injusticia, parcialidad e inequidad contra los grupos minoritarios o menos favorecidos.

En los años 90, diversos autores como Angoff (1993), Cole (1993) y Holland y Wainer (1993) coincidían a la hora de explicar el conflicto existente en torno al sesgo. Para éstos el debate partía de un conflicto semántico, una confusión de los términos en su uso cotidiano. Público y psicólogos estaban usando el mismo término "sesgo" pero para el público tenía grandes connotaciones negativas, de tipo social y político, y para los psicólogos estaba cargado de contenido técnico, y aunque la connotación no era buena, hacía referencia básicamente a 'características técnicas no óptimas' (Cole, 1993) y no a injusticia social.

Fidalgo (1996), en cambio, identifica en este debate a "la falacia igualitaria" que se basa en el supuesto de igualdad entre los hombres. De tal manera que, los tests o instrumentos de medida que pusieran en evidencia diferencias entre grupos de humanos, resultaban discriminatorios y sesgados.

Con el fin de clarificar los conceptos y superar el conflicto semántico existente, Holland y Thayer (1988) sugirieren cambiar el término sesgo, que muestra implicaciones

sociales y de injusticia de la medición, por el de funcionamiento diferencial de los ítems (DIF).

En la actualidad, tal y como reseña Gómez y Navas (1998), existe un consenso bastante generalizado entre los autores que investigan estas cuestiones en reservar el término DIF para el análisis estadístico de la cuestión y el término sesgo para las inferencias sobre la naturaleza de las diferencias observadas (Scheuneman, 1982; Scheuneman y Bleistein, 1989). Como indica Camilli (1992), el sesgo implica DIF y aspectos cualitativos de los ítems del test.

El sesgo puede aparecer de múltiples formas, puede ser de género, cultural, étnico, religioso, etc. Un ítem puede estar sesgado si presenta un contenido o lenguaje que es diferencialmente familiar para subgrupos de examinados, o si la estructura o formato de los ítems es diferencialmente difícil para subgrupos de examinados (Hambleton y Rodgers, 1995). Estos mismos autores, elaboran una guía para revisiones de sesgo, plantean tres cuestiones a considerar cuando se evalúan los ítems: la equidad, el sesgo y los esterotipos (la representación inadecuada o desfavorable de los subgrupos designados).

Una definición de sesgo la ofrecen Gómez-Benito, Hidalgo y Guilera (2010, p.76):

> El sesgo se refiere a la injusticia derivada de uno o varios ítems del test al comparar distintos grupos que se produce como consecuencia de la existencia de alguna característica del ítem o del contexto de aplicación del test que es irrelevante para el atributo medido por el ítem.

Como puede deducirse, la existencia de sesgo en los ítems de un test supone una amenaza a la validez de éste, entendiendo por validez el grado en que la evidencia empírica y el razonamiento teórico apoyan la adecuación e idoneidad de las interpretaciones basadas en

las puntuaciones, de acuerdo con los usos propuestos por el test (Messick, 1989; Prieto y Delgado, 2010). La validez del test puede verse afectada si sus ítems benefician a ciertos grupos de la población en detrimento de otros, de igual nivel en el rasgo que se mide.

Establecer la validez de un test implica también obtener evidencia de que el instrumento con el que se trabaja está libre de sesgo (Navas, 2001), es decir, los ítems del test funcionan del mismo modo para distintos grupos en función de variables sociodemográficas, cognitivas o de cualquier otro tipo que pueda constituir una fuente sistemática de variación ajena al constructo medido por el test (Muñiz, 1996).

A este respecto, cabe distinguir el sesgo del impacto. Ackerman (1992) define el impacto como una diferencia entre grupos en el desempeño en un ítem causada por una diferencia real en la variable medida, mientras que el sesgo se relaciona con diferencias causadas por fuentes sistemáticas de variación ajenas al constructo que mide el test.

En un estudio de sesgo, según Gómez y Navas (1998), lo fundamental es discernir cuándo las diferencias encontradas, en habilidad o conocimiento entre los grupos, son debidas al impacto (reflejan diferencias existentes entre ellos) o al sesgo (reflejan diferencias artifactuales originadas en el proceso de medida en sí).

1.3. Funcionamiento diferencial del ítem (DIF)

De acuerdo con Clauser y Mazor (1998), el DIF se da cuando sujetos con el mismo nivel de habilidad, pero de grupos diferentes, tienen diferentes probabilidades de acertar un

ítem. El DIF representa la interacción entre la pertenencia al grupo y la probabilidad de dar una respuesta a un ítem, condicionada al atributo medido.

Un determinado ítem o test presenta DIF si se comporta diferencialmente para individuos o grupos comparables, que difieren en lengua nativa, género, etnia, cultura, o cualquier otra variable que pueda constituir una fuente sistemática de variación ajena al rasgo medido por la prueba en cuestión, entendiendo por comparables aquellos grupos de sujetos que poseen el mismo nivel en la característica o rasgo medido por el test (Gómez e Hidalgo, 1997).

En términos matemáticos, el DIF se puede definir como (Millsap y Meredith, 1992):

$$P(Y \mid W = w, V = v) \neq P(Y \mid W = w)$$

donde P denota la probabilidad, Y es una variable aleatoria observable relacionada o que pretende medir la variable aleatoria W que es no observada o latente, y V es una variable aleatoria observable que define múltiples poblaciones sujetos de acuerdo con sus valores o categorías.

En términos de la TRI, un ítem presenta DIF si a igual nivel de θ no corresponden iguales valores de P (θ) en las curvas de los grupos considerados, es decir, cuando:

$$T_{jR}(\theta) \neq T_{jF}(\theta)$$

donde:

T_{jR} es la puntuación verdadera del sujeto j, que pertenece al grupo de referencia (R) y tiene una cierta magnitud en la variable latente θ;

T_{jF} es la puntuación verdadera del sujeto j, que pertenece al grupo focal (F) y tiene una cierta magnitud en la variable latente θ.

Antes de entrar a describir los tipos de DIF, es importante detenerse en explicar y definir algunos términos. En los estudios sobre el DIF suelen compararse dos grupos, denominados como, grupo de referencia GR (grupo aventajado o mayoritario) y grupo focal GF (grupo perjudicado o minoritario), siendo normalmente el grupo de especial de interés. Los sujetos de ambos grupos se emparejan o equiparan en función de su puntuación en el atributo medido (variable o criterio de equiparación). La puntuación total del test tiende a usarse como variable de equiparación o igualación dentro de los análisis de DIF, ya que es un medio fácilmente disponible y fiable de emparejar a los examinados. Las características sociodemográficas (sexo, edad, religión, etnia, etc…) son las más empleadas para definir los grupos a comparar.

En ítems de respuesta dicotómica, la función característica del ítem, expresa la probabilidad de responder correctamente a un ítem en función del parámetro de habilidad (θ) de los sujetos y de las características o parámetros del ítem en cuestión. El gráfico de dicha función matemática se denomina Curva Característica del Ítem (CCI).

En la siguiente Figura 1.1 se muestra un ítem con DIF. Como puede verse, la CCI del ítem es diferente para el grupo focal respecto a la del grupo de referencia:

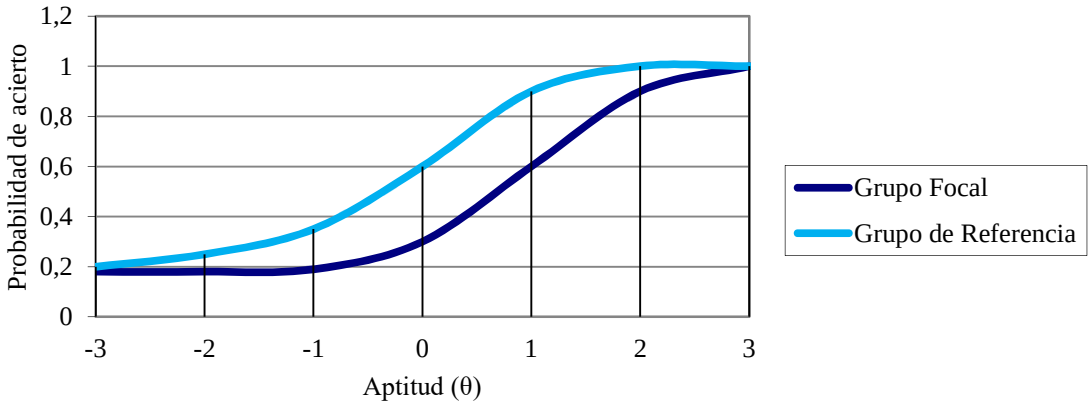


Figura 1.1

Curvas Características del Ítem (CCIs), presencia de DIF

Un ítem no presenta DIF si la curva característica del grupo focal y del grupo de referencia se solapan, como puede observarse en la Figura 1.2:

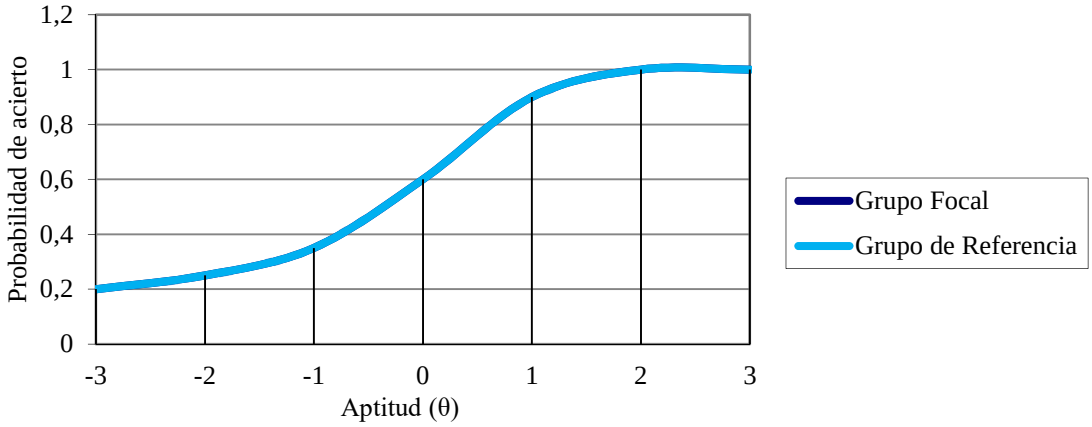


Figura 1.2

Curvas Características del Ítem (CCIs), ausencia de DIF

1.3.1. Tipos de DIF

Existen diversas clasificaciones respecto a los tipos de DIF (Hessen, 2003), aunque la clasificación de Mellenbergh (1982) es la más extendida por su simplicidad. Éste distingue dos tipos de DIF, uniforme y no uniforme, en función de la existencia o no de interacción entre el nivel en el atributo medido y el grupo de pertenencia de los sujetos.

El DIF uniforme se produce cuando la probabilidad de contestar correctamente un ítem es mayor para un grupo que para otro, a través de todos los niveles de la habilidad.

Como puede observarse en las Figuras 1.1 y 1.3, el DIF uniforme tiene lugar cuando las CCIs, de los dos grupos son diferentes, pero no se cruzan. Cuando se da el DIF uniforme existe una ventaja relativa para uno de los grupos que se mantiene constante lo largo de todo el rango se aptitud.

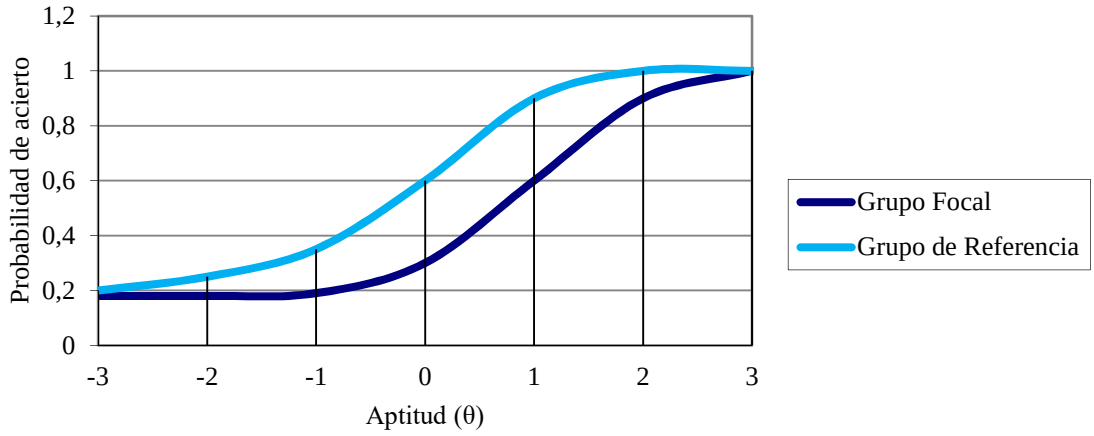


Figura 1.3
DIF uniforme

Fidalgo (1996) resume los casos en los que se puede dar el DIF uniforme, si: (a) los grupos tienen diferentes medidas en la habilidad principal (impacto), siempre que haya una correlación significativa entre la habilidad principal y la espuria (otras variables que no se pretenden medir pero que están siendo medidas y afectan a los resultados), o si (b) existen diferencias en las medidas de la habilidad espuria entre grupos.

El DIF no uniforme se produce cuando la diferencia en la probabilidad de responder correctamente un ítem, entre dos grupos, no es la misma en todos los niveles de habilidad. Como puede observarse en la Figura 1.4, el DIF no uniforme tiene lugar cuando las CCIs de los dos grupos son diferentes y además se cruzan en algún punto de la escala de θ.

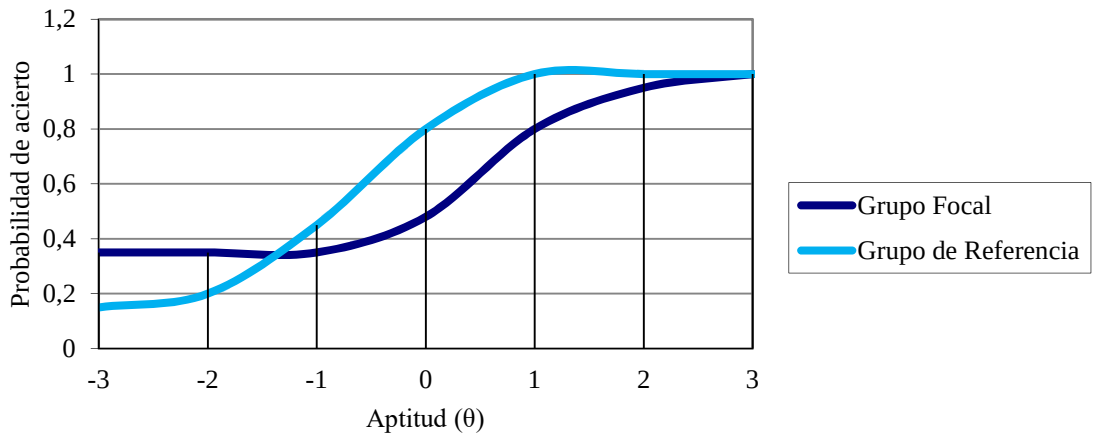


Figura 1.4
DIF no uniforme

Según Fidalgo (1996), el DIF no uniforme puede ocurrir cuando: (a) la varianza de la habilidad espuria no sea la misma entre los grupos, o (b) la magnitud de la correlación entre la habilidad principal y la espuria difieran entre los grupos.

Swaminathan y Rogers (1990) establecen una segunda subdivisión respecto al DIF no uniforme, el simétrico y el mixto.

El DIF no uniforme simétrico se da cuando el parámetro de dificultad se mantiene constante y el parámetro de discriminación varía entre los dos grupos y quedaría representado por un cruzamiento central de las CCIs en el nivel de habilidad, como puede observarse en la Figura 1.5.

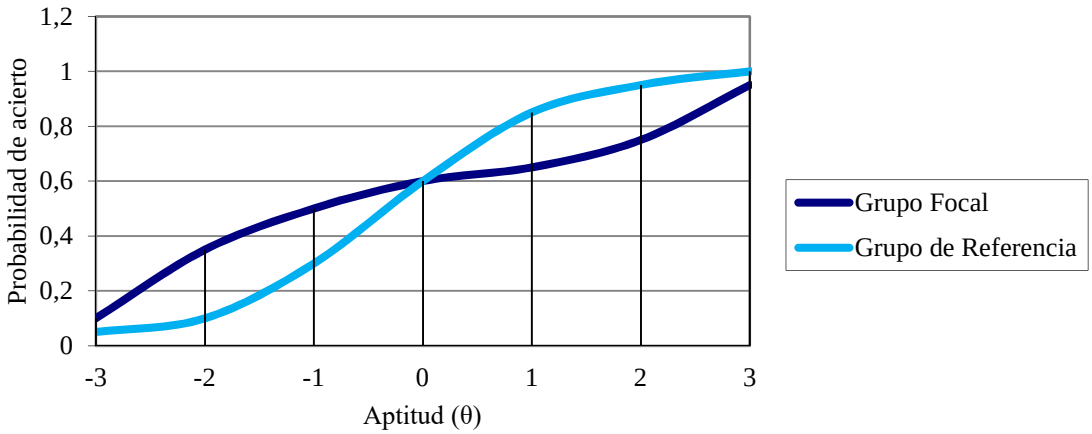


Figura 1.5
DIF no uniforme simétrico

El DIF no uniforme mixto se da cuando los parámetros de dificultad y discriminación son distintos en los dos grupos y se representa por un cruzamiento asimétrico de las CCIs del grupo focal y de referencia, como puede observarse en la Figura 1.6.

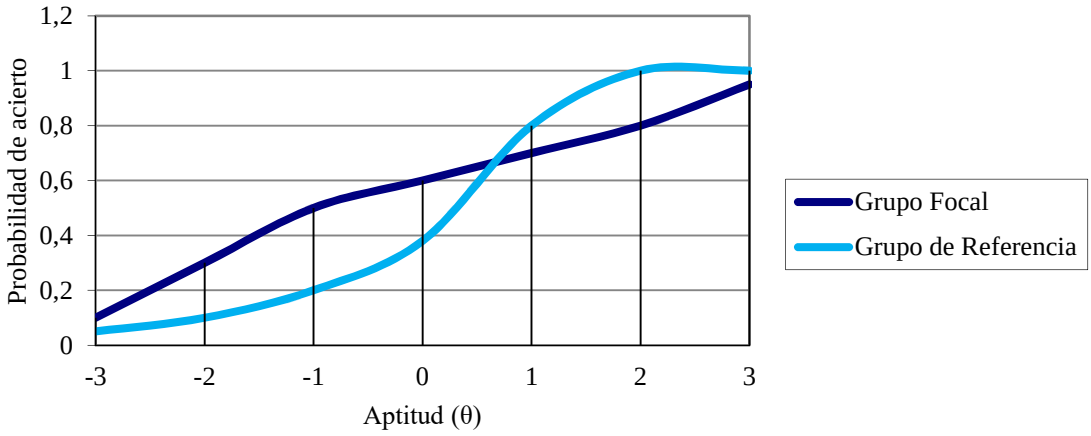


Figura 1.6

DIF no uniforme mixto

Respecto al DIF en ítems politómicos, cada vez se centra más la atención en el estudio de éstos. Para este tipo de formato de respuesta, los procedimientos aplicables son más complejos. El análisis del DIF en este tipo de ítems plantea algunas dificultades:

- Mayor número de categorías, lo que es una dificultad añadida para la comparación de las respuestas.
- Mayor dificultad en su análisis, al tener mayor rango de puntuaciones.
- La dificultad de definir una medida de rendimiento del ítem que se acomode a la medida politómica (Martínez-Arias, Hernández Lloreda y Hernández Lloreda, 2006).

1.3.2. Causas del DIF

Como se ha visto anteriormente, cualquier instrumento de medida tiene que ser objetivo en su medición. Cuando un instrumento no es objetivo y no garantiza resultados idénticos en sujetos que tienen el mismo nivel en el atributo medido, sea cual sea el grupo de pertenencia, se produce el DIF y éste puede tener múltiples causas.

En general, hay que ser cautos a la hora de interpretar las diferencias halladas entre poblaciones distintas. El método de administración, la ubicación en la prueba, la falta de homogeneidad de la población de referencia (De Ayala, Kim, Stapleton y Dayton, 1999), etc., son aspectos que deben de ser controlados con el fin de evitar su interferencia en nuestros resultados.

Aunque hay diferentes teorías que dan explicación a las causas del DIF (Ackerman, 1992; Camilli, 1992; Lord, 1980; Shealy y Stout, 1993), es la teoría multidimensional del DIF la que ofrece una explicación más consistente, aunque hay que tener en cuenta otros aspectos.

Las Directrices elaboradas por la *International Test Comission* (ITC) para la adaptación de tests (Hambleton, Yu, y Slater, 1999) ofrecen recomendaciones en este sentido: invitando a poner mayor atención en los aspectos contextuales, culturales e idiomáticos que puedan influir en el DIF.

La teoría multidimensional advierte sobre la violación del supuesto de unidimensionalidad del test, como causa principal del DIF. Pioneros de esta teoría son Ackerman (1992), Camilli (1992), Kok (1988) y Shealy y Stout (1993). Esta teoría distingue

entre la habilidad principal y las habilidades espurias, otros rasgos o habilidades que el test no pretende medir pero que interfieren en el rendimiento del mismo. No siendo la multidimensionalidad de un ítem la causa directa del DIF, sino las diferencias en las condicionales de las variables espurias. Siendo la multidimensionalidad condición necesaria pero no suficiente para que el DIF ocurra.

1.3.3. El DIF en la adaptación de cuestionarios

La adaptación de tests y cuestionarios de un idioma a otro o de una cultura a otra es una práctica antigua, surge en el campo de la psicología con la traducción de la Escala de Inteligencia para Niños de Binet-Simon de 1905, del francés al inglés en 1916, por la Universidad de Stanford.

Dado el auge y el creciente interés, en los últimos años, por la adaptación de tests y cuestionarios, y consciente de la necesidad de establecer unas directrices y una metodología clara para esta tarea, en 1992, la ITC, formada por 12 psicólogos que representaban 8 organizaciones internacionales de prestigio, abordó el desarrollo de una serie de directrices-guía para la adaptación de tests y cuestionarios.

Este trabajo (Hambleton, 1994, 1996; Muñiz y Hambleton, 1996) da origen a 22 directrices, agrupadas en cuatro apartados (Contexto, Construcción y Adaptación, Aplicación e Interpretación). En la Tabla 1.1 se presentan las 22 directrices elaboradas por la ITC agrupadas en sus cuatro apartados. Estas directrices buscan prevenir las distintas fuentes de error, propias del proceso de adaptación de tests, y ofrecen vías para controlarlas. En 1999 esas directrices se completaron y fueron presentadas en la ITC Conferencia Internacional de

Traducción y Adaptación de Tests Educativos y Psicológicos en Washington, y posteriormente publicadas (Hambleton, 2001).

Tabla 1.1

Directrices elaboradas por la ITC para la adaptación de tests (Hambleton y otros, 1999)

Contexto
➢ Los efectos de las diferencias culturales que no sean relevantes para los objetivos centrales del estudio deberían minimizarse en la medida de lo posible. ➢ Debería de evaluarse la cuantía del solapamiento de los constructos en las poblaciones de interés.

Adaptación del test

- Los constructores/editores de tests deberían de asegurar que el proceso de adaptación tiene en cuenta las diferencias lingüísticas y culturales entre las poblaciones a las que se dirigen las versiones adaptadas del test.
- Los constructores/editores de los tests deberían de proporcionar datos que garanticen que el lenguaje utilizado en las instrucciones, en los propios ítems y en el manual del test, son apropiados para todas las poblaciones culturales e idiomáticas a las que va dirigido el test.
- Los constructores/editores de tests deberían de aportar evidencia de que las técnicas de evaluación elegidas, los formatos de los ítems, las reglas de los tests, y los procedimientos son familiares a todas las poblaciones a las que van dirigidos.
- Los constructores/editores de tests deberían de facilitar evidencia de que el contenido de los ítems y los materiales de los estímulos son familiares para todas las poblaciones a las que van dirigidos.
- Los constructores/editores de tests deberían aportar una justificación racional sistemática, tanto lingüística como psicológica, para mejorar la precisión del proceso de adaptación, así como reunir datos acerca de la equivalencia de todas las versiones en los distintos idiomas.
- Los constructores/editores de tests deberían asegurar que el diseño de recogida de datos permite el uso de técnicas estadísticas apropiadas para establecer la equivalencia entre los ítems correspondientes a las diferentes versiones idiomáticas del test.
- Los constructores/editores de tests deberían aplicar técnicas estadísticas apropiadas para 1) establecer la equivalencia entre las diferentes versiones de un test, y 2) identificar componentes problemáticos o aspectos del test que puedan ser inadecuados para alguna de las poblaciones a las que va destinado el test.
- Los constructores/editores de tests deberían proporcionar información sobre la evaluación de la validez en todas las poblaciones objetivo a las que va dirigido el test adaptado.
- Los constructores/editores de tests deberían aportar datos estadísticos sobre la equivalencia de los tests para todas las poblaciones a las que van dirigidos.
- No deben utilizarse preguntas no equivalentes en todas las versiones dirigidas a diferentes poblaciones cuando se prepara una escala común, o cuando se comparan estas poblaciones. Sin embargo, pueden ser útiles para reforzar la validez de contenido de las puntuaciones de cada población por separado.

Aplicación

- Los constructores y los aplicadores de los tests deberían tratar de prever los tipos de problemas que cabe esperar, y tomar las medidas oportunas para evitarlos mediante la preparación de materiales e instrucciones adecuados.
- Quienes aplican los tests deberían de ser sensibles a cierto número de factores relacionados con los materiales utilizados para los estímulos, los procedimientos de aplicación, y las formas de respuesta, que pueden reducir la validez de las inferencias extraídas de las puntuaciones.
- Aquellos aspectos del entorno que influyen en la aplicación del test deberían de mantenerse lo más parecidos posible para todas las poblaciones a las que va dirigido el test.
- Las instrucciones para la aplicación del test en el idioma fuente y en el objetivo deben minimizar la influencia de fuentes de variación no deseadas.
- El manual del test debería de especificar todos los aspectos del test y de su aplicación que han de revisarse al utilizarlo en un nuevo contexto cultural.
- El aplicador no debe de interferir, debiendo minimizarse su influencia sobre los examinados. Deben de seguirse al pie de la letra las reglas explícitas descritas en el manual del test.

Interpretación de las conclusiones

- Cuando se adapta un test para utilizarlo en otra población, debe de facilitarse la documentación sobre los cambios, así como los datos acerca de la equivalencia entre las versiones.
- Las diferencias entre las puntuaciones obtenidas por las muestras a las que se aplicó el test no deben de tomarse sin más directamente. El investigador tiene la responsabilidad de sustanciar las diferencias con otros datos empíricos.
- Las comparaciones entre poblaciones sólo pueden hacerse al nivel de la invarianza que se haya establecido para la escala en la que se expresan las puntuaciones.
- El constructor del test debería de proporcionar información específica acerca de las distintas formas en las que los contextos socioculturales y ecológicos de las poblaciones pueden afectar al rendimiento en el test, y debería sugerir procedimientos para tener en cuenta estos efectos en la interpretación de los resultados.

La década de los 90, fue testigo de una enorme cantidad de investigación dirigida hacia mitigar el efecto de las fuentes de error propias del proceso de adaptación de tests. Convirtiéndose, el DIF, en parte integral de la validación de pruebas y en la adaptación de tests y cuestionarios, con el fin de asegurar la equivalencia entre versiones.

Durante estos años se incrementa el número de adaptaciones de tests y cuestionarios y se amplían a otras disciplinas, no solo la psicología. Su incremento fue, y es, el reflejo de un medio social marcado por los contactos entre culturas e idiomas y en el que los tests y cuestionarios asisten diariamente en los ámbitos educativo, social, jurídico o clínico, entre otros, en la toma de decisiones individuales o grupales (Muñiz y Hambleton, 1996).

En 1999, tres entidades: la *American Educational Research Association*, la *American Psychological Association* y el *National Council on Measurement in Education*, se unen, en colaboración, para publicar los *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association y National Council on Measurement in Education, 1999). Los *Standards* de 1999 señalan que existe consenso respecto a que las pruebas deben de estar libres de sesgo y comienzan a usarse con frecuencia los términos ítem sesgado y DIF.

Respecto al DIF, los *Standards* indican que el DIF existe cuando los examinados de igual capacidad difieren, de acuerdo con la pertenencia de sus grupos a sus respuestas para un particular ítem.

En ese momento, el DIF ya es un componente clave en los estudios de validez, en prácticamente todas las evaluaciones a gran escala. Éste se han integrado en los estudios de validez y en los *Standards* de 1999, Norma 7.3.

Las directrices de la ITC, como no puede ser de otra manera, se adaptan a las nuevas necesidades, y la ITC consciente de la necesidad de ir mejorando la calidad del proceso de traducción adaptación de tests, presenta en 2013 las nuevas directrices (Muñiz, Elosua y

Hambleton, 2013), ver en Tabla 1.2. Estas directrices constituyen una actualización y reorganización de las publicadas originalmente (Hambleton, 1996; Hambleton, Merenda y Spielberger,, 2005; Muñiz y Hambleton, 1996).

Tabla 1.2

Directrices elaboradas por la ITC para la adaptación de tests (Muñiz, Elosua y Hambleton, 2013)

Directrices previas
➢ Antes de comenzar con la adaptación hay que obtener los permisos pertinentes de quien ostente los derechos de propiedad intelectual del test. ➢ Cumplir con las leyes y prácticas profesionales relativas al uso de tests que estén vigentes en el país o países implicados. ➢ Seleccionar el diseño de adaptación de tests más adecuado. ➢ Evaluar la relevancia del constructo o constructos medidos por el test en las poblaciones de interés. ➢ Evaluar la influencia de cualquier diferencia cultural o lingüística en las poblaciones de interés que sea relevante para el test a adaptar.
Directrices de desarrollo
➢ Asegurarse, mediante la selección de expertos cualificados, de que el proceso de adaptación tiene en cuenta las diferencias lingüísticas, psicológicas y culturales entre las poblaciones de interés. ➢ Utilizar diseños y procedimientos racionales apropiados para asegurar la adecuación de la adaptación del test a la población a la que va dirigido. ➢ Ofrecer información y evidencias que garanticen que las instrucciones del test y el contenido de los ítems tienen un significado similar en todas las poblaciones a las que va dirigido el test. ➢ Ofrecer información y evidencias que garanticen que el formato de los ítems, las escalas de respuesta, las reglas de corrección, las convenciones utilizadas, las formas de aplicación y demás aspectos son adecuados para todas las poblaciones de interés. ➢ Recoger datos mediante estudios piloto sobre el test adaptado, y efectuar análisis de ítems y estudios de fiabilidad y validación que sirvan de base para llevar a cabo las revisiones necesarias y adoptar decisiones sobre la validez del test adaptado.

Directrices de confirmación

- Definir las características de la muestra que sean pertinentes para el uso del test, y seleccionar un tamaño de muestra suficiente que sea adecuado para las exigencias de los análisis empíricos.
- Ofrecer información empírica pertinente sobre la equivalencia del constructo, equivalencia del método y equivalencia entre los ítems en todas las poblaciones implicadas.
- Recoger información y evidencias sobre la fiabilidad y la validez de la versión adaptada del test en las poblaciones implicadas.
- Establecer el nivel de comparabilidad entre las puntuaciones de distintas poblaciones por medio de análisis de datos o diseños de equiparación adecuados.

Directrices sobre la aplicación

- Preparar los materiales y las instrucciones para la aplicación de modo que minimicen cualquier diferencia cultural y lingüística que pueda ser debida a los procedimientos de aplicación y a los formatos de respuesta, y que puedan afectar a la validez de las inferencias derivadas de las puntuaciones.
- Especificar las condiciones de aplicación del test que deben seguirse en todas las poblaciones a las que va dirigido.

Directrices sobre puntuación e interpretación

- Interpretar las diferencias de las puntuaciones entre los grupos teniendo en cuenta la información demográfica pertinente.
- Comparar las puntuaciones entre poblaciones únicamente en el nivel de invarianza establecida para la escala de puntuación utilizada en las comparaciones.

Directrices sobre la documentación

- Proporcionar documentación técnica que recoja cualquier cambio en el test adaptado, incluyendo la información y las evidencias sobre la equivalencia entre las versiones adaptadas.
- Proporcionar documentación a los usuarios con el fi n de garantizar un uso correcto del test adaptado en la población a la que va dirigido.

En la actualidad, son los estudios transculturales y transidiomáticos, los que han cobrado especial relevancia y reclaman, en mayor medida, la adaptación de tests y cuestionarios. Como señalan Muñiz, Elosua y Hambleton (2013), existe cada vez más demanda, por parte de las corporaciones multinacionales y organismos internacionales, de disponer de pruebas de acreditación o de selección que puedan utilizarse en distintos países o en distintos idiomas. En este sentido, cabe mencionar el estudio PISA (Programme for International Student Assessment), para medir el rendimiento académico a nivel mundial y el estudio TIMMS (Trends in International Mathematics and Science Study) para evaluar internacionalmente ciertos conocimientos en estudiantes, empleando, ambos, pruebas adaptadas en más de cuarenta idiomas, siendo sus resultados de gran impacto social.

Como puede verse, existen diversos motivos para la adaptación de test y cuestionarios. Entre todos, encontramos motivos económicos, dado que la adaptación es más rápida, práctica y económica que construir un nuevo test. Pero también existen motivos sociales, dado que la adaptación de un cuestionario nos permite realizar estudios comparativos entre países, y también técnicos, con el fin de mejorar la imparcialidad de la evaluación (Ribeiro, Gómez-Conesa, e Hidalgo, 2010, p.256):

> Permitiendo que las personas puedan utilizar los tests en el idioma que tenga mayor dominio, eliminando el sesgo de las puntuaciones del examen asociado a la obligación de tener que realizarlo en la segunda o tercera lengua de un individuo, potenciando así su validez.

1.4. Procedimientos de análisis del DIF

Se han desarrollado numerosos procedimientos estadísticos para evaluar el DIF, principalmente desde finales de los años 80. El principal objetivo de estos procedimientos es la detección del DIF uniforme y no uniforme, evitando confundir el DIF con el impacto.

Ante la creciente utilización de formatos de respuesta politómica y la necesidad de detectar con precisión el DIF en este tipo de ítems, sin recurrir a la dicotomización, con la consiguiente pérdida de información, las investigaciones recientes se encaminan a desarrollar procedimientos de evaluación adecuados para ítems con formato de respuesta politómica.

Algunas de las técnicas evalúan la presencia de DIF en ítems politómicos; y otras han sido modificadas, con el fin de poder ser empleadas con ítems de respuesta politómica. En ambos casos, hablamos de técnicas complejas, ya que deben de tener en cuenta que el DIF puede existir para cada categoría de respuesta del ítem.

En este apartado se presenta una aproximación a dichos procedimientos, tanto para ítems dicotómicos como politómicos.

1.4.1. Clasificación de los procedimientos de detección del DIF

Potenza y Dorans (1995), teniendo en cuenta la distinción sobre la naturaleza de los ítems (dicotómicos/politómicos), clasifican los diferentes métodos en función del tipo de criterio de igualación de los grupos (puntuación observada/variable latente) y de la relación entre la puntuación del ítem y la variable de igualación (paramétrica/no paramétrica).

Basándose en esta taxonomía, Hidalgo y Gómez-Benito (2010) ofrecen una clasificación de todos los procedimientos actuales de detección del DIF.

- Teniendo en cuenta la naturaleza del tipo de respuesta (dicotómicos/politómicos):

    - En el caso de *ítems politómicos* hay que tener en cuenta que el DIF puede estar presente en las diferentes categorías de respuesta del ítem, y no necesariamente en la misma dirección ni en todas las categorías.

    - En el caso de *ítems dicotómicos*, hablamos de técnicas más sencillas conceptual y computacionalmente.

- Teniendo en cuenta el tipo de criterio de igualación de los grupos (puntuación observada/variable latente):

    - El método que emplea *la variable latente* como criterio de igualación de los grupos emplea una estimación de la habilidad latente, tal y como explica la teoría de respuesta al ítem (TRI).

    - El método que emplea la *puntuación observada* utiliza la puntuación total observada del test.

- Teniendo en cuenta la relación entre la puntuación del ítem y la variable de igualación (paramétrica/no paramétrica):

- Métodos *paramétricos*, utilizan una función matemática que relaciona la puntuación del ítem con el nivel de habilidad, que representan gráficamente la probabilidad de obtener una determinada puntuación en el ítem en función del nivel de habilidad de los individuos como las CCIs de la Figura 1.1. El DIF serían las diferencias en las CCIs de los grupos y para que esto ocurra los parámetros que definen las correspondientes CCIs han de ser diferentes.

- Métodos *no paramétricos*, no utilizan ninguna función matemática para relacionar la respuesta al ítem con el nivel de habilidad, sino que simplemente tienen en consideración la puntuación observada al ítem en cada uno de los niveles de habilidad para cada grupo. En este caso, la presencia de DIF está determinada por la obtención de diferencias entre grupos en la puntuación observada.

Según la propuesta de clasificación que tiene en cuenta la naturaleza de los ítems, el tipo de criterio de igualación de los grupos y la relación entre la puntuación del ítem y la variable de igualación, se presenta en la Tabla 1.3 una clasificación sobre los principales métodos de detección del DIF.

Tabla 1.3

Cuadro resumen de los principales métodos de detección del DIF (Guilera, 2009)

Variables de igualación	**Tipos de relación**	
	Paramétrica	**No paramétrica**
	DICOTÓMICA	
Puntuación observada	Regresión Logística Modelos log-lineales Modelos de clase latente	Mantel-Haenszel Estandarización
Variable latente	IRT log-likelihood ratio χ^2 de Lord Medidas del área IRT-Differential Item Funcioning Test (DIFT) Análisis Factorial Confirmatorio Múltiple indicador múltiple-causa (MIMIC)	SIBTEST CATSIB
	POLITÓMICA	
Puntuación observada	Regresión Logística Multinomial Regresión Logística Discriminante Modelos log-lineales Modelos de clase latente	Mantel-Haenszel Generalizado Mantel Estandarización
Variable latente	IRT Log-likelihood ratio χ^2 de Lord Medidas del área IRT- Differental Item Functioning Test (DIFT) Análisis Factorial Confirmatorio Múltiple indicador múltiple-causa (MIMIC)	POLYSISBTEST

A continuación serán explicadas, resumidamente, las técnicas más populares en el estudio del DIF, describiendo en profundidad la RL y sus variaciones en el próximo capítulo.

1.4.1.1. Mantel-Haenszel (MH)

El procedimiento de Mantel-Haenszel (MH) (Mantel y Haenszel, 1959) es uno de los métodos más empleados para detectar DIF, por su sencillez conceptual y facilidad de cálculo (Guilera, Gómez-Benito e Hidalgo, 2009; Sireci y Rios, 2013).

Este estadístico fue aplicado, al estudio de DIF, por Holland y Thayer (1988), con ítems dicotómicos. Compara la ejecución de un ítem entre el grupo de referencia (R) y el grupo focal (F) a través de *k* niveles de un determinado criterio (generalmente la puntuación total del test); se asume que en cada nivel los sujetos de uno y otro grupo son comparables. Si lo ejecutan por igual los sujetos de uno u otro grupo, quiere decir que el ítem no presenta DIF.

El modo de operar para el estadístico MH, comienza con las respuestas de los examinados al ítem, la puntuación de los mismos en el test y su grupo de pertenencia.

Así, lo primero es elaborar *k* tablas de contingencia de 2 x 2, siendo *k* el número de intervalos en los que se divide la puntuación en el test, como se puede observar en la Tabla 1.4.

Tabla 1.4

Tabla de contingencia 2 x 2 para el nivel de puntuación k

Grupo	**1**	**0**	**Total**
G_R	A_k	B_k	N_{Rk}
G_F	C_k	D_k	N_{Fk}
	GN_{1k}	N_{0k}	N_k

Los valores de las celdas A_k, B_k, C_k y D_k indican el número de examinados en cada categoría. Los valores marginales N_{Rk} y N_{Fk} representan el número de examinados en el grupo de referencia y focal, respectivamente. N_{1k} y N_{0k} representa el número de examinados que ha respondido correcta e incorrectamente el ítem, respectivamente. Y N_k es el número total de examinados en el nivel de puntuación k.

El estadístico de MH viene dado por:

$$\chi^2_{MH} = = \frac{(|\sum_{k=1}^{m} A_k - \sum_{k=1}^{m} E(A_k)| - 0.5)^2}{\sum_{k=1}^{m} Var\,(A_k)}$$

donde E(A_k) es el valor esperado de A_k y Var (A_k) es su varianza, siendo iguales a:

$$\mathrm{E}\,(A_k) = (N_{Rk}N_{1k})/N_k$$

$$\mathrm{V}\,(\mathrm{A_k}) = \frac{\mathrm{N_{Rk}N_{Fk}N_{1k}N_{0k}}}{\mathrm{N_k^2(N_k-1)}}$$

El estadístico χ^2_{MH} sigue una distribución χ^2 con un grado de libertad.

El ítem tiene DIF al nivel de significación α[1], cuando $\chi^2_{MH} \geq \chi^2_{1-\alpha}$. Además Mantel y Haenszel (1959) proporcionan un estimador de α, *common odds ratio*, $\hat{\alpha}_{MH}$, dado por:

$$\hat{\alpha}_{MH} = \frac{\sum_{k=1}^{m} A_k D_k / N_k}{\sum_{k=1}^{m} B_k C_k / N_k}$$

$\hat{\alpha}_{MH}$ es un estimador del DIF y sus valores van de 0 a ∞.

La hipótesis nula de ausencia de DIF vendría representada por un valor de 1, si el valor es mayor de 1, indica que el ítem está favoreciendo al grupo de referencia, si este valor es menor que 1, indica que el ítem está favoreciendo al grupo focal.

La principal limitación que tiene este procedimiento es la incapacidad para detectar el DIF no uniforme (Rogers y Swaminathan, 1993), aunque esto es corregido por Mazor, Clauser y Hambleton (1994). Este procedimiento supone calcular dos estadísticos $\chi 2$ de forma separada, uno para el grupo de sujetos con las puntuaciones más bajas (los sujetos con una puntuación total menor o igual a la media de la distribución de las puntuaciones en el test) y en el grupo de sujetos con las mayores puntuaciones (sujetos con una puntuación total mayor que la media de la distribución de las puntuaciones en el test).

Del procedimiento MH se han propuesto varios estadísticos para ítems politómicos: por una parte, el estadístico MH generalizado (Mantel y Haenszel, 1959; Zwick, Donoghue y Grima, 1993) que considera las categorías de respuesta del ítem como una variable nominal (especificando la hipótesis alternativa de DIF que la distribución de las respuestas al ítem difiere entre los grupos); de otra parte, el estadístico de Mantel (Mantel, 1963; Zwick,

[1] El α de la medida del tamaño del efecto de Mantel-Haenszel, no es igual al α de la significación estadística.

Donoghue y Grima, 1993) que considera la naturaleza ordenada de las categorías de respuesta del ítem politómico (especificando la hipótesis alternativa de DIF que la media de las puntuaciones correspondientes a las categorías de respuesta difiere entre los grupos).

1.4.1.2. Método de estandarización

Este método fue propuesto por Dorans y Kulick (1983, 1986) y se basa en una aproximación empírica, que utiliza la puntuación total del test como variable de emparejamiento. Se analiza el posible DIF mediante tablas de contingencia que recogen las respuestas a un ítem en un determinado intervalo de habilidad en función del grupo de pertenencia.

Este procedimiento se basa en la diferencia en la proporción de sujetos que aciertan el ítem en el grupo focal y en el grupo de referencia en cada nivel de la puntuación.

Este procedimiento ofrece un índice para cuantificar el DIF, la diferencia en proporciones estandarizadas (DPE):

$$DPE: \frac{\sum_{k=1}^{m} W_k (P_{Fk} - P_{Rk})}{\sum_{k=1}^{m} W_k}$$

donde, W_k es el factor de ponderación en el nivel de puntuación *k*, que se emplea para ponderar las diferencias en la proporción de respuestas correctas entre el grupo focal y el de referencia. P_{Fk} y P_{Rk} son las proporciones de examinados que responden correctamente al ítem en el nivel de puntuación *k* en el grupo focal y de referencia, siendo:

$$P_{Fk}=C_k/N_{Fk} \quad \text{y} \quad P_{Rk}= A_k/N_{Rk}$$

El índice DPE varía entre -1 y 1. Si los valores son positivos, el ítem favorece al grupo focal, y si son negativos favorece al grupo de referencia.

Dorans y Holland (1993) proponen una serie de valores para su interpretación:

- Ausencia de DIF: valores entre -.05 y .05.
- Revisión de ítems aconsejada: valores entre -.10 y .05 y entre .05 y .10.
- Ítems altamente sospechosos: valores del rango -.10, .10.

Respecto al método de estandarización, se han propuesto diversas extensiones para el análisis del DIF en ítems politómicos (Dorans, Schmitt y Bleinstein, 1992).

1.4.1.3. Procedimiento SIBTEST

SIBTEST, desarrollado por Shealy y Stout (1993), es uno de los procedimientos para detectar tanto el DIF, como el funcionamiento diferencial del test (FDT); simultáneamente, también, el DIF presente en uno o más ítems del test, tal y como lo indica su nombre (SIB significa simultaneous item bias).

Comprobar la presencia o ausencia de DIF en varios ítems a la vez presenta varias ventajas, pero la principal es la posibilidad de comprobar si se da FDT. Si los ítems con DIF actúan juntamente afectando a las puntuaciones en el test, de manera diferente según que los examinados pertenezcan a un grupo u otro, se da el fenómeno de amplificación. El fenómeno de cancelación, por el contrario, se da cuando los efectos del DIF presentes en varios ítems se anulan unos a otros no llegando a provocar FDT.

Una limitación del SISTEST es su incapacidad para detectar el DIF/FDT no uniforme, ya que ha sido diseñado para detectar el DIF/FDT uniforme. Además, sólo puede aplicarse a tests de más de 25 ítems, que no presenten impacto. Por ello, Li y Stout (1996) proponen el procedimiento Crossing SIBTEST, que permite evaluar el FDT/DIF no uniforme en ítems dicotómicos. El procedimiento propuesto calcula primero la puntuación del subtest, en el que se produce el cruce de las CCIs (en términos de la TRI), usando el análisis de regresión por mínimos cuadrados.

Al igual que el resto de procedimientos, SIBTEST tuvo su versión para ítems con respuestas ordinales. Poly-SIBTEST (Chang, Mazzeo, y Roussos, 1996) permite evaluar el FDT/DIF en ítems politómicos; éste se considera un procedimiento no paramétrico porque requiere pocos supuestos. Se supone un subconjunto de ítems, unidimensionales y libres de DIF, y los grupos referencia y focal, independientes, pero no se asume ninguna distribución de la θ.

1.4.1.4. Análisis de regresión logística

Ver capítulo 2: Regresión Logística.

1.4.1.5. Procedimientos de análisis basados en la Teoría de Respuesta al Ítem (TRI)

Estos procedimientos establecen un modelo de medida que relaciona las respuestas al ítem con la variable latente que pretende medir el test. Hay procedimientos de TRI que, comparan las CCIs, basándose en cálculo del área entre éstas, otros comparan los parámetros

estimados de los ítems, en dos o más grupos (χ^2 de Lord) y otros comprueban si los parámetros de las CCIs difieren en un mismo ítem, administrado a grupos distintos, mediante la comparación de modelos (IRTLR).

1.4.1.5.1. Medidas de Área

Si recordamos, las CCIs de un ítem con DIF son diferentes para el grupo focal y para el grupo de referencia. Teniendo en cuenta esto, el DIF es cuantificado, mediante las medidas de área, en función del área existente entre las CCIs de los grupos comparados, para determinado intervalo de θ $[\theta_1, \theta_2]$, como puede observarse gráficamente en la Figura 1.7.

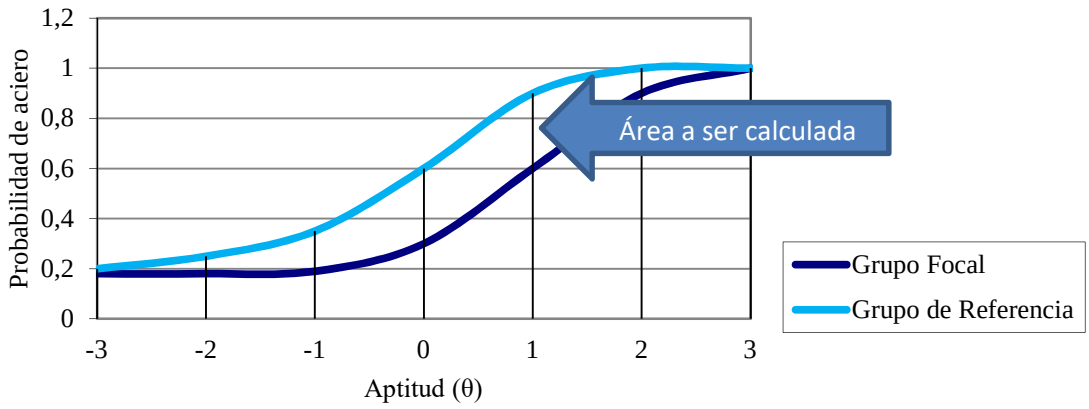


Figura 1.7
CCI de un ítem y el área que las separa

Existen diversas propuestas para las medidas del área, que difieren según sean (Fidalgo, 1996, p. 416):

- Con signo o sin signo.
- El intervalo de θ sea finito o infinito.

- La aproximación sea continua (integrando) o discreta (utilizando el sumatorio).
- Se ponderen o no las diferencias entre las probabilidades.

Rudner (1977) y Rudner, Getson y Knight (1980a, 1980b) proponen una medida de área sin signo, basada en la comparación de las CCIs teóricas según los parámetros estimados en cada grupo por separado. Rudner, Getson y Knight (1980b) proponen la siguiente fórmula para el cálculo del área comprendida entre ambas CCIs:

$$A = \sum_{\theta=-4}^{\theta=+4} |P_R(\theta_j) - P_F(\theta_j)\,|\Delta\theta$$

donde:

$P_R(\theta_j)$ es la probabilidad de acertar el ítem del grupo de referencia, dado θ_j;

$P_F(\theta_j)$ es la probabilidad de acertar el ítem del grupo focal, dado θ_j;

$\Delta\theta$ es el valor de la base de un rectángulo ($\Delta\theta$=0,005) y altura $|P_R(\theta_j - P_F(\theta_j)|$.

En este procedimiento las áreas son calculadas para los distintos valores de θ, comprendidos en el intervalo -4 hasta +4, con el incremento $\Delta\theta$.

Linn, Levine, Hastings y Wardrop (1981) proponen otro procedimiento, derivado del anterior:

$$A = \sum_{\theta=-3}^{\theta=+3} \sqrt{[P_R(\theta_j) - P_F(\theta_j)]^2 \Delta\theta}$$

La única diferencia, respecto a la propuesta de Rudner, Getson y Knight (1980), es el valor del intervalo adoptado para las distintas magnitudes de θ que, en este caso, está comprendida entre -3 y +3.

Linn, et al. (1981) propone cuatro índices de DIF:

a) Área base superior: es el área entre la CCI del grupo R y la CCI del grupo F, rango de (-3, +3), en aquellos intervalos en los que la CCI del grupo R está por encima de la CCI del grupo F. Esta medida proporciona información sobre la dirección del DIF.

b) Área base inferior: es el área entre la CCI del grupo de referencia y la CCI del grupo focal, en aquellos intervalos en los que la CCI del grupo F está por encima de la CCI del grupo R. Esta medida proporciona información sobre la dirección del DIF.

c) Área total: se obtiene sumando las dos medidas de área anteriores, y esta medida proporciona información sobre la cantidad de DIF.

d) RDMC o raíz cuadrada de la diferencia media cuadrática: esta medida proporciona información sobre la cantidad de DIF.

Raju (1988, 1990), asumiendo que la θ tiene una naturaleza continua, desarrolla un conjunto de medidas de área exactas con signo (ACS) y sin signo (ASS), acompañadas de una

prueba de significación que permite probar, a un nivel de confianza establecido, si el área entre dos CCIs es significativamente diferente de cero.

Uno de los procedimientos propuestos por Raju (1988) es el siguiente:

$$A = (1 - C)|\frac{2(a_2 - a_1)}{D\ a_1 a_2}\ln\left[1 + e\ \frac{D\ a_1 a_2 (b_2 - b_1)}{a_2 - a_1}\right] - (b_2 - b_1)|$$

donde:

a_1 y a_2 son los parámetros de discriminación del grupo de referencia y focal;

b_1 y b_2 son los parámetros de dificultad del grupo de referencia y focal;

c es la probabilidad de acierto al azar;

D es una constante de valor 1.7;

e es el número 2.7182, base de los logaritmos neperianos.

Para el uso de esta fórmula se asume que el valor del parámetro *c* es el mismo para los grupos analizados. Cuando, además, el parámetro *a* también tiene el mismo valor para los dos grupos, según Swaminathan y Rogers (1990), la fórmula se reduce a:

$$A = (1 - c)|b_2 - b_1|$$

Kim y Cohen (1991), asumiendo que la θ tiene una naturaleza continua, proponen las medidas del área de intervalos cerrados con signo (ACCS) y sin signo (ACSS).

1.4.1.5.2. Medidas basadas en la comparación de parámetros (χ^2 de Lord)

Para las medidas basadas en la comparación de parámetros, lo que se compara son los parámetros del ítem que definen las CCIs, no el área existente entre las curvas, como en el procedimiento anterior.

Bajo esta premisa, Lord (1980) propone estimar conjuntamente (para ambos grupos) los parámetros de los ítems, estandarizando sobre el parámetro de dificultad, fijando los parámetros de pseudoazar a los valores obtenidos y estimando los parámetros de dificultad y discriminación, para cada grupo por separado. Su formulación matemática es:

$$\chi^2 = V \sum^{-1} V'$$

donde:

χ^2 tiene dos grados de libertad;

V es el vector de dimensión (1 x 2) de las diferencias entre los parámetros a y b de los grupos de referencia y focal;

V' es el vector traspuesto de V;

Σ^{-1} es la inversa de la matriz suma de varianzas-covarianzas de V para los grupos de referencia y focal, cuya dimensión es 2 x 2.

En el caso de aplicación al modelo logístico de un parámetro, su formulación matemática es:

$$\chi^2 = \frac{b_F - b_R}{Var\,(b_F) - Var\,(b_R)}$$

donde:

b_F y b_R son los valores de los parámetros b en cada grupo;

$Var\ (b_F)$ y $Var\ (b_R)$ las varianzas estimadas de dichos parámetros.

Por lo que no existiría DIF si los parámetros del ítem estimados, en cada grupo, coincidieran. Este estadístico se calcula fácilmente y proporciona una prueba de significación; sin embargo presenta algunas deficiencias, tal y como indican Gómez e Hidalgo (1997): en primer lugar, la estimación de la θ se realiza sobre todos los ítems del test, de tal modo que se incluyen también los ítems con DIF y, en segundo lugar, se asume que θ es conocida, sin embargo los valores de θ son estimados.

1.4.1.5.3. Medidas basadas en la comparación de modelos (IRTLR)

Thissen, Steinberg y Gerrard (1986) proponen un procedimiento que comprueba si los parámetros de las CCIs de un ítem, al ser administrado a grupos distintos, difieren.

En este procedimiento de análisis, basado en la teoría de respuesta al ítem, la hipótesis nula indica que no hay diferencia entre los parámetros de los ítems entre los grupos; esto se prueba mediante una estrategia de comparación de modelos. Estos modelos son:

- El modelo aumentado (A), en el que se especifica que los parámetros de los ítems no son los mismos para los dos grupos.

- El modelo compacto (C), que establece la restricción de igualdad de los parámetros de los ítems en los dos grupos.

Dado que el modelo compacto (C) está anidado dentro del modelo aumentado (A), se calcula la bondad de ajuste G^2 para cada modelo. La prueba de significación de la hipótesis nula se obtiene comparando el estadístico de ajuste del modelo C y del modelo A de la siguiente manera:

$$G^2 = -2LL_c - (-2LL_A)$$

Que sigue una distribución central de χ^2 con 1 grados de libertad (*gl*), donde 1 es la diferencia entre el número de parámetros de ítem estimados en el modelo C y el modelo A, respectivamente. Si el valor obtenido es mayor que el valor teórico de la distribución χ^2 con l *gl*, entonces rechazamos la hipótesis nula y, por implicación, el modelo C, concluyéndonos que el ítem o ítems especificados muestran DIF.

Tal y como indican Gómez-Benito, Hidalgo y Guilera (2010), las técnicas que emplean la puntuación observada en el test como variable de igualación pueden resultar imprecisas en la detección del DIF principalmente cuando el test contiene ítems de discriminación diversa, mientras que los métodos de la variable latente superan este inconveniente a base de incrementar la sofisticación de los modelos matemáticos de estimación de la habilidad. Respecto a los métodos no paramétricos, las citadas autoras indican que una ventaja de éstos es que los supuestos del modelo son escasos, por lo que el DIF no suele confundirse con la falta de ajuste del modelo; en el caso de los métodos paramétricos es necesario asegurar una adecuada estimación de los parámetros del test precisamente para evitar esta confusión, por tanto, se requieren tamaños muestrales del grupo de referencia y focal mucho más elevados que con los modelos no paramétricos.

Hidalgo y López-Pina (2000), desde el punto de vista teórico-estadístico, nos dan dos pautas para elegir la técnica más adecuada. En primer lugar, debe de proporcionar una medida del DIF en términos claros y sencillos. Y en segundo lugar, debe de disponer de un estimador adecuado para dicha medida, que sea potente y eficiente.

1.5. Métodos de purificación

Hay que tener en cuenta, a la hora de elegir el criterio de igualación para el análisis del DIF, que éste esté libre de sesgo. Es frecuente que en la mayoría de situaciones la única evidencia empírica de equiparación, o igualación, que se dispone es el propio test o subtests (generalmente la puntuación total del test o de los subtests), y ésta se encuentra contaminada por la presencia de ítems con DIF, a la vez que forman parte del criterio, junto a los ítems sin DIF. Gómez, Hidalgo y Guilera (2010) definen este problema como un problema endémico a los métodos de detección del DIF, que reside en que sufre de una cierta circularidad en su forma de proceder ya que el ítem estudiado también contribuye a la definición de la variable de igualación de los grupos.

Con el fin de reducir el efecto producido por los ítems con DIF, se han propuesto algunas técnicas de purificación que, en dos etapas o iterativamente, eliminan del criterio aquellos ítems que previamente han sido detectados con DIF (French y Maller, 2007, Gómez-Benito y Navas, 1996; Hidalgo y Gómez-Benito, 2003; Holland y Thayer, 1988; Navas-Ara y Gómez-Benito, 2002; Wang, Shih y Yang, 2009).

Como ejemplo de algunas técnicas, están Holland y Thayer (1988) que emplean el método de purificación bietápica para el estadístico MH, Gómez y Navas (1996) que aplican

la purificación paso a paso a la RL dicotómica, e Hidalgo y Gómez (2003) con la RL politómica.

1.6. Medidas del tamaño del efecto

Tal y como resaltan Gómez-Benito, Hidalgo y Guilera (2010, p. 79) "Hay que tener en cuenta que detectar un ítem con DIF mediante una prueba de significación estadística no necesariamente implica que su efecto sea destacable, es decir, puede que su efecto sea de escasa relevancia. En este sentido, es importante examinar la magnitud del DIF porque los efectos de la presencia de ítems con DIF pueden ser triviales, cancelarse o pueden realmente poner en duda las decisiones basadas en el test".

Cada vez más, nos encontramos con trabajos que recomiendan el uso de medidas del tamaño del efecto como complemento a las pruebas de significación, como por ejemplo el trabajo de Monahan, McHorney, Stump y Perkins (2007), con el objetivo de evaluar la magnitud del efecto observado.

En otras palabras, el uso de las medidas del tamaño del efecto es recomendable:

- Para no dejarnos llevar por la influencia que pueden ejercer los diferentes tamaños muestrales, en los resultados. Como es en el caso de tamaños muestrales grandes, que pueden indicar hallazgos estadísticamente significativos, pero con un tamaño del efecto muy pequeño e insignificante (Kirk, 1996).

- Para poder ir alejándonos de los estándares, algo arbitrarios, establecidos por Cohen (1992). Esto se conseguirá, según Zumbo y Hubley (1998), informando a la comunidad psicométrica de los tamaños de los efectos encontrados, ya sean, estadísticamente significativos o no significativos, consiguiendo así un archivo de efectos.

Por todo ello, las medidas del tamaño del efecto cobran gran importancia en los estudios psicométricos.

Existen diversas medidas del efecto, en función de la técnica de detección del DIF que se emplee. Para el procedimiento MH, Dorans y Holland (1993) proponen el estadístico Delta-DIF como medida del efecto. Para la RL, Zumbo y Thomas (1997) sugieren el incremento en R^2; Gómez-Benito e Hidalgo (2007) y Monahan, et al. (2007) han propuesto el uso de la odds-ratio como medida del tamaño del efecto, para ítems dicotómicos e Hidalgo, Gómez-Benito y Zumbo (2008) para ítems politómicos.

1.7. Programas para detectar el DIF

Existe una gran variedad de programas informáticos que permiten implementar la mayoría de procedimientos de detección del DIF.

Entre ellos, los programas que han sido diseñados específicamente para la detección de los ítems con DIF son: DIF (Klieme y Strumpf, 1991), IRTDIF (Kim y Cohen, 1992), DICHODIF software (Rogers, Swaminathan y Hambleton, 1993), SIBTEST (Li y Stout, 1994), MHDIF (Fidalgo, 1994), MH (Rogers y Hambleton, 1994), DFITPU (Raju, 1995),

SIBTEST (Stout y Roussos, 1995), EZDIF (Waller, 1998a), LINKDIF (Waller, 1998b), DFITP5 (Raju, 1999), DIF/DBF (Stout y Roussos, 1999), TESTGRAPH (Ramsay, 2000), IRTLRDIF (Thissen, 2001, 2003), STDIF (Robin, 2001), RLDIF (Gómez, Hidalgo, Padilla, y González, 2005), DIFAS (Penfield, 2005), XS-DIF (Ordóñez y Romero, 2007) y EASY-DIF (González, Padilla, Hidalgo, Gómez-Benito y Benítez, 2011).

Pueden utilizarse también programas estándares de análisis estadístico, como por ejemplo SPSS (IBM Corp. Released 2013) para MH y RL, LISREL (Jöreskog y Sörbom, 2006) o MPLUS (Muthén y Muthén, 2007) para procedimientos basados en modelos de ecuaciones estructurales.

Magis, Béland, Tuerlinck y De Boeck (2010), presentan un nuevo paquete para el software R (R Development Core Team, 2008), denominado difR (versión 2.2), que puede realizar varios procedimientos de detección DIF tradicionales para los ítems dicotómicos (el usuario puede elegir entre varios métodos basados en IRT o no basados en IRT).

CAPÍTULO 2

REGRESIÓN LOGÍSTICA

La regresión logística (RL) fue propuesta por Swaminathan y Rogers en los años 90 como técnica de análisis del DIF (Swaminathan y Rogers, 1990; Rogers y Swaminathan, 1993). Básicamente, como en otros modelos de regresión, este procedimiento pretende predecir los valores de una variable, la variable dependiente (VD), a partir de los valores conocidos de una o varias variables predictoras (VIs).

Esta técnica se basa en el modelado estadístico de la probabilidad de obtener una respuesta correcta al ítem, en función de dos variables: la pertenencia al grupo (referencia o focal) y el nivel de habilidad de los sujetos (puntuación empírica u observada en el test o bien el nivel de habilidad estimado bajo algún modelo de respuesta al ítem). Para ello se iguala a los sujetos respecto de la habilidad medida por el test mediante la puntuación observada en el test.

El análisis de RL es uno de los métodos de comprobada eficacia para su uso en la detección del DIF uniforme y no uniforme (Clauser y Mazor, 1998). En el caso de ítems con

DIF uniforme, la RL produce resultados similares a los obtenidos con MH, aunque presenta una mayor potencia estadística para la detección de ítems con DIF no uniforme (Clauser, Nungester, Mazor y Ripkey, 1996; Ferreres, Fidalgo y Muñiz, 2000; Hidalgo y Gómez, 2000; Hidalgo y López-Pina, 2004; Narayanan y Swaminathan, 1996; Rogers y Swaminathan, 1993).

2.1. Extensiones

2.1.1. Regresión Logística Dicotómica (RLD)

Cuando los ítems tienen dos categorías de respuesta, se presenta el análisis de la RLD (Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990), que puede ser expresado en los siguientes términos:

$$\ln\left[\frac{\Pr(Y=1|G,X,XG)}{1-\Pr(Y=1|G,X,XG)}\right]=\beta_0+\beta_1X+\beta_2G+\beta_3GX$$

Esta expresión relaciona la probabilidad de dar una respuesta positiva a un ítem (Y=1) con la variable de agrupamiento categórico (G), el nivel en el rasgo o habilidad medido por el test (variable de equiparación), que suele ser la puntuación total observada (X), y la interacción del grupo y los resultados del test (GX). Bajo esta formulación un ítem muestra DIF uniforme si $\beta2\neq0$ y $\beta3=0$, y DIF no uniforme si $\beta3\neq0$.

Existen varias estrategias para la detección de DIF mediante la RL (Hidalgo, Gómez-Benito y Padilla, 2005; Paek, 2012), siendo la más usual la comparación de modelos usando el estadístico de razón de verosimilitud G^2. Ésta consiste:

- En primer lugar, en introducir la puntuación total del test en el modelo (Modelo 1), siendo éste el modelo base con ausencia de DIF.
- En segundo lugar se incluye el término de grupo en el modelo (Modelo 2).
- Y, en tercer lugar, se introduce la interacción en el modelo (Modelo 3).

La evaluación del DIF se lleva a cabo comprobando la presencia o ausencia de significación en las variables que se introducen en el modelo. De tal modo que:

- Se obtiene una prueba para el DIF, comparando los valores de G^2 del Modelo 1 (sin DIF) con los del Modelo 2; siguiendo una distribución $\chi 2$ con 1 grado de libertad.
- Se confirma la presencia de DIF no uniforme, comparando el valor de G^2 del modelo, en el segundo paso con el obtenido en el tercer paso; siguiendo una distribución $\chi 2$ con 1 grado de libertad.
- Por último, se confirma la presencia de DIF uniforme y no uniforme, comparando el valor de G^2 del modelo, en el primer paso con el obtenido en el tercer paso, esto sigue una distribución $\chi 2$ con 2 grados de libertad.

En este proceso de comparación de modelos, lo que se evalúa es la mejora explicativa al introducir un nuevo término al modelo.

La significación estadística del efecto de las variables explicativas sobre la variable dependiente puede evaluarse mediante (Gómez e Hidalgo, 1997):

- El test de Wald: evalúa la hipótesis nula de que un determinado parámetro logístico β_p es igual a cero, con una distribución χ^2 (para variables con un grado de libertad, equivale al cuadrado de la razón entre el parámetro estimado y su error estándar).

- Prueba de razón de verosimilitud: como se ha visto anteriormente.

2.1.2. Regresión Logística Multinomial (RLM)

Es utilizada en modelos con una variable dependiente nominal, con más de dos categorías. El modelo de RLM ha de ajustarse a modelos de RLD *m-1*, donde *m* es el número o las categorías de respuestas para un ítem dado (Miller y Spray, 1993; Miller y Spray y Wilson, 1992).

Para cada combinación de niveles de las variables explicativas, los modelos asumen que la respuesta para las categorías de Y tienen una distribución multinomial. La formulación del modelo depende de la definición del *logit*. Agresti (1984, 1990) discutió varios enfoques de construcción de *logits* implementando diferentes esquemas de codificación: *logits acumulativos, logits de relación de continuación* y *logits de categorías adyacentes.* Por lo tanto, para los tres *logits,* para una variable de respuesta con *m* categorías hay *m-1* modelos de RL que pueden ser evaluados para detectar DIF.

$$
\begin{aligned}
L_1 &= \beta_0^1 + \beta_1^1 X + \beta_2^1 G + \beta_3^1 XG \\
L_2 &= \beta_0^2 + \beta_1^2 X + \beta_2^2 G + \beta_3^2 XG \\
&\vdots \qquad \vdots \\
L_{m-1} &= \beta_0^{m-1} + \beta_1^{m-1} X + \beta_2^{m-1} G + \beta_3^{m-1} XG
\end{aligned}
$$

Si la variable de respuesta se mide en una escala ordinal, los *logits* pueden incorporar directamente el orden. Las probabilidades acumulativas son las probabilidades de que la respuesta Y caiga en la categoría *j* o inferior, para cada posible *j* (Agresti, 1984). En esta situación, es posible calcular modelos de RLM asumiendo un modelo de pendientes paralelas o un modelo de probabilidades proporcional. En este modelo sólo el parámetro de intercepción es diferente para cada probabilidad acumulativa; mientras que se supone que los

efectos de las variables predictoras son constantes a lo largo de las comparaciones (Menard, 1995), es decir, los predictores tienen el mismo efecto en las probabilidades de las categorías, independientemente del punto de corte para las probabilidades. Así que para la evaluación DIF esto significa que:

$$\hat{\beta}_0^1 \neq \hat{\beta}_0^2 \neq \hat{\beta}_0^3 \neq .... \neq \hat{\beta}_0^{m-1}$$
$$\hat{\beta}_1^1 X = \hat{\beta}_1^2 X = \hat{\beta}_1^3 X = ... = \hat{\beta}_1^{m-1} X$$
$$\hat{\beta}_2^1 G = \hat{\beta}_2^2 G = \hat{\beta}_2^3 G = ... = \hat{\beta}_2^{m-1} G$$
$$\hat{\beta}_3^1 XG = \hat{\beta}_3^2 XG = \hat{\beta}_3^3 XG = ... = \hat{\beta}_3^{m-1} XG$$

El modelo que describe simultáneamente los m-1 logits puede expresarse como:

$$L = \boldsymbol{\beta}_0 + \beta_1 X + \beta_2 G + \beta_3 XG$$

donde las ecuaciones de regresión tienen más de un coeficiente de intersección que depende del punto de corte y sólo una pendiente. Para probar la significación estadística de DIF, se utiliza la misma estrategia de modelado que con la RLD (Swaminathan y Rogers, 1990), porque los efectos se interpretan sólo en la RLD.

2.1.3. Regresión Logística Discriminante (DLR)

Miller y Spray (1993) proponen DLR como alternativa al análisis de RLM. El DLR tiene sus antecedentes en el análisis de RLD, sin embargo, al contrario de éste la variable dependiente no es la respuesta al ítem sino la variable de grupo, y considera la respuesta al

ítem como una variable independiente más del modelo. Miller y Spray (1993) definen la función discriminante logística como:

$$P(G/\,X,Y) = \frac{\exp(G-1)\,(\beta_0 + \beta_1 X + \beta_2 Y + \beta_3 XY)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 Y + \beta_3 XY)}$$

donde, G es la variable de grupo, Y es la respuesta al ítem (dicotómica o politómica), β_0 es el peso asociado a la intercepción, β_1 es el peso asociado a la variable X (puntuación observada en el test), β_2 el peso asociado a la variable Y (respuesta al ítem) y β_3 representa el peso asociado a la interacción puntuación observada en el test por la respuesta al ítem (XY).

La presencia de DIF, aquí, es un indicador de que al menos una de las categorías del ítem funciona diferencialmente, pero no aporta información adicional sobre la dirección del DIF (si favorece al grupo de referencia o al grupo focal) ni sobre la ubicación del mismo (por ejemplo, si se da en la primera categoría del ítem o en la segunda).

Lo que sí evalúa el DLR es la presencia de DIF uniforme y no uniforme, que puede ser modelado en la misma ecuación, pudiéndose probar por separado los coeficientes para cada uno. Estas hipótesis, normalmente, se someten a prueba mediante la comparación de modelos empleando una prueba de razón de verosimilitud condicional.

El modelo de ausencia de DIF implica que la pertenencia de un sujeto a uno u otro grupo se puede explicar por la habilidad en el test, y matemáticamente vendría dado por:

$$P(G/X) = \frac{\exp(G-1)(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$$

Si se compara la estadística de razón de verosimilitud estimada en ausencia de DIF (el modelo nulo) con la obtenida cuando se ajusta el modelo para la presencia de DIF (el modelo completo):

- Si la diferencia entre las dos estadísticas es significativa, el ítem muestra DIF:

    - Si el efecto de la variable de respuesta al ítem (Y) es significativo pero el de la interacción no lo es, el ítem muestra DIF uniforme.
    - Si la puntuación de la prueba observada y la interacción de la respuesta del ítem (XY) son significativas, entonces el ítem muestra DIF no uniforme.

Si se evalúa una ausencia de DIF, es decir, cuando el modelo que mejor se ajusta a los datos es el modelo nulo, la probabilidad de pertenencia a un grupo sólo depende del rasgo medido por la prueba.

En condiciones DIF, el análisis debe completarse trazando las funciones discriminantes estimadas de los niveles de puntuación de los ítems relevantes para observar la dirección del DIF y donde, a lo largo de la escala de prueba, ha ocurrido (Miller y Spray, 1993).

Este procedimiento resulta ventajoso para evaluar DIF en ítems politómicos frente a la regresión logística RLM porque:

- No necesita ajustar tantos modelos de RLD como categorías del ítem menos una.

- Detecta mayor porcentaje de ítems con DIF que la RLM, aun presentando una tasa de falsos positivos similar (Hidalgo y Gómez-Benito, 2010).

2.2. Medidas de tamaño del efecto en la RL

Como ya se ha comentado anteriormente, es recomendable el uso de medidas del tamaño del efecto como complemento a las pruebas de significación, con el fin de evaluar la magnitud del efecto observado.

Algunos autores proponen diversos procedimientos para el cálculo de las medidas del tamaño del efecto en la RL. Zumbo y Thomas (1997) sugieren el incremento en R^2; Gómez-Benito e Hidalgo (2007); Monahan, et al. (2007) proponen el uso de la odds-ratio para ítems dicotómicos e Hidalgo, Gómez-Benito y Zumbo (2008) para ítems politómicos.

2.3.1. Medidas R^2

Para ítems dicotómicos hay tres opciones disponibles de R^2: Nagelkerke R^2, mínimos cuadrados ponderados ΔR^2 y R^2 ordinal; para ítems politómicos la opción es el cálculo de R^2 ordinal, como puede verse en la Tabla 2.1:

Tabla 2.1

Medidas R^2 disponibles para medir la magnitud de DIF

Ítems	**Medida**	**Autores**
Ordinales	R^2 ordinal	McKelvey y Zavoina (1975)
Dicotómicos (nominales)	R^2 de Nagelkerke	Thomas y Zumbo (1998)
Dicotómicos (nominales)	ΔR^2	Thomas y Zumbo (1998)
Dicotómicos (ordinales)	R^2 ordinal	McKelvey y Zavoina (1975)

Cada medida de R^2 muestra sus ventajas:

- R^2 de Nagelkerke se obtiene fácilmente mediante el paquete estadístico SPSS (Zumbo, 1999).

- ΔR^2, Weighted least squares (WLS), una medida ponderada del tamaño del efecto, de mínimos cuadrados, para emplear cuando se usa la RL, con el fin de cuantificar la magnitud del DIF uniforme o no uniforme (Zumbo y Thomas (1997).

- R^2 ordinal proporciona una estrategia uniforme para modelar ítems politómicos ordinales y dicotómicos, porque asume una variable latente continua (Zumbo, 1999)

Zumbo (1999) indica como práctica útil, el cálculo del efecto R^2 para (a) el DIF uniforme, y (b) un test simultáneo de DIF uniforme y no uniforme. Tal y como indica Zumbo (1999), esta estrategia es útil porque es capaz de aprovechar la naturaleza jerárquica del modelado DIF y de comparar la R^2 del DIF uniforme con la del simultáneo DIF uniforme y no uniforme, para medir el sentido de la magnitud del DIF o el DIF no uniforme.

2.3.2. Medidas Odds Ratio

El Odds Ratio (OR) es un tipo muy intuitivo de medida del tamaño del efecto y puede derivarse fácilmente de la RLD (Monahan, et al. 2007).

Por lo tanto, la transformación exponencial de los coeficientes de regresión logística ($\hat{\beta}_j$) proporciona un tamaño del efecto. Otros tamaños del efecto basados en $\hat{\beta}_j$ incluyen el sistema de clasificación del ETS, los índices de estandarización basados en modelos de las diferencias condicionales en las proporciones, el OR ajustado y el estadístico delta, entre otros.

Los coeficientes de LR ($\hat{\beta}_j$) se estiman en la escala log odds. El exponencial de $\hat{\beta}_j$ proporciona la odds ratio estimada de probabilidad máxima del evento de interés para cada incremento de una unidad en el j-ésimo predictor, ajustado para otras covariables en el modelo LR (Hosmer y Lemeshow, 2000). Así, el exponencial de $\hat{\beta}_2$ proporciona de referencia a focal el odds ratio del ítem:

$$\widehat{\propto}_{LR} = \exp(\hat{\beta}_2)$$

Las OR van desde 0 hasta ∞. Los valores de $\widehat{\propto}_{LR}$ adicionales de 1.0 representan una magnitud de DIF mayor. Una odds ratio y su recíproco son equivalentes en intensidad pero no simétricos en la distancia desde el valor nulo de 1.0 (Monahan, et al., 2007).

De acuerdo con DeMars (2011), el OR como una medida del tamaño del efecto es estable a través de diferentes niveles de dificultad del ítem, pero se ve afectado por un aumento en el parámetro de discriminación, que produce mayores tamaños de efecto.

Otra opción para medir el tamaño del efecto es la transformación de $\widehat{\propto}_{LR}$ a la definición logística de la escala delta, utilizada por ETS para medir la dificultad del ítem. Utilizamos la fórmula de Holland y Thayer (1988) para convertir la odds ratio de Mantel-Haenszel ($\widehat{\propto}_{MH}$) en el estadístico MH delta-DIF (MH-D-DIF o $\widehat{\Delta}_{MH}$):

$$\text{LR-D-DIF o } \widehat{\Delta}_{LR} = -2.35 \ln(\widehat{\propto}_{LR}) = -2.35(\hat{\beta}_2)$$

Además, se puede calcular el sistema de clasificación ETS (Dorans y Holland, 1993):

- Categoría A. Ítems con DIF insignificante o no significativo. Definido por LR-D-DIF no significativamente diferente de cero o valor absoluto inferior a 1.0.

- Categoría B. Ítems con magnitud leve a moderada de DIF estadísticamente significativo. Definido por LR-D-DIF significativamente diferente de cero y valor absoluto de al menos 1.0 y menos de 1.5 o no significativamente mayor que 1.0.

- Categoría C. Ítems con magnitud moderada a grande de DIF estadísticamente significativo. Definido por el valor absoluto de LR-D-DIF de al menos 1.5 y significativamente mayor que 1.0.

Asignar las categorías (Monahan, et al., 2007):

- A y B implica usar LR para probar H_0: $\beta_2 = 0$.
- B y C requiere probar H_0: $|LR - D - DIF| \leq 1.1$.

A este respecto, Gómez-Benito, Hidalgo y Zumbo (2013, p. 895) concluyen que:

> Una manera de clasificar un ítem que muestra DIF es usar una regla de decisión combinada, de tal manera que tanto el valor p como el tamaño del efecto (con correspondiente criterio) se tengan en cuenta al decidir si el DIF está presente o no. Este enfoque combinado es en realidad una prueba estadística diferente y, por lo tanto, los valores obtenidos para la tasa de error Tipo I, la tasa de falsos positivos o la potencia estadística son muy diferentes de los obtenidos mediante medidas de tamaño de efecto usadas convencionalmente.

2.4. Purificación de la variable de equiparación

No es común el empleo de la purificación en estudios de simulación; sólo unos pocos autores incluyen este procedimiento dentro de sus estudios. Trabajos en los que se compara la RL con otros procedimientos para el análisis del DIF con datos simulados, como los de French y Maller (2007) y Shih, Liy y Wang (2014) emplean la técnica de purificación iterativa. Otros trabajos, como los de Navas-Ara y Gómez-Benito (2002), Welkenhuysen-Gybels y Billiet (2002) e Hidalgo y Gómez-Benito (2003; 2006) emplean el procedimiento de purificación de dos pasos.

Emplear técnicas de purificación, tal y como concluyen Hidalgo y Gómez-Benito (2003; 2006) mejora las tasas de falsos positivos y puede mejorar la detección correcta de

ítems con DIF. Navas-Ara y Gómez (2002) tras emplear el procedimiento de purificación en dos pasos, encontraron que la purificación de la escala de habilidad mejoró considerablemente la detección del sesgo en los ítems, proporcionando tasas de identificación correctas cercanas al 100% con la RL. French y Maller (2007) también indican que, en su estudio, fue beneficioso el empleo de la técnica de purificación iterativa, aunque bajo ciertas condiciones, la potencia total y las tasas de error Tipo I no mejoraron sustancialmente; sin embargo, el uso combinado de la prueba estadística y el criterio de tamaño del efecto dio como resultado tasas de error bien controladas.

Son múltiples los autores que recomiendan la técnica de purificación como una fase más, en el proceso de detección de ítems con DIF. Zumbo (1999) muestra resumidamente los pasos para purificar "el criterio de equiparación" en el proceso de análisis del DIF y lo recomienda: eliminado primero los ítems detectados con DIF, en una primera aplicación de la RL, para calcular, nuevamente, la puntuación total, para finalmente emplear esa puntuación total (libre de ítems con DIF) como criterio de equiparación.

CAPÍTULO 3

REVISIÓN BIBLIOMÉTRICA

El comportamiento de la productividad científica ha sido, tradicionalmente, un tema de interés. Conocer el estado de la investigación en un campo determinado, el curso de su desarrollo y su tendencia de crecimiento, ofrece una información importante a la comunidad científica; que a su vez, puede verse beneficiada de este conocimiento, para dirigir u orientar su tarea investigadora hacia una mayor eficacia y aprovechamiento de los recursos.

Este conocimiento llega a nuestras manos mediante los estudios bibliométricos. Estos estudios pretenden describir el curso de una disciplina o área científica, presentando, a su vez, un gran impacto en la sociedad, en la ciencia y en la política, tal y como destaca Andrés (2009). Actualmente los estudios bibliométricos están presentes en la mayoría de disciplinas científicas ya que son considerados de gran utilidad para la comunidad científica, ya que supone un eficaz instrumento para una correcta distribución de recursos económicos para la investigación (Zulueta y Bordons, 1999).

Igualmente, no se debe de confundir que el objetivo de los estudios bibliométricos es ofrecer indicadores del nivel de difusión entre la comunidad científica y no evaluar la calidad de los artículos y de las revistas científicas, ya que para esto es necesario emplear otras herramientas e indicadores. Buela-Casal (2003) critica que el factor de impacto y otros índices bibliométricos son utilizados en la actualidad en diversos países para evaluar la producción y/o la calidad de la investigación científica, por lo que propone dos factores de impacto nuevos, el factor de impacto medio de las revistas donde se producen las citas (FIMRC) y el factor de impacto ponderado (FIP). Para Pelechano (2000) el uso de los estudios bibliométricos para la evaluación de la calidad de un artículo o una revista es confundir la ciencia con la sociología de la ciencia. De manera similar, Sternberg (2001) dice que hay que diferenciar entre lo que se publica y dónde se publica, ya que no todo lo que se publica en una misma revista tiene la misma calidad.

Con este trabajo se pretende dar una visión sobre el estado de la producción científica en el estudio y aplicación de la RL, como técnica de análisis y detección del DIF.

Los análisis llevados a cabo se centran en conocer la evolución de la producción científica de los trabajos que hayan empleado datos empíricos o simulados, en el estudio de la RL como técnica de análisis y detección del DIF.

Dada la necesidad de conocer el estado, el curso y la evolución de la producción científica en esta área, se realiza este trabajo.

3.1. Objetivos

El objetivo de este trabajo ha sido analizar la producción científica relacionada con el uso de la regresión logística como técnica de análisis y detección del DIF, con el objetivo de ofrecer una visión general de la actividad de investigación en este campo y caracterizar sus aspectos más importantes y su evolución durante la última década del siglo XX y principios del siglo XXI. Para tal fin se han establecido los siguientes objetivos específicos:

1. Estudiar la evolución temporal de las publicaciones.
2. Analizar la productividad de los autores y autoras: Ley de Lotka.
3. Analizar las colaboraciones entre autores.
4. Analizar la productividad de las instituciones.
5. Analizar la productividad de los países.
6. Analizar la productividad de las revistas: Ley de Bradford.

3.2. Metodología

3.2.1. Procedimiento

3.2.1.1. Obtención de los documentos

Para la identificación de documentos se ha empleado las bases de datos PsycInfo, Education Resources Information Center (ERIC) y Web of Science, por ser grandes bases de datos de la educación, la psicología y de información científica.

Para identificar los trabajos se ha tenido en cuenta los siguientes criterios de inclusión:

a) *Terminología:* Para la obtención de los documentos se ha empleado los términos "*Logistic Regression technique"* y "*Differential Item Functioning"*, debiendo ser mencionados en el título, resumen o en las palabras clave, siguiendo esta estrategia de búsqueda: ("differential item functioning" OR DIF) AND ("logistic regression" OR LR).

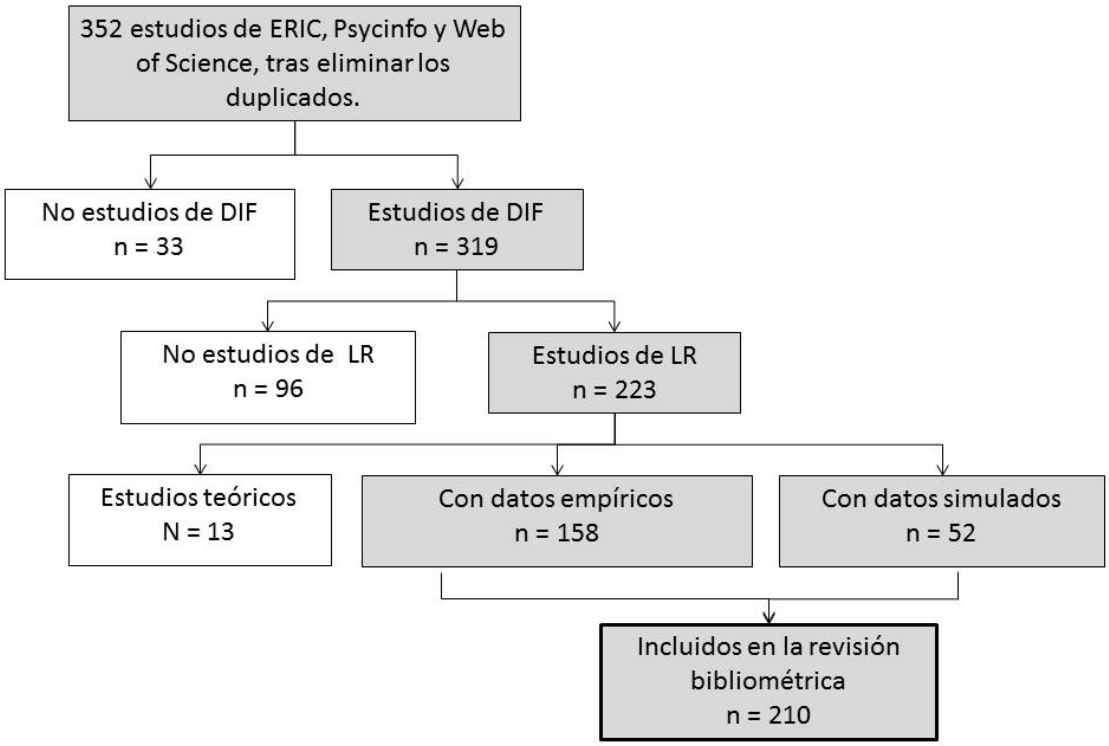
b) *Período de tiempo:* Los estudios son identificados el 9 de enero de 2017, cubriendo el período entre 1990 y 2016.

c) *Tipo de documento:* La búsqueda se limitó a artículos de revistas escritos en inglés, eliminándose los artículos en otros idiomas, las tesis, libros, capítulos, etc.

3.2.1.2. Criterios de inclusión y exclusión de documentos

Han sido incluidos los estudios empíricos, que han usado RL con el fin de detectar el DIF en instrumentos de medida, en el ámbito de las ciencias sociales y de la salud, así como los estudios de simulación, que han analizado la eficacia de esta técnica en términos de tasa de error Tipo I y/o potencia estadística. Han sido excluidos del presente estudio los artículos puramente teóricos, por no ser objeto de este trabajo, al igual que, los artículos que no sean en inglés, las tesis, libros, capítulos, etc.

Siguiendo esta estrategia, finalmente, han pasado a formar parte de la revisión bibliométrica 210 artículos (ver Figura 3.1).



DIF: Funcionamiento diferencial del ítem; LR: Regresión Logística.

Figura 3.1

Diagrama de flujo de información

3.3. Análisis de datos

Los datos se han tabulado con SPSS v.22.0.0.0 (2013), y con el paquete Excel de Microsoft Office Professional Plus (2010).

El resumen de variables codificadas es el siguiente:

- Año de publicación del artículo.
- Revista en la que se publica.
- Título del artículo .
- Tipo estudio (Simulado/Empírico).
- Nombre del autor o autora principal.
- Nombre del resto de autores y/o autoras.
- Filiación principal de cada autor o autora.
- Filiaciones secundarias de cada autor o autora.
- Región del país correspondiente a la filiación de cada autor.
- Número de autores y autoras por artículo.
- Colaboración entre autores (Colaboración Nacional/Colaboración Internacional/No existe colaboración)

3.3.1. Indicadores analizados

Con el fin de conocer la evolución y el estado de la producción científica en el uso y estudio de la RL, como técnica de detección del DIF, en este trabajo se presentan frecuencias y porcentajes de publicaciones, en relación a diversos indicadores bibliométricos.

En primer lugar se ha codificado como “tipo de estudio (simulado/empírico)” a los estudios por la naturaleza de los datos empleados.

Cuando se dice que se ha empleado, en un estudio, datos empíricos, se refiere a situaciones aplicadas en las que se ha estudiado el DIF en un instrumento dado con respecto a una o más variables de agrupamiento con datos empíricos. El objetivo de estos trabajos es evaluar la calidad, en este caso y específicamente, la validez, mediante el análisis del DIF, de la prueba objeto de estudio.

Cuando se dice que se ha empleado datos simulados, se refiere a los estudios en los que se ha empleado datos simulados mediante vectores de respuestas, para diferentes tamaños de muestra, diferente número de ítems, diferentes parámetros de los ítems, etc., mediante un modelo previamente definido. El objetivo de estos trabajos es evaluar la eficacia relativa de un procedimiento, en este caso, la eficacia de la RL en el análisis del DIF. En ocasiones, se compara la eficacia de la RL frente a otros procedimientos y, en ocasiones, se evalúa la eficacia de ésta frente a otras condiciones (diferente tamaño del test, parámetros de los ítems, tamaño de la muestra, etc.).

Igualmente, se ha codificado el año de publicación para cada uno de los artículos, con el fin de presentar la evolución del número de publicaciones a lo largo del tiempo.

Respecto a los autores de los trabajos, se ha descrito el número de autores y autoras en cada artículo, su evolución temporal y su productividad, como primer firmante y respecto al número de participaciones en los artículos.

De la misma forma, se ha llevado cabo un análisis de la producción científica por institución, teniendo en cuenta la filiación del autor o autora principal, permitiéndonos comprobar las instituciones más productivas e impulsoras de estudios, en el tema que nos aborda. Igualmente se ha analizado a los países más productivos, teniendo en cuenta el mismo criterio empleado para el análisis de las instituciones.

Asimismo, se ha realizado el estudio de la productividad de las revistas que han publicado los trabajos seleccionados para este estudio y su evolución temporal.

Los análisis realizados, en el presente estudio, son los más usados en estudios bibliométricos, por aportar gran información. Se analizan frecuencias y porcentajes de las variables seleccionadas. Además, se describe la productividad de autores y revistas usando la ley de Lotka (Lotka, 1926) y la ley de Bradford de dispersión (Bradford, 1934, 1948), respectivamente.

3.3.2. Productividad de los autores y autoras: Ley de Lotka

Lotka (1926) estudió los patrones de productividad de los autores y desarrolló una de las principales leyes de la bibliometría, la ley de Lotka. Éste, observó que, en un área dada de la ciencia, hay muchos autores que publican sólo un estudio, mientras que un pequeño grupo de autores prolíficos contribuye con un gran número de publicaciones. Esta premisa es la base de la ley de Lotka, también conocida como la ley del cuadrado inverso sobre la productividad del autor.

La ley se expresa en forma de:

$$y = C \text{ x } x^{-n} \qquad x = 1, 2, \dots, x_{max} \quad (1)$$

donde,

x es el número de publicaciones de interés

n es un exponente que es constante para un conjunto de datos

y es el porcentaje esperado de autores con frecuencia x de publicaciones

C es una constante.

Esto significa que la productividad no corresponde al número de artículos publicados por un autor sino a su logaritmo.

El exponente n se fija a menudo en 2, en cuyo caso la ley se conoce como la "Ley inversa del cuadrado de la productividad científica". Sin embargo, dado que el exponente n predice el número relativo de autores en cada nivel de productividad parecería útil calcularlo. En el presente estudio se utilizó el método de los mínimos cuadrados, cuya expresión es:

$$n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum x)^2}$$

donde,

N es el número de pares de datos considerados

X es el logaritmo de x (x = número de artículos)

Y es el logaritmo de y (y = número de autores).

La constante C se calcula utilizando la siguiente fórmula:

$$C = \frac{1}{\sum 1/x^n}$$

Para verificar que la distribución observada de la productividad del autor se ajusta a la distribución estimada, Pao (1985) sugiere aplicar la prueba no paramétrica de bondad de ajuste Kolmolgorov-Smirnov (K-S). Para ello, se calcula la diferencia máxima entre las frecuencias acumuladas real y estimada, comparándose este valor con el valor crítico (c.v.) obtenido a partir de la siguiente fórmula:

$$c.v. = \frac{1{,}63}{(\sum y_x + ((\sum y_x/10)^{1/2})^{1/2}}$$

3.3.3. Productividad de las revistas: Ley de Bradford

Respecto a la productividad de las revistas, cabe mencionar, y aplicar, la ley de Bradford. Esta ley, Bradford (1934, 1948), que lleva su nombre, nos ofrece un indicador de dispersión científica, basándose en un modelo de zonas de productividad concéntricas con una densidad de información decreciente. Explicado de otra manera, gráficamente puede representarse de manera más sencilla, tal y como puede verse en la Figura 3.2.

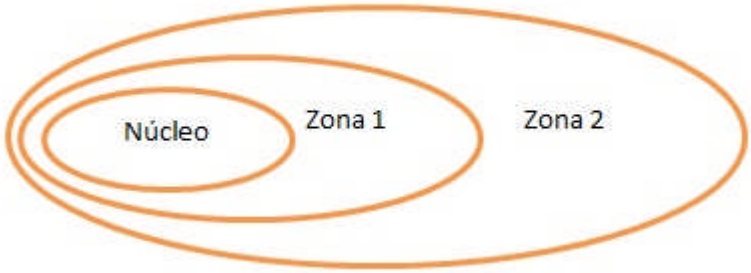


Figura 3.2
Zonas Bradford

Las llamadas zonas de Bradford son grupos o categorías en las que se clasifican las revistas sometidas a análisis. El número de artículos en cada zona es el mismo, sin embargo, el número de revistas que publican estos artículos no será el mismo en cada zona, el número de revistas aumentará a través de las zonas, ya que algunas revistas serán más productivas que otras. Se ha observado que la relación entre el número de revistas en zonas subsiguientes es aproximadamente $1, n, n^2, \ldots$ Así que, un pequeño grupo de artículos se ubicará en la primera zona central, mientras que un número creciente de revistas se encontrarán en las zonas posteriores.

Mediante este procedimiento se puede distinguir un grupo de revistas dedicadas específicamente al tema de interés (las que se concentran en el núcleo central) y que ejercen cierta influencia sobre las otras revistas.

Para el cálculo de la ley, primero de todo, hay que elaborar una tabla donde se recojan las frecuencias de las revistas y de los artículos publicados en las revistas.

A partir de estos datos, es necesario calcular el valor de la constante k de Bradford, que es el multiplicador que explicará cómo el número de revistas crece de una

zona a la siguiente. La siguiente fórmula para el multiplicador k de la ley de Bradford ha sido formulada por Egghe (1986, 1990) y Egghe y Rousseau (1990):

$$k = (e^{\gamma} \times Y_m)^{1/P}$$

donde,

γ es el número de Euler ($\gamma = 0{,}5772$)

Y_m es la productividad máxima de la revista de rango uno

P es el número de zonas o grupos de Bradford.

En segundo lugar, se calcula el número de revistas que pertenecerán al primer grupo de Bradford, al grupo núcleo, representado como r_0, y se optiene mediante la siguiente fórmula:

$$r_0 = \frac{T(k-1)}{(k^P - 1)}$$

donde,

T representa el número total de revistas que publican artículos en una materia dada

k es la constante de Bradford

P es el número de grupos de Bradford.

Conociendo los valores de k y r_0, se puede obtener la distribución teórica de las revistas a través de las zonas de Bradford (correspondientes a r_0, r_1, r_2 y r_3).

Donde,

$$r_0 = r_0 \times 1$$

$$r_1 = r_0 \times k$$

$$r_2 = r_0 \times k^2$$

$$r_3 = r_0 \times k^3$$

3.4. Resultados

Tras la búsqueda en las bases de datos y eliminar los artículos duplicados, se obtienen 352 artículos publicados en revistas científicas, en inglés, entre 1990 y 2016. Entre estos artículos, es en 319 en los que se estudia el DIF y en 223, de éstos, se emplea la RL como técnica de análisis de DIF. No se analizan en este estudio los 223 trabajos por contener artículos teóricos, por lo que finalmente son incluidos en esta revisión 210.

3.4.1. Publicaciones

3.4.1.1. Tipo de datos que emplean

Se distinguen dos tipos de estudios, aquellos que emplean datos empíricos y los que trabajan con datos simulados. Respecto a esta distinción, y en base a la búsqueda realizada, se ha encontrado 158 trabajos que emplean datos empíricos, un 75.2% del total de los trabajos seleccionados, y 52 que emplean datos simulados, un 24.8%, como se puede ver en la Figura 3.3:

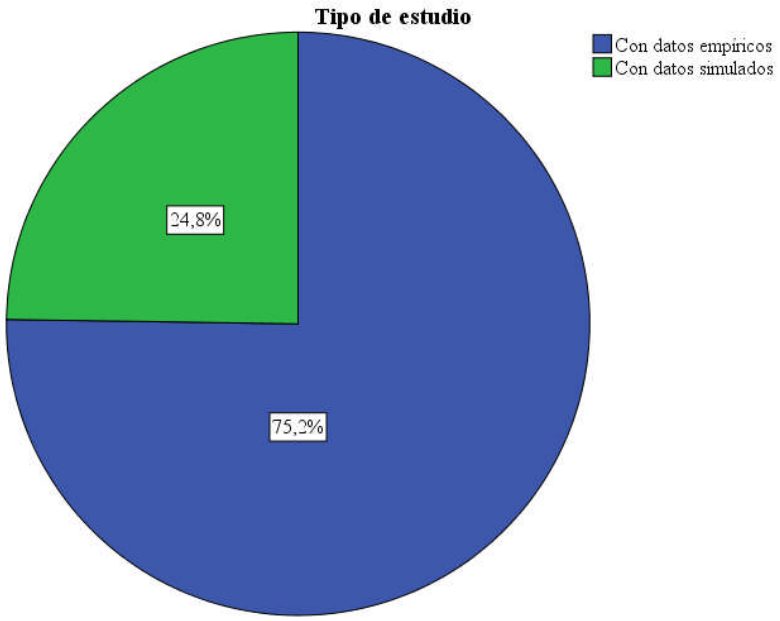


Figura 3.3

Porcentaje de estudios que emplean datos empíricos o simulados

3.4.1.2. Año de publicación

Como se puede observar en la Figura 3.4, el interés por el estudio y aplicación de la RL como técnica de detección del DIF ha ido creciendo exponencialmente, desde los años noventa hasta el pasado año.

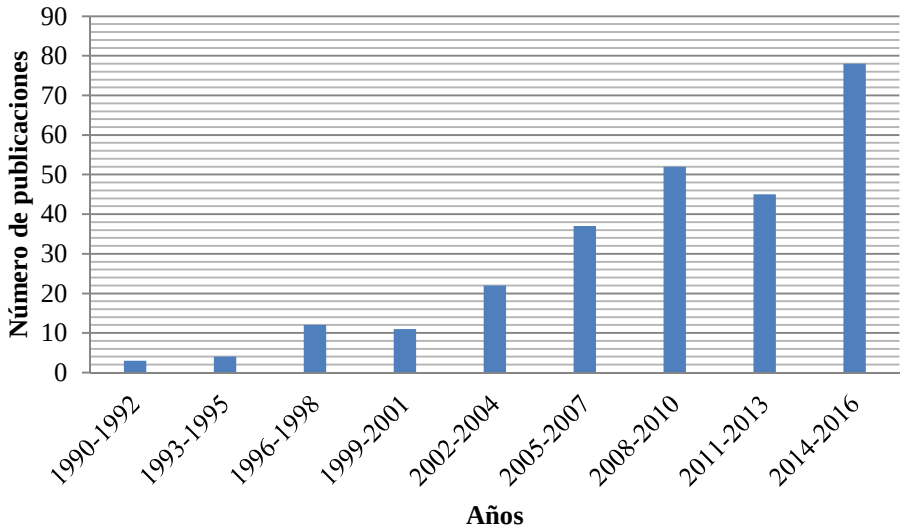


Figura 3.4

Número de publicaciones por año

Se observa un salto importante a partir de 2006, ver Tabla 3.1, pasando de 3 artículos publicados en 2005 a 11 en 2006, alcanzando su máximo auge en 2014, con 24 artículos publicados.

Tabla 3.1

Número de publicaciones por año

AÑO	Nª de artículos	Porcentaje
1990	1	.5
1992	1	.5
1993	2	1.0
1995	1	.5
1996	5	2.4
1997	1	.5
1998	1	.5
1999	3	1.4
2000	1	.5
2001	4	1.9
2002	5	2.4
2003	5	2.4
2004	7	3.3
2005	3	1.4
2006	11	5.2
2007	17	8.1
2008	9	4.3
2009	19	9.0
2010	12	5.7
2011	16	7.6
2012	7	3.3
2013	17	8.1
2014	24	11.4
2015	19	9.0
2016	19	9.0
Total	**210**	**100.0**

El interés por la aplicación y el estudio de la RL ha generado y genera interés en la comunidad científica, tal y como, se puede observar en el creciente número de publicaciones en esta área.

Como se puede ver en la Figura 3.5, los trabajos con datos empíricos, frente a los trabajos con datos simulados, son más numerosos, siendo, igualmente, más

significativo su crecimiento a lo largo de los años. Al igual, los trabajos que emplean datos simulados, van creciendo a lo largo de los años, alcanzando su máximo en los tres últimos años, de 2014 a 2016.

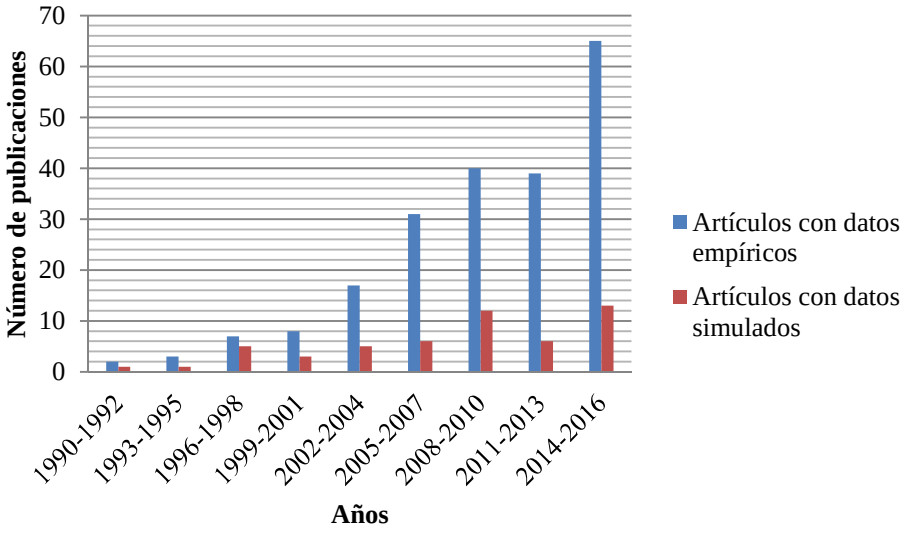


Figura 3.5

Número de publicaciones, según tipo de datos empleado, por año

3.4.2. Autores y Autoras

Un total de 569 autores y autoras participan en los artículos revisados.

3.4.2.1. Participación de los autores y autoras

Primero de todo, se analiza la producción de estos autores respecto al tema objeto de estudio. Para ello, se ha registrado a todos los firmantes, en términos de número de participaciones en artículos.

En la Tabla 3.2 se muestra el listado de autores junto con el número total de publicaciones realizadas por cada uno de ellos, se ha excluido, de esta tabla a los autores con menos de tres publicaciones.

Tabla 3.2

Autores y autoras con mayor número de publicaciones

Autor/a	Nª de artículos	%
Crane, PK	13	1.6
Gómez-Benito, J	10	1.2
Zumbo, BD	9	1.1
French, BF	8	1.0
Gibbons, LE	8	1.0
Groenvold, M	8	1.0
Hidalgo-Montesinos, MD	8	1.0
Aaronson, NK	6	.7
Bjorner, JB	6	.7
Fayers, PM	6	.7
Finch, WH	6	.7
Petersen, MA	6	.7
Scott, NW	6	.7
Bottomley, A	5	.6
Clauser, BE	5	.6
De Graeff, A	5	.6
Ercikan, K	5	.6
Hart, DL	5	.6
Jafari, P	5	.6
Koller, M	5	.6
Sprangers, MAG	5	.6
Bagheri, Z	4	.5
Cook, KF	4	.5
Ravens-Sieberer, U	4	.5
Swaminathan, H	4	.5
Amtmann, D	3	.4
Auquier, P	3	.4
Bonnema, SJ	3	.4
Cella, D	3	.4
De Boeck, P	3	.4
Deutscher, D	3	.4
Erhart, M	3	.4
Feldt-Rasmussen, U	3	.4

Gundy, C	3	.4
Kim, J	3	.4
Manly, JJ	3	.4
Mazor, K	3	.4
McHorney, CA	3	.4
Mukherjee, S	3	.4
Mungas, D	3	.4
Narasimhalu, K	3	.4
Navas-Ara, MJ	3	.4
Nungester, RJ	3	.4
Oliveri, ME	3	.4
Sierra, JC	3	.4
Stump, TE	3	.4
Terluin, B	3	.4
Tutz, G	3	.4
Vallejo-Medina, P	3	.4
Van Belle, G	3	.4
Wang, Y	3	.4
Watt, T	3	.4

** El porcentaje de productividad de cada autor y autora se obtuvo teniendo en cuenta el número total de publicaciones.*

Cabe destacar a autores y autoras como P. K. Crane, J. Gómez-Benito, B. D. Zumbo, L. E. Gibbons, M. Groenvold, B. F. French y M. D. Hidalgo-Montesinos por el gran número de participaciones en las publicaciones seleccionadas.

Cuando se refiere a autores y autoras que participan en estudios con datos empíricos, los porcentajes de participación cambian. Los autores y autoras con más participaciones se exponen en la Tabla 3.3, tras excluir a los autores con menos de tres publicaciones en esta clasificación. En la Figura 3.6 se ve más gráficamente a los autores y autoras con mayor número de publicaciones en estudios realizado con datos empíricos.

Tabla 3.3

Autores y autoras con mayor número de publicaciones en estudios realizados con datos empíricos

Autor/a	Nª de artículos	%
Crane, PK	13	1.9
Gibbons, LE	8	1.2
Groenvold, M	7	1.0
Bjorner, JB	6	.9
Fayers, PM	5	.7
Hart, DL	5	.7
Petersen, MA	5	.7
Scott, NW	5	.7
Zumbo, BD	5	.7
Aaronson, NK	4	.6
Bottomley, A	4	.6
Cook, KF	4	.6
De Graeff, A	4	.6
Ercikan, K	4	.6
French, BF	4	.6
Jafari, P	4	.6
Koller, M	4	.6
Ravens-Sieberer, U	4	.6
Sprangers, MAG	4	.6
Amtmann, D	3	.4
Auquier, P	3	.4
Bagheri. Z	3	.4
Bonnema. SJ	3	.4
Cella. D	3	.4
Clauser. BE	3	.4
Deutscher. D	3	.4
Erhart. M	3	.4
Feldt-Rasmussen. U	3	.4
Manly. JJ	3	.4
McHorney. CA	3	.4
Mukherjee. S	3	.4
Mungas. D	3	.4
Narasimhalu. K	3	.4
Reeve. BB	3	.4
Sierra. JC	3	.4
Stump. TE	3	.4
Terluin. B	3	.4
Vallejo-Medina. P	3	.4

Van Belle. G	3	.4
Wang. Y	3	.4
Watt, T	3	.4

** El porcentaje de productividad de cada autor y autora se obtuvo teniendo en cuenta el número total de publicaciones.*

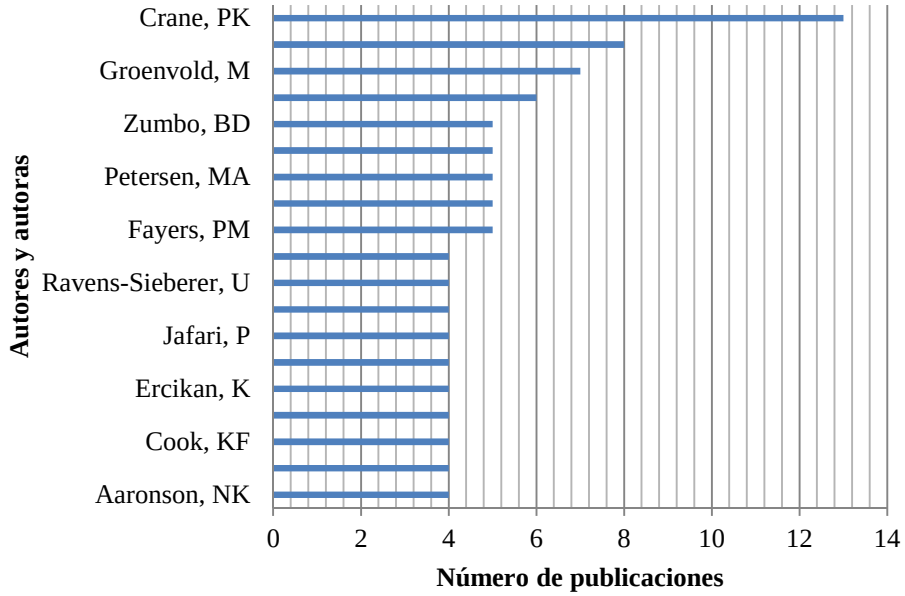


Figura 3.6

Autores y autoras con mayor número de publicaciones en estudios realizados con datos empíricos

En la Figura 3.6 se observa que los autores P. K. Crane y L. E. Gibbons son los más productivos, seguidos de P. M. Fayers, D. L. Hart, M. A. Petersen, N. W. Scott, y B. D. Zumbo. En este caso se puede observar como B. D. Zumbo es uno de los autores más productivos en el estudio de la RL como técnica de detección del DIF, ya sea en estudios que emplean datos simulados como datos empíricos.

En la Tabla 3.4, se puede observar a los autores y autoras que tienen un mayor número de participaciones en estudios con datos simulados, tras excluir a los autores con menos de tres publicaciones en esta clasificación.

Tabla 3.4

Autores y autoras con mayor número de publicaciones en estudios realizados con datos simulados

Autor/a	Nª de artículos	%
Gómez-Benito, J	8	6.4
Hidalgo-Montesinos, MD	6	4.8
Finch, WH	5	4.0
French, BF	4	3.2
Zumbo, BD	4	3.2
Swaminathan, H	3	2.4
Tutz, G	3	2.4

** El porcentaje de productividad de cada autor y autora se obtuvo teniendo en cuenta el número total de publicaciones.*

En este caso, son J. Gómez-Benito y M. D. Hidalgo-Montesinos, las autoras con mayor número de publicaciones, tras ellas encontramos a W. H. Finch, B. F. French, H. Swaminathan y G. Tutz.

3.4.2.2. Autores y autoras como primer firmante

Y finalmente, nos vamos a centrar en la clasificación de los autores y autoras que figuran como primer firmante. En la Tabla 3.5, tras excluir a los autores con menos de tres publicaciones, se puede ver la ordenación de éstos, de mayor a menor, según número de publicaciones como primer firmante. Entre ellos, P. K. Crane, W. H. Finch, N. W. Scott, J. Gómez-Benito y M. D. Hidalgo-Montesinos.

Tabla 3.5

Autores y autoras como primer firmante en mayor número de trabajos

Autor/a	Nª de artículos	%
Crane, PK	7	3.3
Finch, WH	5	2.4
Scott, NW	5	2.4
Gómez-Benito, J	4	1.9
Hidalgo-Montesinos, MD	4	1.9
French, BF	3	1.4
Hart, DL	3	1.4
Jafari, P	3	1.4
Oliveri, ME	3	1.4

** El porcentaje de productividad de cada autor y autora se obtuvo teniendo en cuenta el número total de publicaciones.*

Entre los autores y autoras como primer firmante, en mayor número de trabajos, se encuentra P. K. Crane, en primer lugar, seguido de W. H. Finch, N. W. Scott, J. Gómez-Benito y M. D. Hidalgo-Montesinos.

Teniendo en cuenta los trabajos que emplean datos empíricos, como se observa en la Tabla 3.6, y tras excluir a los autores con menos de tres, se ve la ordenación de éstos, de mayor a menor, según número de publicaciones como primer firmante.

Tabla 3.6

Autores y autoras como primer firmante en los trabajos que emplean datos empíricos

Autor/a	Nª de artículos	%
Crane, PK	7	4.4
Scott, NW	4	2.5
Hart, DL	3	1.9
Jafari, P	3	1.9
Crane, PK	7	4.4
Scott, NW	4	2.5
Hart, DL	3	1.9

** El porcentaje de productividad de cada autor y autora se obtuvo teniendo en cuenta el número total de publicaciones.*

En este caso, los autores como primer firmante, en mayor número de trabajos que emplean datos empíricos son P. K. Crane y N. W. Scott.

Cuando se tienen en cuenta los trabajos que emplean datos simulados, como se puede ver en la Tabla 3.7, y tras excluir a los autores con menos de dos publicaciones en esta clasificación, se puede ver la ordenación de éstos según mayor o menor número de publicaciones firmadas como primer autor o autora.

Tabla 3.7

Autores y autoras como primer firmante en los trabajos que emplean datos simulados

Autor/a	Nª de artículos	%
Finch, WH	4	7.7
Hidalgo-Montesinos, MD	4	7.7
Gómez-Benito, J	3	5.8
French, BF	2	3.8
Welkenhuysen-Gybels, J	2	3.8

** El porcentaje de productividad de cada autor y autora se obtuvo teniendo en cuenta el número total de publicaciones.*

Para los trabajos que emplean datos simulados, encontramos que los autores, como primer firmante, con mayor número de trabajos son W. H. Finch y M. D. Hidalgo-Montesinos, seguidos de J. Gómez-Benito.

3.4.2.3. Productividad de los autores y autoras: Ley de Lotka

Con el fin de analizar y describir la productividad de los autores y las autoras en los trabajos seleccionados, se aplica la ley de Lotka.

El primer paso en la aplicación de la ley de Lotka, es decidir qué datos serán considerados para el análisis. Existen diferentes opciones, desde el recuento completo, donde cada participación de un autor es reconocida y recibe tratamiento igual, el recuento únicamente del primer autor y el recuento según número de autores. La opción seleccionada en este caso es la del recuento completo.

Los autores y autoras han contribuido entre uno y trece artículos. Sin embargo, la mayoría de éstos han contribuido con un pequeño número de artículos, mientras que un pequeño grupo de autores y autoras han sido muy prolíficos. Como puede verse, en la Tabla 3.8, el 82.2% de los autores y autoras contribuyen en un artículo, y solo un 3.8% contribuyen con más de 5 artículos.

Tabla 3.8

Productividad de los autores y autoras

Nº de artículos	Nº de autores	%
1	468	82.2
2	49	8.6
3	27	4.7
4	4	.7
5	8	1.4
6	6	1.1
8	4	.7
9	1	.2
10	1	.2
13	1	.2

A continuación se muestra, en la Tabla 3.9, la aplicación de la ley de Lotka. Teniendo en cuenta que N corresponde al número de pares de datos, y que según Lotka, cuando $x_y = 1$ se encuentra al final de la distribución, correspondiente a los valores más altos de x (número de artículos). En estos casos, este pequeño grupo de autores más prolíficos se excluyen del análisis a fin de no sobrestimar los resultados. Por lo que hemos excluido a los tres autores que tienen 9, 10 y 13 artículos.

Tabla 3.9

Aplicación de la ley de Lotka

x	y_x	$X=lg\ x$	$Y=lg\ y$	X^2	XY	$y_x/\sum y_x$	$\sum(y_x/\sum y_x)$	$f_e = C(\frac{1}{x^n})$	$\sum f_e$	D
1	468	.000	2.670	.000	.000	.827	.827	.736	.729	.098
2	49	.301	1.690	.091	.509	.087	.913	.145	.874	.039
3	27	.477	1.431	.228	.683	.048	.961	.056	.931	.031
4	4	.602	.602	.362	.362	.007	.968	.029	.959	.009
5	8	.699	.903	.489	.631	.014	.982	.017	.976	.006
6	6	.778	.778	.606	.606	.011	.993	.011	.987	.005
8	4	.903	.602	.816	.544	.007	1.000	.006	.993	.007
∑	566	3.760	8.677	2.590	3.335					

Después del análisis mostrado en la Tabla 3.9, el valor de n calculado por el método de mínimos cuadrados fue de 2.33, dando un valor C de .734. Como el valor de la diferencia máxima entre las frecuencias acumuladas real y estimada fue .093, es decir, mayor que el valor crítico (c.v. = .068), la hipótesis nula tiene que ser rechazada. Por lo tanto, podemos concluir que la productividad del autor en esta hipotética área de investigación no se ajusta a la ley de Lotka.

Respecto a los trabajos que emplean datos empíricos, los autores y autoras han contribuido entre uno y trece artículos. Sin embargo, la mayoría de éstos han contribuido con un pequeño número de artículos, mientras que un pequeño grupo de autores y autoras han sido muy prolíficos. Como se puede ver, en la Tabla 3.10, el 83.9% de éstos contribuyen en un artículo, y solo un 1.8% contribuyen con cinco o más artículos.

Tabla 3.10

Productividad de los autores y autoras que realizan estudios con datos empíricos

Nº de artículos	Nº de autores	%
1	433	83.9
2	42	8.1
3	22	4.3
4	10	1.9
5	5	1
6	1	.2
7	1	.2
8	1	.2
13	1	.2

Teniendo en cuenta que N corresponde al número de pares de datos, y que según Lotka, cuando $x_y = 1$ se encuentra al final de la distribución, correspondiente a los valores más altos de x (número de artículos). En estos casos, este pequeño grupo de

autores más prolíficos se excluyen del análisis a fin de no sobrestimar los resultados. Por lo que hemos excluido a los cuatro autores que tienen 6, 7, 8 y 13 artículos. La aplicación de la ley de Lotka, bajo estas condiciones, se puede observar en la Tabla 3.11.

Tabla 3.11

Aplicación de la ley de Lotka para los autores de trabajos con datos empíricos

x	y_x	X=lg x	Y=lg y	X^2	XY	$y_x/\sum y_x$	$\sum(y_x/\sum y_x)$	$f_e = C(\frac{1}{x^n})$	$\sum f_e$	D
1	433	.000	2.636	.000	.000	.846	.822	.802	.802	.020
2	42	.301	1.623	.091	.489	.082	.905	.125	.927	-.023
3	22	.477	1.342	.228	.640	.043	.947	.042	.969	-.022
4	10	.602	1.000	.362	.602	.020	.967	.020	.989	-.022
5	5	.699	.699	.489	.489	.010	.977	.011	1.000	-.023
∑	91	1.380	3.289	.681	.559					

Después del análisis mostrado en la Tabla 3.11, el valor de n calculado por el método de mínimos cuadrados fue de 2.68, dando un valor C de .80. Como el valor de la diferencia máxima entre las frecuencias acumuladas real y estimada fue .020, es decir, menor que el valor crítico (c.v. = .072), la hipótesis nula se acepta. Por lo tanto, podemos concluir que la productividad del autor en esta hipotética área de investigación se ajusta a la ley de Lotka.

Respecto a los trabajos que emplean datos simulados, los autores y autoras han contribuido entre uno y trece artículos. Sin embargo, la mayoría estos han contribuido con un pequeño número de artículos, mientras que un pequeño grupo de autores y autoras han sido muy prolíficos. Como se puede ver, en la Tabla 3.12, el 86.2% de los

autores contribuyen en un artículo, y solo un 3.3% contribuyen con cinco o más artículos.

Tabla 3.12

Productividad de los autores y autoras que realizan estudios con datos simulados

Nº de artículos	Nº de autores	%
1	81	86.2
2	6	6.4
3	2	2.1
4	2	2.1
5	1	1.1
6	1	1.1
8	1	1.1

Teniendo en cuenta que N corresponde al número de pares de datos, y que según Lotka, cuando $x_y = 1$ se encuentra al final de la distribución, correspondiente a los valores más altos de x (número de artículos). En estos casos, este pequeño grupo de autores más prolíficos se excluyen del análisis a fin de no sobrestimar los resultados. Por lo que hemos excluido a los tres autores que tienen 5, 6, y 8 artículos La aplicación de la ley de Lotka, bajo estas condiciones, se puede observar en la Tabla 3.13.

Tabla 3.13

Aplicación de la ley de Lotka para los autores de trabajos con datos simulados

x	y_x	$X=lg\ x$	$Y=lg\ y$	X^2	XY	$y_x/\sum y_x$	$\sum(y_x/\sum y_x)$	$f_e = C(\frac{1}{x^n})$	$\sum f_e$	D
1	81	.000	1.908	.000	.000	.890	.890	.434	1	81
2	6	.301	.78	.091	.234	.066	.956	.247	2	6
3	2	.477	.301	.228	.144	.022	.978	.178	3	2
4	2	.602	.301	.362	.181	.022	1.000	.141	4	2
∑	91	1.380	3.289	.681	.559					91

Después del análisis mostrado en la Tabla 3.13, el valor de n calculado por el método de mínimos cuadrados fue de .81, dando un valor C de .43. Como el valor de la diferencia máxima entre las frecuencias acumuladas real y estimada fue .456, es decir, mayor que el valor crítico (c.v. = .168), la hipótesis nula tiene que ser rechazada. Por lo tanto, podemos concluir que la productividad del autor en este hipotético área de investigación no se ajusta a la ley de Lotka.

3.4.2.4. Colaboración entre autores y autoras

En esta sociedad, son cada vez más comunes las colaboraciones entre autores y autoras, formando una amplia y heterogénea red de investigadores e investigadoras que trabajan de forma cooperativa. Estos investigadores pueden estar a miles de kilómetros o en el despacho de al lado, y pueden pertenecer a la misma institución o no. La cooperación genera sinergias que van más allá de lo que puede aportar la suma de las partes consideradas de forma individual (Hara, Solomon, Kim, y Sonnenwald, 2003). Es por ello que la tendencia, como se muestra más adelante, es hacia la colaboración entre, cada vez, más autores y autoras.

En los trabajos seleccionados para este estudio, el número clave de autores y autoras, que trabajan conjuntamente, es de dos, como se puede ver en la Tabla 3.14 y en Figura 3.7. Se observa que lo más frecuente es la cooperación entre dos autores, para la elaboración de trabajos conjuntamente.

Tabla 3.14

Número de autores y autoras por artículo

Número de autores y autoras	Nª de artículos	Porcentaje
2	60	28.6
3	47	22.4
4	36	17.1
1	15	7.1
5	14	6.7
6	13	6.2
7	8	3.8
8	5	2.4
9	4	1.9
11	3	1.4
10	2	1.0
12	1	.5
15	1	.5
23	1	.5
Total	**210**	**100.0**

Un 28.6% de los autores y autoras se agrupan en parejas para llevar a cabo este tipo de trabajos, seguidos de los que se agrupan en tríos, 22.4%, o cuartetos, 17.1%. Siendo poco frecuentes, con un 7.1% los autores y autoras que deciden liderar trabajos en solitario.

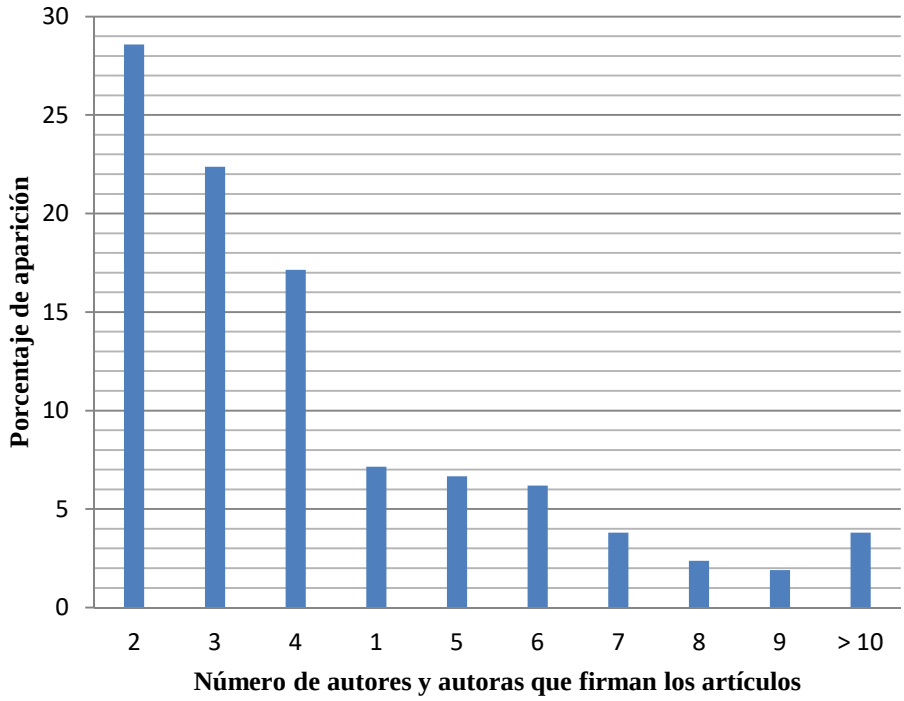


Figura 3.7

Porcentaje de aparición de número de autores y autoras que firman los artículos

Si distinguimos entre los trabajos que emplean datos empíricos o aplicados estos porcentajes varían.

En el caso de los trabajos con datos empíricos, se observa, en la Tabla 3.15, con más frecuencia, autores y autoras que deciden llevar a cabo sus trabajos en colaboración con otros colegas. Siendo lo más frecuente, con un 22.2%, los trabajos elaborados por tres autores y/o autoras y las colaboraciones entre dos y cuatro, con un 19% y 20.9% respectivamente. Igualmente se ha encontrado trabajos con hasta veintitrés autores firmantes, aunque son los menos frecuentes. Por el contrario, se puede observar como sólo un 5,7%, los trabajos son realizados por un autor o autora.

Tabla 3.15

Porcentaje de trabajos elaborados según número de autores y autoras

		Estudios con datos empíricos		Estudios con datos simulados	
		Nª de artículos	%	Nª de artículos	%
Número de autores y autoras	**1**	10	**5.7%**	6	**11.5%**
	2	30	**19.0%**	30	**57.7%**
	3	35	**22.2%**	12	**23.1%**
	4	33	**20.9%**	3	**5.8%**
	5	15	**8.9%**	0	**0%**
	6	13	**8.2%**	0	**0%**
	7	8	**5.1%**	0	**0%**
	8	5	**3.2%**	0	**0%**
	9	4	**2.5%**	0	**0%**
	10	2	**1.3%**	0	**0%**
	11	3	**1.9%**	0	**0%**
	12	0	**0%**	1	**1.9%**
	15	1	**.6%**	0	**0%**
	23	1	**.6%**	0	**0%**

En el caso de los trabajos con datos simulados, se observa en la Tabla 3.15 que más de la mitad, el 57.7%, están elaborados por dos investigadores y/o investigadoras. A esta preferencia, le sigue la de colaborar entre tres autores y/o autoras, con un 23.1%. Cabe destacar que, en este tipo de trabajos, es poco o casi nada usual encontrar trabajos realizados por más de cinco autores, en cambio es frecuente, con un 11.5%, encontrar trabajos elaborados individualmente.

En resumen, y tal y como se muestra en la Tabla 3.16, la media de autores y autoras que firman los trabajos que emplean datos empíricos es de 4.28, con una desviación de 2.81. En el caso de autores y autoras que firman los trabajos en los que se emplean datos simulados, nos encontramos con una menor media, estando en una media de 2.42 firmantes, con una desviación de 1.54.

Tabla 3.16

Estadísticos descriptivos respecto al número de autores y/o autoras firmantes de los trabajos, según tipo de estudio

Número de autores			
Tipo de estudio	**Media**	**N**	**Desviación estándar**
Con datos empíricos	4.28	158	2.81
Con datos simulados	2.42	52	1.54
Total	**3.79**	**210**	**2.69**

Dada la lista de autores y autoras, con mayor número de artículos firmados como autor o autora principal, cabe analizar el tipo de colaboraciones que tienen en sus trabajos, colaboraciones de tipo internacional (autores con filiaciones de diferentes países), nacional (autores con filiaciones del mismo país que el autor principal) o si, por el contrario, no suelen colaborar con otros investigadores, en sus trabajos. Para conocer la tendencia de estos autores y autoras a colaborar de una manera u otra, o incluso a no colaborar con otros autores se realiza un análisis de frecuencias.

Respecto a los autores y autoras con mayor número de artículos, como primeros firmantes, se observa que los autores N. W. Scott, P. K. Crane y D. L. Hart son los tres que con mayor frecuencia, respecto al resto, se decantan por colaborar con autores y autoras de filiaciones de tipo internacional, como se puede ver en la tabla 3.17.

Tabla 3.17

Autores y autoras como primeros firmante en mayor número de trabajos y que colaboran internacionalmente en estos

Autor/a	Nª de artículos	% de sus trabajos
Scott, NW	5	100
Crane, PK	4	57
Hart, DL	3	100

Además, N. W. Scott y D. L. Hart, en el 100% de los trabajos que firman como primero autor, de los seleccionados para este estudio, son colaboraciones de tipo internacional.

Respecto a los autores y autoras con mayor número de artículos, como primeros firmantes, tenemos que los autores P. K. Crane, J. Gómez-Benito, M. D. Hidalgo-Montesinos, W. H. Finch y B. F. French son los cinco que con mayor frecuencia, respecto al resto, sus colaboraciones son nacionales, como se puede ver en la Tabla 3.18.

Tabla 3.18

Autores y autoras como primeros firmante en mayor número de trabajos y que colaboran nacionalmente en estos

Autor/a	Nª de artículos	% de sus trabajos
Crane, PK	3	42.9
Gómez-Benito, J	3	75
Hidalgo-Montesinos, MD	3	75
Finch, WH	3	60
French, BF	3	100

Además, en el 75% de los trabajos que firman como primeras autoras, J. Gómez-Benito e M. D. Hidalgo-Montesinos, de los seleccionados para este estudio, son de tipo nacional, W. H. Finch en el 60% de los trabajos y B. F. French en el 100%.

Si se analiza a los autores, con mayor número de artículos como primer firmante, encontramos a W. H. Finch, con mayor número de artículos, respecto al resto de primeros firmantes, siendo para él, en el 40% de sus trabajos.

3.4.2.5. Número de autores por artículo y su evolución temporal

Respecto a la evolución a lo largo de estos años, de 1990 a 2016, de colaborar con mayor o menor número de investigadores, ésta queda reflejada en la Figura 3.8. La evolución del número de firmantes a lo largo de los últimos años es de tendencia ascendente, los investigadores cada vez tienden más a la colaboración y cooperación en red con otros colegas. Se observa, como del año 2002 al 2004, en los trabajos seleccionados, se produce un gran ascenso en el número de autores y autoras firmantes.

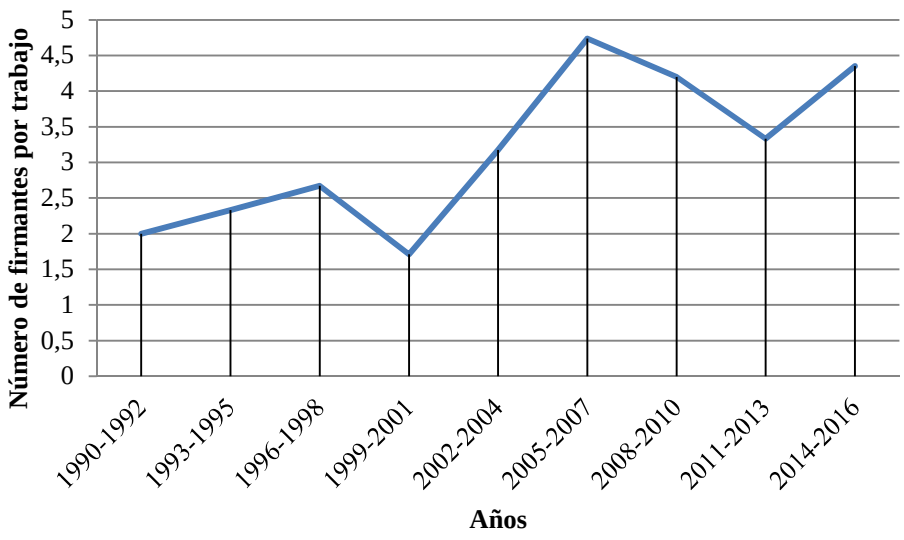


Figura 3.8

Evolución del promedio de firmantes, desde 1990 hasta 2016

Si se tienen en cuenta, únicamente, los trabajos que emplean datos empíricos, se puede observar, igualmente, una tendencia ascendente, situando el promedio, máximo, de autores colaboradores en 5, como se puede ver en la Figura 3.9.

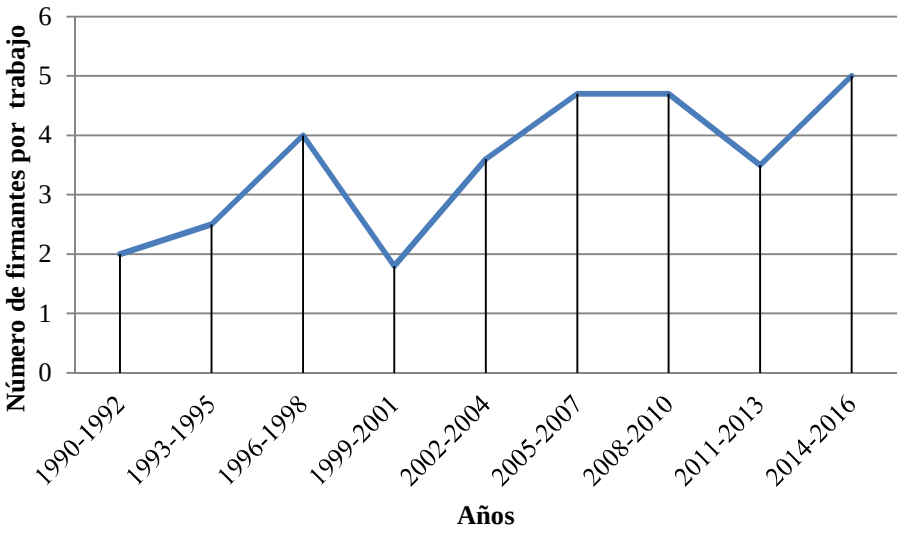


Figura 3.9
Evolución del promedio de firmantes, en los trabajos con datos empíricos, desde 1990 hasta 2016

Por otro lado, si se tienen en cuenta, los trabajos que emplean datos simulados, se puede observar, una tendencia ascendente, como en otros casos, situando el promedio, máximo, de autores colaboradores en 2.6, como se puede ver en la Figura 3.10.

Tal y como se ha comentado anteriormente la tendencia es al alza y cabe esperar que los grupos de investigadores colaboradores sean cada vez mayores. Teniendo en cuenta que el crecimiento no es igual para ambos tipos de estudios, con datos empíricos y simulados.

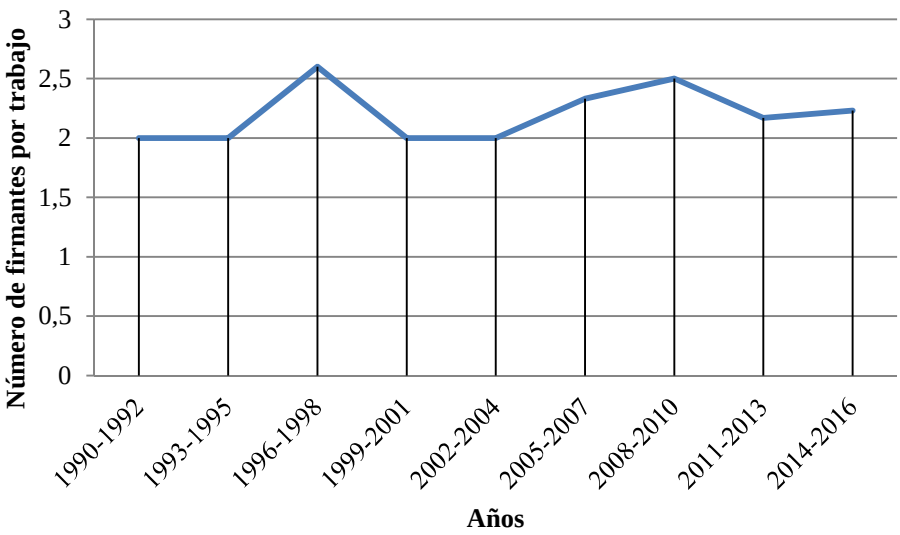


Figura 3.10

Evolución del promedio de firmantes, en los trabajos con datos simulados, desde 1990 hasta 2016

3.4.2.6. Colaboración entre autores

Como es sabido, cada autor y/o autora presenta a una afiliación, ya sea a una universidad, a una empresa, a un organismo público, a un centro de investigación, etc. En este caso, vamos a centrar nuestra atención en el país de filiación de los autores y las autoras, con el fin de indagar sobre el modo de colaboración entre éstos. Es decir, si colaboran, más frecuentemente con compañeros de su mismo país o de distinto, según el tipo de estudio.

Como se puede observar en la Tabla 3.19, sin distinción del tipo de estudio, lo más frecuente son las colaboraciones nacionales, entre los autores y/o las autoras de los trabajos seleccionados. Es indiferente que el estudio emplee datos empíricos como aplicados, los firmantes colaboran con colegas nacionales, con más frecuencia.

Tabla 3.19

Porcentaje de trabajos, según tipo de estudio, en los que los autores colaboran internacionalmente, nacionalmente o no colaboran con otros

	Tipo de estudio			
	Con datos empíricos		Con datos simulados	
Colaboración	Nª de artículos	%	Nª de artículos	%
Internacional	65	41.1	14	26.9
Nacional	84	53.2	32	61.5
No colabora	9	5.7	6	11.5

En el caso de las colaboraciones internacionales y en los estudios con datos empíricos, son los autores y /o las autoras de estos trabajos los que colaboran en mayor medida con colegas internacionales, con un algo más de un 40% de sus trabajos, frente a los que emplean datos simulados, con casi un 27% de sus trabajos.

Un aspecto a destacar, en los trabajos seleccionados, es la preferencia por colaborar con otros y/u otras colegas, más que trabajar de manera individual.

Finalmente, la Figura 3.11 muestra la distribución de colaboraciones con respecto al número de autores y/o autoras. Podemos ver como cuando las colaboraciones son entre pocos autores son en mayor medida colaboraciones nacionales y a medida que aumentan el número de autores el tipo de colaboraciones son mayormente internacionales. En este análisis no se incluyen los autores y autoras que deciden liderar el estudio en solitario y no colaboran con otros u otras.

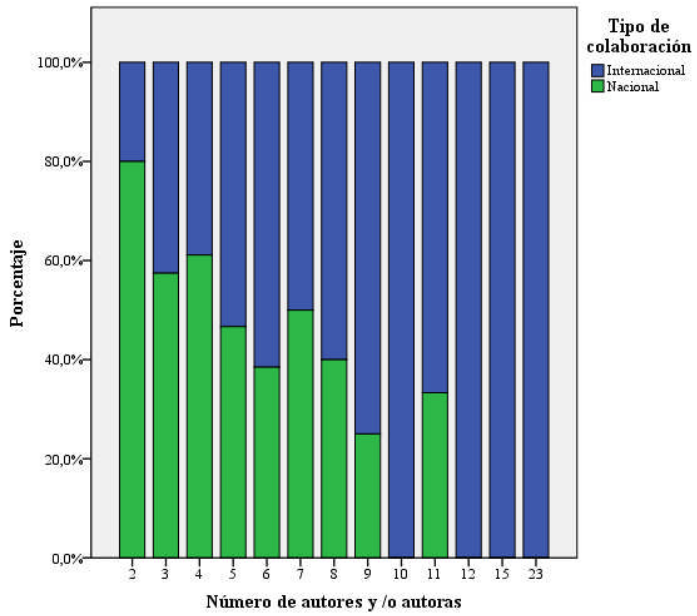


Figura 3.11

Distribución de colaboraciones con respecto al número de autores y/o autoras

3.4.3. Instituciones

Se ha realizado una revisión de las instituciones a las que están vinculados los distintos autores y a autoras, para comprobar cuáles tienen mayor número de producciones. Cada autor y autora pertenece, al menos a una filiación, como se ha dicho anteriormente, la mayor parte de las filiaciones son universidades, aunque también se encuentran organismos públicos y privados. La mayor parte de los autores y autoras presentan una filiación, pero en otros casos presentan dos y hasta tres filiaciones.

Se ha encontrado un total de 328 filiaciones distintas. Cada autor y autora pertenece, al menos, a una institución, 44 de ellos y ellas a dos filiaciones, y 3 de éstos pertenecen a 3 filiaciones.

Primero de todo, se va a analizar la producción de estas instituciones respecto al tema objeto de estudio. Para ello, se ha registrado todas las filiaciones, en término de número de participaciones en artículos.

En la Tabla 3.20 se muestra el listado de las 20 instituciones, que más participan en los estudios seleccionados, para ello se ha eliminado a las instituciones con menos de 8 participaciones.

Tabla 3.20

Listados de instituciones de mayor a menor participación en los estudios seleccionados

Institución	Frecuencia	Porcentaje
University of Washington	47	5.6
University of British Columbia	26	3.1
University of Massachusetts	19	2.2
University of Barcelona	16	1.9
The Netherlands Cancer Institute	15	1.8
University of Copenhagen	15	1.8
Shiraz University of Medical Sciences	13	1.5
National Board of Medical Examiners	12	1.4
University of Aberdeen	12	1.4
University of California	12	1.4
Pennsylvania State University	11	1.3
Ludwig-Maximilians-University	9	1.1
University Medical Center	9	1.1
University of Granada	9	1.1
University of Murcia	9	1.1
Washington State University	9	1.1
Columbia University	8	.9
Johns Hopkins Bloomberg School of Public Health	8	.9
Mayo Clinic	8	.9
University of North Carolina at Chapel Hill	8	.9

En la Figura 3.12 puede observarse gráficamente el número de trabajos por institución.

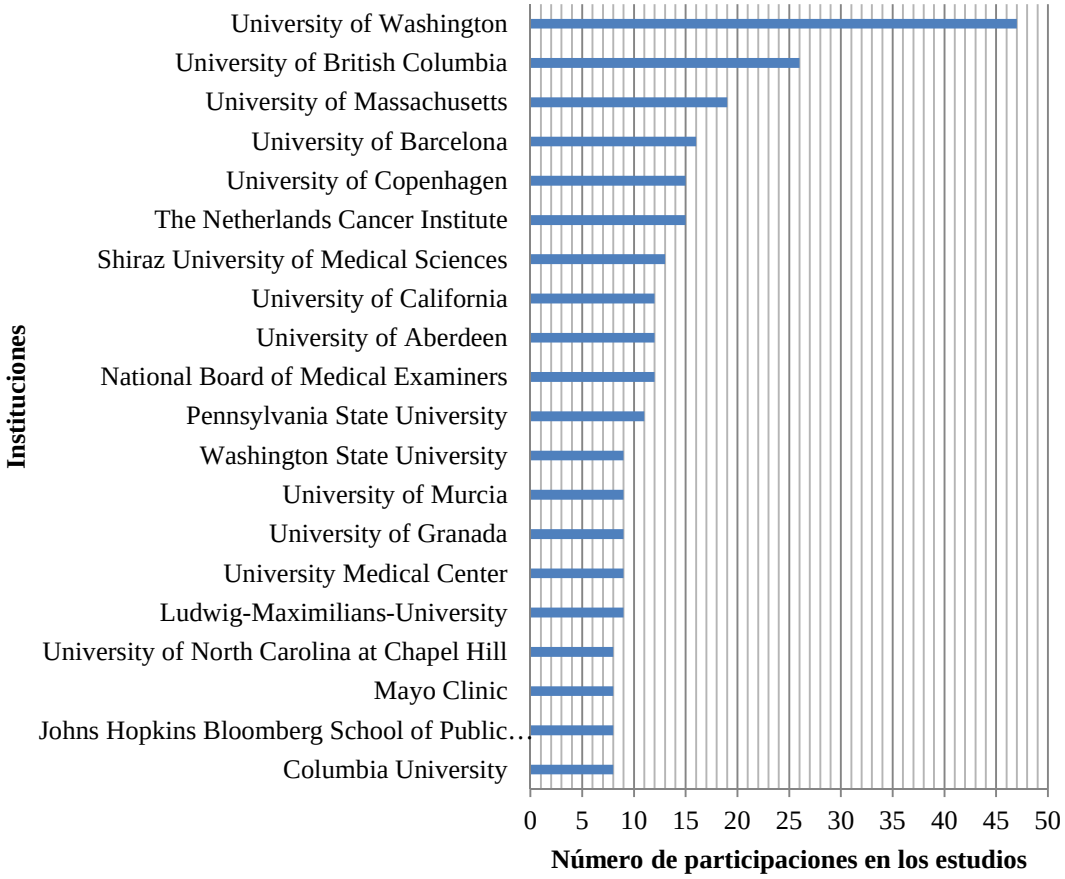


Figura 3.12

Listados de instituciones de mayor a menor participación en los estudios seleccionados

Liderando la lista encontramos a la Universidad de Washington con su participación en 47 trabajos, la Universidad de British Columbia con 26 participaciones, la Universidad de Massachusetts con 19 participaciones y la Universidad de Barcelona con 16.

Y finalmente, se ha focalizado la atención en las instituciones a las que pertenecen los autores que figuran como primeros firmantes. En la tabla 3.21, tras excluir a los autores con menos de tres publicaciones en esta clasificación, se puede ver la ordenación de éstos según tengan más o menos publicaciones firmadas como primer autor de trabajos que emplean datos empíricos.

Tabla 3.21

Filiaciones de los primeros autores de estudios empíricos

Institución	Nª de artículos	%
University of Washington	11	7.0
University of Aberdeen	7	4.4
University of British Columbia	6	3.8
University of Massachusetts	5	3.2
Focus On Therapeutic Outcomes	3	1.9
University of Granada	3	1.9
University of the Basque Country	3	1.9

Respecto a los centros a los que están adscritos los primeros autores de los trabajos que emplean datos empíricos, cabe otorgar el primer lugar a la Universidad de Washington, con 11 trabajos firmados en primer lugar. Seguido de esta universidad se sitúan la Universidad de Aberdeen con 7 trabajos, la Universidad de Columbia Británica y la de Massachusetts, con 6 y 5 trabajos, respectivamente.

En la Tabla 3.22, tras excluir a los autores con menos de tres publicaciones en esta clasificación, se puede ver la ordenación de éstos según tengan más o menos publicaciones firmadas como primer autor de trabajos que emplean datos simulados.

Tabla 3.22

Filiaciones de los primeros autores de estudios simulados

Institución	Nª de artículos	%
Ball State University	4	7.7
University of Murcia	4	7.7
Educational Testing Service	3	5.8
Ludwig-Maximilians-University	3	5.8
University of Barcelona	3	5.8
University of Massachusetts	3	5.8

Si se habla de los centros a los que están adscritos los primeros autores de los trabajos que emplean datos simulados, cabe otorgar los dos primeros lugares a la Universidad Estatal Ball y a la Universidad de Murcia, con 4 trabajos firmados, ambos, en primer lugar. Seguido de estas dos universidades se sitúan la Universidad de Barcelona, de Massachusetts, de Munich y el Educational Testing Service con 3 trabajos.

3.4.4. Países

Cada autor y autora pertenecen a una filiación o varias y éstas están afincadas en países. Se ha encontrado un total de 49 países distintos entre todas las filiaciones de los autores y autoras de los trabajos seleccionados.

Primero de todo, se va a analizar la participación de estos países en los trabajos seleccionados. Para ello, se ha registrado todos los países, en término de número de participaciones en artículos.

En la Tabla 3.23 se muestra el listado de los doce países, que más participan en los estudios seleccionados, para ello se ha eliminado de la lista a los países con menos de quince participaciones.

Tabla 3.23

Listados de países de mayor a menor participación en los estudios seleccionados

País	Nª de artículos	%
Estados Unidos	338	40.1
Canadá	68	8.1
España	68	8.1
Dinamarca	46	5.5
Reino Unido	46	5.5
Países Bajos	41	4.9
Alemania	40	4.7
Francia	23	2.7
Bélgica	21	2.5
Australia	16	1.9
Turquía	16	1.9
Irán	15	1.8

En la Figura 3.13 puede observarse de países con mayor participación en los estudios seleccionados.

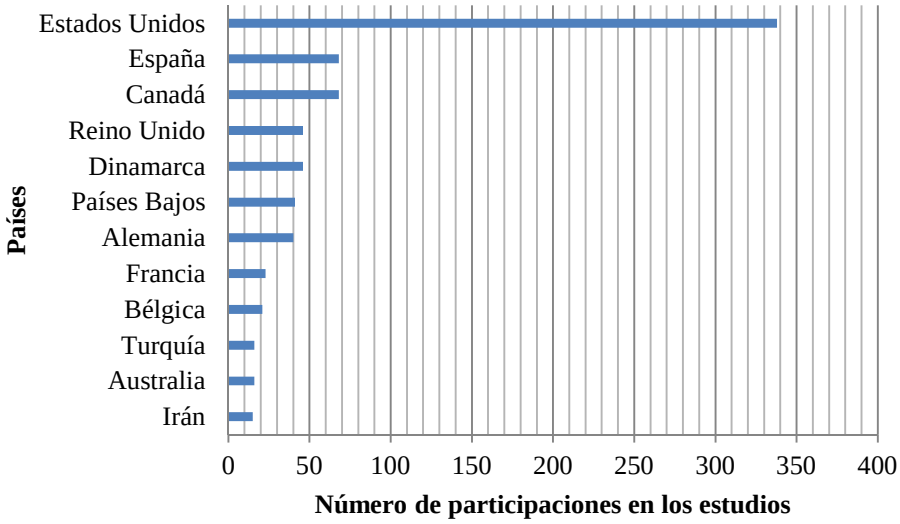


Figura 3.13

Listados de países de mayor a menor participación en los estudios seleccionados

Liderando la lista se encuentra a Estados Unidos con 338 participaciones, seguido de Canadá y España con 68 participaciones, tras éstos, Dinamarca y Reino Unido con 46 participaciones, Países Bajos con 41 y Alemania con 40 participaciones en los estudios seleccionados.

Y finalmente, cabe centrarse en los países a los que pertenecen los autores que figuran como primeros firmantes. En la Tabla 3.24, tras excluir a los países con menos de seis publicaciones en esta clasificación, se puede ver la ordenación de éstos según tengan más o menos publicaciones, como país del primer autor, de trabajos que emplean datos empíricos.

Tabla 3.24

Filiaciones de los primeros autores de estudios empíricos

País	Nª de artículos	%
Estados Unidos	67	42.4
Canadá	12	7.6
España	12	7.6
Reino Unido	11	7.0
Dinamarca	7	4.4
Alemania	6	3.8

Particularmente en los trabajos encontrados, que emplean datos empíricos, los resultados no difieren mucho de la clasificación general. Continuamos encontrando en primer lugar a Estados Unidos, con 67 apariciones como filiación del primer autor, a Canadá y a España, en segundo lugar, con 12 apariciones y en tercer lugar a Reino Unido con 11.

En la tabla 3.25, tras excluir a los países con menos de cuatro publicaciones en esta clasificación, se puede ver la ordenación de éstos según tengan más o menos publicaciones como país del primer autor de trabajos que emplean datos simulados.

Tabla 3.25

Filiaciones de los primeros autores de estudios simulados

País	Nª de artículos	%
Estados Unidos	26	50
España	9	17.3
Alemania	4	7.7

En este caso, los resultados difieren un poco más de la clasificación general, respecto a los trabajos que emplean datos empíricos. Se continúa encontrando, en primer lugar, a Estados Unidos, con 26 apariciones como filiación del primer autor que firma los trabajos, que emplean datos simulados, destacando que esto corresponde al

50% de los trabajos seleccionados con estas características, seguido de España, con 9 apariciones y de Alemania con 4.

3.4.5. Revistas

Los 210 artículos que han sido analizados en este trabajo, han sido publicados en 103 revistas.

Primero de todo, se muestra el listado de las revistas que más trabajos han publicado, del tema seleccionado. Para ello, se ha registrado todas las revistas, en término de número de participaciones en artículos.

En la Tabla 3.26 se muestra el listado de las nueve revistas, que más artículos, de los seleccionados, han publicado. Para ello, se ha eliminado de la lista las revistas que han publicado menos de cinco de los artículos seleccionados.

Tabla 3.26

Revistas con mayor producción en los artículos seleccionados

Revista	Nª de artículos	%
Quality of Life Research	25	11.9
Educational and Psychological Measurement	17	8.1
International Journal of Testing	12	5.7
Journal of Educational Measurement	10	4.8
European Journal of Psychological Assessment	6	2.9
Quality y Quantity	6	2.9
Applied Measurement in Education	5	2.4
Applied Psychological Measurement	5	2.4
Journal of Educational and Behavioral Statistics	5	2.4

Como se puede observar, de los trabajos que han sido seleccionados, la revista que más artículos ha publicado es Quality of Life Research, con 25 artículos publicados, seguida de Educational and Psychological Measurement con 17, International Journal of Testing con 12 y Journal of Educational Measurement con 10.

En la Tabla 3.27 se muestra el listado de las siete revistas, que más artículos, de los que emplean datos empíricos, han publicado. Para ello, se ha eliminado de la lista las revistas que han publicado menos de tres de los artículos seleccionados.

Tabla 3.27

Revistas con mayor producción en los artículos seleccionados que emplean datos empíricos

Revista	Nª de artículos	%
Quality of Life Research	25	15.8
International Journal of Testing	12	7.6
European Journal of Psychological Assessment	4	2.5
Journal of the International Neuropsychological Society	4	2.5
Language Testing	4	2.5
Journal of Educational Measurement	3	1.9
Journal of Psychoeducational Assessment	3	1.9

En el caso de seleccionar solo los artículos que han trabajado con datos empíricos, como se puede ver en la Tabla 3.26, es la misma revista, *Quality of Life Research*, la que se encuentra en primer lugar con 25 artículos publicados. Seguida de *International Journal of Testing* con 12, y *European Journal of Psychological Assessment*, *Journal of the International Neuropsychological Society* y *Language Testing*, con 4 artículos publicados. En este caso, es diferente el interés de las revistas, por lo que son otras las que se interesan más, por publicar artículos que empleen datos empíricos.

En la tabla 3.28 se muestra el listado de las ocho revistas, que más artículos, de los que emplean datos simulados, han publicado. Para ello, se ha eliminado de la lista las revistas que han publicado menos de dos de los artículos seleccionados.

Tabla 3.28

Revistas con mayor producción en los artículos seleccionados que emplean datos simulados

Revista	Nª de artículos	%
Educational and Psychological Measurement	15	28.8
Journal of Educational Measurement	7	13.5
Quality & Quantity	6	11.5
Applied Psychological Measurement	5	9.6
Applied Measurement in Education	3	5.8
Journal of Educational and Behavioral Statistics	3	5.8
European Journal of Psychological Assessment	2	3.8
Psychometrika	2	3.8

Y si se tienen en cuenta solo los trabajos que emplean datos simulados, se tiene, como se puede ver en la tabla 18, que es la revista Educational and Psychological Measurement, la más interesada, con 15 trabajos publicados. Seguidos de la revista Journal of Educational Measurement con 7, la revista Quality & Quantity con 6 y la revista Applied Psychological Measurement con 5.

3.4.5.1. Productividad de las revistas: Ley de Bradford.

Con el fin de analizar la productividad de las revistas que han publicado artículos en la citada área de estudio, se aplica la ley de Bradford.

Los 220 artículos analizados en el presente estudio fueron publicados en 103 revistas de diversos campos científicos, como se puede ver en la Tabla 3.29.

Tabla 3.29

Dispersión de la literatura científica en el área de estudio

Nº de revistas	***Nº de artículos***	***Revistas Acumuladas***	***Artículos acumulados***	***Ln (Revistas acumuladas)***
1	75	1	75	0
2	16	3	107	1.10
3	1	6	110	1.79
4	2	10	118	2.30
5	2	15	128	2.71
6	3	21	146	3.04
10	1	31	156	3.43
12	1	43	168	3.76
17	1	60	185	4.09
25	1	85	210	4.44

Después de aplicar la ley de Bradford de dispersión con respecto a la variable "revista", se definieron tres zonas concéntricas. El núcleo ha contenido un total de 42 artículos de los 210 totales, el 20% de los documentos, que fueron publicados en dos revistas, *Quality of Life Research*, con 25 artículos publicados y *Educational and Psychological Measurement* con 17. La zona 1 ha contenido 45 artículos, el 21,4% de los documentos totales, que fueron publicados en seis revistas, *International Journal of Testing* (con 12 artículos publicados), *Journal of Educational Measurement* (con 10 artículos), *European Journal of Psychological Assessment* (con 6 artículos), *Journal of Educational and Behavioral Statistics* (con 6 artículos), *Quality & Quantity: International Journal of Methodology* (con 6 artículos) y *Applied Measurement in Education* y *Applied Psychological Measurement* (con 5 artículos). La zona 2 ha contenido 50 artículos, el 23,8% de los documentos totales, que fueron publicados en veintidós revistas, publicando cada una de ellas de 1 a 5 artículos. Y la zona 3 ha contenido 73 artículos, el 34,8% de los documentos totales, todas las revistas pertenecientes a esta zona publican 1 artículo, como se puede ver en la Tabla 3.30.

Tabla 3.30

Zonas Bradford

Zona Bradford	***Nº de artículos***	***Porcentaje de artículos***
Núcleo	2	20
Zona 1	6	21.4
Zona 2	22	23.8
Zona 3	73	34.8

En la Figura 3.14, se pueden ver de manera más gráfica las zonas de dispersión de Bradford.

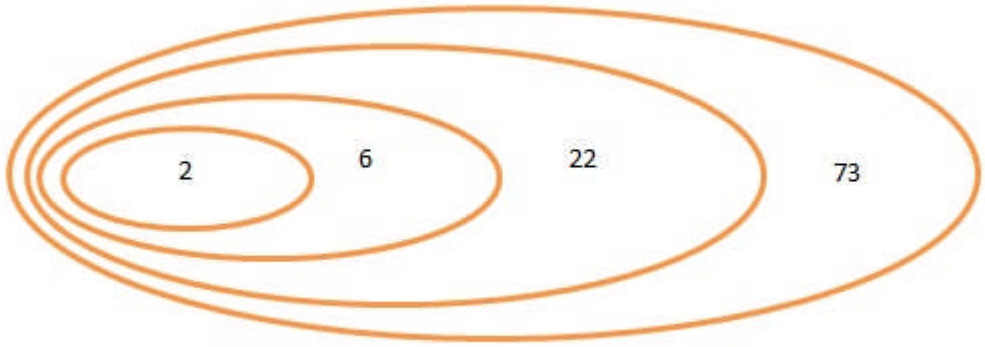


Figura 3.14

Zonas Bradford

Más gráficamente, se puede observar, en la Figura 3.15, las revistas con mayor producción en los artículos seleccionados.

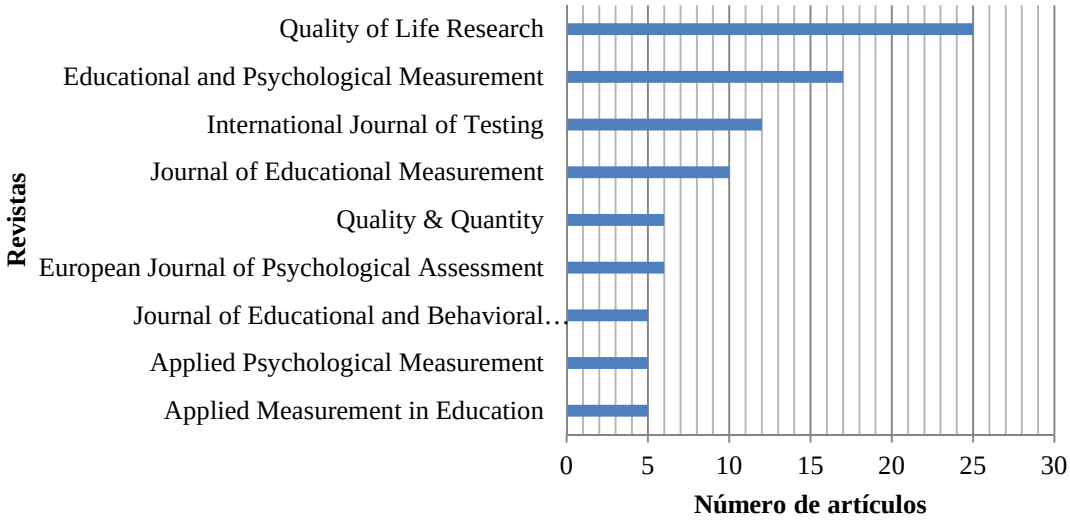


Figura 3.15

Revistas con mayor producción en los artículos seleccionados

Como se puede ver, las revistas que han publicado la mayor proporción de los artículos, en los que se emplea la RL como técnica de detección del DIF, ya sea con datos empíricos o simulados, son *Quality of Life Research*, con 25 artículos publicados, representado casi el 12% de los artículos publicados y *Educational and Psychological Measurement* con 17, con un 8%.

Estas revistas, junto con *International Journal of Testing* y *Journal of Educational Measurement* representan casi un tercio de los artículos recopilados.

3.5. Conclusiones

A lo largo de este capítulo se ha realizado un análisis de la producción científica respecto al uso de la RL como técnica de detección del DIF, ya sea en la práctica, con datos empíricos, o en la investigación, con datos simulados. Sin embargo, debe

recordarse que una mayor productividad no implica necesariamente mayor calidad (Rey-Rocha, Martin-Sempere y Garzón, 2002). Teniendo en cuenta esta cuestión y volviendo a los objetivos específicos del estudio, se pueden extraer las siguientes conclusiones.

El número de trabajos aplicados en el que se usa RL como técnica para detectar el DIF ha ido aumentando progresivamente a lo largo del tiempo. Parece resultar evidente el mayor interés en la aplicación de esta técnica en la práctica. Por otro lado, encontramos, en los últimos años, un mayor número de trabajos que estudian la eficacia de esta técnica en cuanto a su capacidad para detectar de manera correcta el DIF.

Autores como, P. K. Crane y L. E. Gibbons son los autores más productivos en la publicación de trabajos que emplean datos empíricos en la aplicación de la técnica para la detección del DIF; J. Gómez-Benito y M. D. Hidalgo-Montesinos son las autoras con mayor número de publicaciones al emplear datos simulados. Siendo, B. D. Zumbo uno de los autores más productivo en el estudio de la RL como técnica de detección del DIF, ya sea en estudios que emplean datos simulados como datos empíricos.

Debido a la gran variedad de contextos en los que se puede aplicar la RL, como técnica de detección del DIF, encontramos un gran número de autores y autoras en este estudio, en total 569. Un 82.2% de éstos, participan en un solo artículo, y un únicamente 3.8% contribuyen con más de 5 artículos. Cabe esperar que los autores y autoras que contribuyen en más de un trabajo sea en el estudio de la RL con datos simulados, ya que quien la emplea con datos empíricos, la emplea como herramienta para la identificación de ítems que funcionan diferencialmente, no como objeto de estudio en sí misma.

En términos generales, se observa que la tendencia es colaborar con otros autores y autoras. Un 92.9% de los autores colaboran con otros, siendo lo más frecuente colaborar en parejas, con un 28.6%. Cuando se trata de trabajos que emplean datos empíricos, lo más habitual es encontrarse con artículos firmados por dos, tres y cuatro autores, con un 62.1%; en cambio en los trabajos que emplean datos simulados, lo más común es encontrarse con artículos firmados por dos investigadores y/o investigadoras, con 57.7%.

Como se ha podido ver, la preferencia a la hora de liderar trabajos en solitario o colaborar con más autores es bastante distinta según el tipo de estudio que se lleve a cabo. Si se trata de estudios con datos empíricos, lo más frecuente es colaborar con mayor número de autores y autoras que si se trata de estudios con datos simulados. Podría dar explicación a este suceso, el hecho de necesitar, para los trabajos que manejan datos empíricos, la aplicación de la prueba de medición a una muestra o población, siendo más beneficiosas, para el resultado, las colaboraciones múltiples. En cambio, esta cifra se ve reducida en los trabajos con datos simulados, ya que no es necesaria la recogida de datos empíricos y el trabajo está centrado en el manejo y análisis de datos mediante un soporte informático.

Respecto a los autores y autoras con mayor número de artículos, como primeros firmantes, son N. W. Scott, P. K. Crane y D. L. Hart los tres autores que con mayor frecuencia, respecto al resto, colaboran con autores y autoras de filiaciones de tipo internacional. El otro caso, son P. K. Crane, J. Gómez-Benito, M. D. Hidalgo-Montesinos, W. H. Finch y B. F. French, los cinco que con mayor frecuencia, respecto al resto, colaboran con autores y autoras de filiaciones de tipo nacional.

La evolución del número de firmantes a lo largo de los últimos años muestra una tendencia ascendente, los investigadores, cada vez más, tienden a la colaboración y cooperación en red con otros colegas. Se observa, como del año 2002 al 2004, en los trabajos seleccionados, se produce un gran ascenso en el número de autores y autoras firmantes.

El número de colaboraciones internacionales aumenta a medida que aumenta el número de autores y autoras firmantes, lo es sistemáticamente en el caso de artículos entre dos y cinco firmantes. Sin embargo, los resultados de los artículos con seis, siete y doce firmantes no pueden ser interpretados, ya que corresponden a sólo dos publicaciones, en el caso de seis firmantes y un solo artículo, para grupos de siete y doce autores y autoras firmantes.

Lo más frecuente son las colaboraciones nacionales. En el caso de las colaboraciones internacionales y en los estudios con datos empíricos, son los autores y/o las autoras de estos trabajos los que colaboran en mayor medida con colegas internacionales, con algo más de un 40% de sus trabajos, frente a los que emplean datos simulados, con casi un 27% de sus trabajos.

Respecto a las instituciones a las que están filiados los autores de los trabajos seleccionados, encontramos 328 instituciones distintas. Liderando la lista, con mayor participación en los trabajos seleccionados, encontramos a la Universidad de Washington, la Universidad de British Columbia, la Universidad de Massachusetts y la Universidad de Barcelona. Respecto a las instituciones a las que pertenecen los primeros firmantes de trabajos que emplean datos empíricos, encontramos que las que

tienen mayor número de trabajos firmados, en primer lugar, son la Universidad de Washington, Universidad de Aberdeen, la Universidad de Columbia Británica y la de Massachusetts. Respecto a los trabajos que emplean datos simulados, encontramos que las que tienen mayor número de trabajos firmados, en primer lugar, son la Universidad Estatal Ball, la Universidad de Murcia, la Universidad de Barcelona, de Massachusetts, y de Munich.

Son 49 los países a los que pertenecen los autores y autoras que participan en los trabajos seleccionados. Liderando la lista de países con mayor número de participaciones, se encuentran los Estados Unidos, Canadá y España. Si se tiene en cuenta a los países de los primeros firmantes de trabajos que emplean datos empíricos, los tres primeros países son, igualmente, Estados Unidos, Canadá y España; en cambio si se tiene en cuenta a los países de los primeros firmantes de trabajos que emplean datos simulados, los primeros países son Estados Unidos, España y Alemania.

Son 103 revistas diferentes las que publican los trabajos seleccionados. La revista que más artículos ha publicado es *Quality of Life Research*, seguida de *Educational and Psychological Measurement*, *International Journal of Testing y Journal of Educational Measurement*. Las revistas que más trabajos han publicado, empleando datos empíricos, son *Quality of Life Research*, *International Journal of Testing*, *European Journal of Psychological Assessment*, *Journal of the International Neuropsychological Society* y *Language Testing.* Las revistas que más trabajos han publicado, empleando datos simulados, son *Educational and Psychological Measurement, Journal of Educational Measurement, Quality & Quantity* y *Applied Psychological Measurement*.

Después de aplicar la ley de Bradford de dispersión con respecto a la variable "revista", se definieron tres zonas concéntricas. El núcleo ha contenido un total de 42 artículos de los 210 totales, el 20% de los documentos, que fueron publicados en dos revistas, *Quality of Life Research*, con 25 artículos publicados y *Educational and Psychological Measurement* con 17. La zona 1 ha contenido 45 artículos, el 21.4% de los documentos totales, que fueron publicados en seis revistas, la zona 2 ha contenido 50 artículos, el 23.8% de los documentos totales, que fueron publicados en veintidós revistas, y la zona 3 ha contenido 73 artículos, el 34.8% de los documentos totales.

CAPÍTULO 4

REGRESIÓN LOGÍSTICA EN LA DETECCIÓN DEL DIF: UNA REVISIÓN SISTEMÁTICA

Tras realizar un análisis de la producción científica respecto al uso de la RL como técnica de detección del DIF, y observar el interés existente por evaluar su eficacia mediante estudios de simulación, se hace imprescindible realizar una revisión sistemática de los trabajos de simulación encontrados, con el fin de obtener un mapa de los estudios realizados, que nos agrupe la información existente sobre el funcionamiento de la RL bajo distintas condiciones en la detección de los ítems con DIF.

La eficacia en la detección correcta de ítems con DIF depende de diversas variables, tales como el tamaño muestral de los grupos que se comparan, la longitud del test, las características de los ítems, el formato de respuesta del ítem, el tipo de DIF (uniforme o no uniforme) o la cantidad de ítems con DIF en el test, por citar algunas de las más relevantes. De este modo, encontramos numerosos estudios que analizan la efectividad y eficacia de la RL en la identificación del DIF, usando estudios de simulación Monte Carlo. La simulación Monte Carlo es un método muy usado para la investigación ya que permite conocer el funcionamiento del modelo o prueba estadística

bajo distintas condiciones establecidas y controladas (Cohen, Kane y Kim, 2001; Harwell, Stone, Hsu y Kirisci, 1996).

Por otro lado, la revisión sistemática de la literatura científica es una herramienta metodológica muy útil para obtener evidencia científica que apoye las decisiones tanto en el contexto aplicado como en el de la investigación. Las principales ventajas de las revisión sistemática son que facilita la identificación de lagunas en un ámbito de investigación, facilita la toma de decisiones dado que integra de forma eficiente la información/resultados obtenidos en dicha área, y en definitiva es un método que permite integrar, evaluar y sintetizar las investigaciones llevadas a cabo en un ámbito disciplinar (Cooper, 2016).

En esta revisión sistemática sobre la RL se han incluido un total de 50 artículos publicados en revistas científicas que analizan el funcionamiento de la técnica mediante la simulación de datos cuando ésta se utiliza en la detección del DIF. Con esta revisión se ha pretendido, por un lado, analizar en profundidad las condiciones de simulación en las que se ha estudiado el funcionamiento de la RL hasta la actualidad y, por otro lado, conocer posibles brechas en el estudio de la técnica que deberían ser abordadas en futuros estudios de simulación.

4.1. Método

4.1.1. Obtención de los documentos

Para la identificación de documentos se ha empleado las bases de datos PsycInfo, Education Resources Information Center (ERIC) y Web of Science, por ser grandes bases de datos de la educación, la psicología y de información científica.

Para identificar los trabajos se ha tenido en cuenta los siguientes criterios de inclusión:

d) *Terminología:* Para la obtención de los documentos se ha empleado los términos "*Logistic Regression technique"* y "*Differential Item Functioning"*, debiendo ser mencionados en el título, resumen o en las palabras clave, siguiendo esta estrategia de búsqueda: ("differential item functioning" OR DIF) AND ("logistic regression" OR LR).

e) *Período de tiempo:* Los estudios son identificados el 9 de enero de 2017, cubriendo el período entre 1990 y 2016.

f) *Tipo de documento:* La búsqueda se limitó a artículos de revistas escritos en inglés, eliminándose los artículos en otros idiomas, las tesis, libros, capítulos, etc.

Tras eliminar los artículos duplicados, se obtuvieron 352 artículos publicados en revistas científicas.

4.1.2. Criterios de inclusión y exclusión de documentos

Se incluyeron únicamente los estudios de simulación que pusieron a prueba la RL como técnica de detección de DIF en términos de tasa de error Tipo I y/o potencia estadística, excluyendo las revisiones teóricas y los estudios empíricos que únicamente aplicaron la técnica en datos reales. Siguiendo esta estrategia, y tras eliminar dos artículos de simulación que no informaban de las tasas de detección, finalmente pasaron a formar parte de la revisión sistemática 50 artículos.

4.1.3. Extracción de datos

Las variables que se tuvieron en cuenta fueron:

a) Método de detección del DIF con el que se compara la RL (por ejemplo, Mantel-Haenszel)
b) Tasa de detección estudiada (es decir, error Tipo I y/o potencia)
c) Modelo de simulación de datos (por ejemplo, modelo logístico de tres parámetros)
d) Parámetros de los ítems (por ejemplo, discriminación o dificultad)
e) Longitud del test
f) Tamaño muestral del grupo de referencia
g) Tamaño muestral del grupo focal
h) Ratio entre tamaños muestrales (es decir, R/F)
i) Cantidad de impacto (es decir, diferencia en términos de desviaciones típicas entre las medias del nivel de habilidad del grupo focal y de referencia)

j) Tipo de DIF (por ejemplo, uniforme)

k) Cantidad de DIF (por ejemplo, áreas entre las curvas características de los ítems)

l) Porcentaje de ítems con DIF en el criterio de equiparación

m) Procedimiento de purificación empleado.

Algunas variables potencialmente de interés finalmente no se codificaron dado que fueron muy escasos los artículos que las tuvieron en cuenta (por ejemplo, distribución de las puntuaciones, porcentaje de datos perdidos, etc.).

Todas estas variables quedaron recogidas en un libro de codificación, con la finalidad de que los codificadores siguieran la misma estrategia a la hora de extraer los datos de los estudios. Dos codificadores (M. D. L. y M. D. H.) extrajeron independientemente la información del 10% de los artículos, coincidiendo en un 88.4% de los casos, y llegando a un acuerdo absoluto después de discutir los casos de discordancia (11.6%). Tras esta primera fase, un único codificador (M. D. L.) extrajo la información del resto de artículos, consultando al segundo codificador en caso de duda.

4.2. Resultados

Se incluyeron en la revisión sistemática un total de 50 artículos, 41 centrados en el estudio de la LR en ítems dicotómicos (82%) y 9 en ítems politómicos (18%). Los años de publicación oscilan entre el año 1990 y 2016, aunque el mayor número de artículos científicos se concentra en el año 2009 con un total de 6 artículos (12%).

4.2.1. Técnica de comparación

En prácticamente todos los estudios la RL es comparada con otros procedimientos de detección del DIF (o incluso con variaciones de la misma técnica) con el fin de determinar cuál presenta una mayor eficacia, en términos de tasa de error Tipo I y/o potencia estadística, bajo determinadas condiciones de simulación. En los 41 artículos centrados en ítems dicotómicos, la técnica más comparada con la RL es el estadístico Mantel-Haenszel (Holland y Thayer, 1988; n = 24), mientras que en el caso de ítems politómicos, la técnica de comparación más utilizada es IRT likelihood-ratio test (Thissen, 2001; Thissen, Steinberg y Wainer, 1988; n = 2).

4.2.2. Tasa de detección estudiada

En cuanto a tasas de detección, en el estudio de la eficacia de cualquier técnica de detección del DIF pueden analizarse tanto la tasa de error Tipo I (detectar un ítem con DIF cuando realmente no lo presenta) como la potencia estadística (detectar un ítem con DIF cuando realmente funciona diferencialmente). Una técnica será eficaz si controla la tasa de error Tipo I y a la vez es capaz de clasificar con DIF los ítems que lo presentan.

De los 50 estudios incluidos en la revisión sistemática, 40 (80%) analizan simultáneamente la tasa de error y la potencia (32 en ítems dicotómicos y 8 en politómicos), mientras que son 8 (16%) los que analizan únicamente la tasa de error Tipo I (todos ellos en ítems dicotómicos) y 2 (4%) solamente la potencia tanto en ítems dicotómicos como politómicos.

4.2.3. Modelo de simulación de datos

Del modelo de simulación empleado para generar las respuestas de los sujetos a los ítems se desprende en qué tipo de ítems se aplica la técnica. El modelo logístico de 3 parámetros es el más empleado para simular las respuestas en el caso de ítems dicotómicos (n = 19), y el modelo de respuesta graduada en el caso de ítems politómicos (n = 6).

4.2.4. Parámetros de los ítems

Numerosos estudios sobre DIF indican que los valores de los parámetros de los ítems, tanto de los que muestran DIF como de los insesgados, pueden explicar las tasas de error y de potencia de las técnicas de detección del DIF (DeMars, 2011; Li, Brooks y Johanson, 2012; Jin, Myers y Ahn, 2014; Narayanan y Swaminathan, 1994, 1996; Rogers y Swaminathan, 1993; Zwick, 1990). En el caso de ítems dicotómicos, los estudios incluidos en la presente revisión analizan diversos valores de parámetros de discriminación y dificultad de los ítems con DIF (baja, moderada y alta), en combinación con ítems del test caracterizados por presentar valores también diversos de los parámetros de discriminación y dificultad. El valor del parámetro de pseudoazar más ampliamente estudiado es de c = 0.20. En los estudios de simulación con ítems politómicos, también se observa esta misma tendencia con valores bajos, moderados y altos tanto de los parámetros de pendiente, como de los parámetros de localización.

4.2.5. Longitud del test

Aunque la relación entre longitud del test y fiabilidad es compleja, es razonable afirmar que deberían obtenerse puntuaciones de equiparación más fiables a medida que aumenta la longitud del test, dado que con un número más elevado de ítems podemos conseguir una medición más precisa. En los estudios de DIF, diversos trabajos han encontrado relación entre la longitud del test y las tasas de detección, tanto del error Tipo I (Hidalgo, López-Martínez, Gómez-Benito y Guilera, 2016; Paek y Wilson, 2011; Wang y Yeh, 2003), como de la potencia estadística (Hidalgo, López-Martínez, Gómez-Benito y Guilera, 2016; Paek y Wilson, 2011; Wang y Yeh, 2003).

En la presente revisión sistemática, la longitud del test más ampliamente estudiada en ítems dicotómicos es aquella que contempla 20 ítems (n = 20), lo que puede considerarse una longitud corta, siendo de 10 ítems la longitud mínima analizada (Gómez-Benito, Hidalgo, y Padilla, 2009) y de 100 la máxima (Herrera y Gómez, 2008; Li, 2014). En el caso de ítems politómicos, el número de ítems del test oscila entre 2 ítems (Scott, et al., 2009) y 30 ítems (Hidalgo y Gómez-Benito, 2003, 2006), siendo de alrededor de 25 ítems la longitud del test más escogida (n = 3). En ambos formatos de respuesta de los ítems, son escasos los estudios que analizan el funcionamiento de la RL en tests muy cortos (por ejemplo, en menos de 20 ítems) o largos (por ejemplo, en más de 60 ítems).

4.2.6. Tamaño muestral de R y F y ratio R/F

Respecto al tamaño muestral del grupo de referencia en ítems dicotómicos, se constata que mayoritariamente se emplean grupos de 1,000 sujetos (*n* = 30), con un rango que oscila entre 50 (Li, 2014) y 10,000 (French y Finch, 2010). Y también en el grupo focal los tamaños oscilan entre 50 y 10,000, concentrándose la mayoría en el estudio de grupos de 1,000 sujetos (*n* = 25). En el caso de los estudios que emplean ítems politómicos, el tamaño muestral del grupo de referencia y focal más estudiado es de 500 sujetos (*n* = 7 en ambos), con valores que van desde 100 sujetos (Scott, et al., 2009) hasta 3,200 sujetos en el grupo de referencia (Kristjansson, Aylesworth, McDowell, y Zumbo, 2005) y hasta 2,000 en el grupo focal (French y Miller, 1996; Hidalgo y Gómez-Benito, 2003; Kristjansson, Aylesworth, McDowell, y Zumbo, 2005).

Que el tamaño muestral del grupo focal y de referencia no coincidan (y normalmente es el grupo focal el de menor tamaño muestral) es algo bastante frecuente en los estudios de DIF, y como apuntan varios autores (por ejemplo, Penfield, 2001) esta situación puede afectar a los resultados del análisis del DIF. Así, por ejemplo, Tian (1999) indicó que tamaños muestrales desiguales produjeron una disminución de la potencia en algunos procedimientos DIF. En ítems dicotómicos, la mayoría de estudios incluidos en la revisión contemplan, entre las condiciones de simulación analizadas, la igualdad de tamaños muestrales entre el grupo focal y de referencia (ratio 1:1; *n* = 33), pero también analizan diferencias tan dispares como que el grupo focal presente un tamaño muestral 10 veces inferior (ratio 10:1) al del grupo de referencia (French y Maller, 2007; Magis, Tuerlinckx, y De Boeck, 2015). En el caso de los estudios con ítems politómicos, todos contemplan la igualdad de tamaños muestrales en sus

condiciones de simulación, a excepción del trabajo de Goméz-Benito, Hidalgo, y Zumbo (2013), y la diferencia más dispar multiplica por 4 el tamaño del grupo de referencia respecto al focal (Goméz-Benito, Hidalgo, y Zumbo, 2013; Kristjansson, Aylesworth, McDowell, y Zumbo, 2005).

4.2.7. Cantidad de impacto

La distribución de las diferencias de medias también se considera un factor significativo que puede influir en la detección de DIF (Jodoin y Gierl, 2001; Kristjansson, Aylesworth, McDowell, y Zumbo, 2005). En la presente revisión sistemática, los estudios de simulación mayoritariamente analizan la RL en condiciones de inexistencia de impacto (n = 37 para ítems dicotómicos, y n = 9 para ítems politómicos), es decir, cuando las medias del nivel de habilidad de ambos grupos son iguales. Ésta debería ser la situación más favorable para estudiar una técnica, dado que se minimiza la posibilidad de confundir DIF con impacto en términos de control de la Tasa de Error Tipo I (DeMars, 2009; Li, Brooks y Johanson, 2012). En presencia de diferencias en las distribuciones, se estudian impactos con valores desde .1 (Li, Brooks, y Johanson, 2012) hasta 1.5 (Aguerri, Galibert, Atorresi, y Prieto, 2009) en ítems dicotómicos, y desde .1 (Allahyari, Jafari, y Bagheri, 2016) hasta 1 (Goméz-Benito, Hidalgo, y Zumbo, 2013) en ítems politómicos.

4.2.8. Tipo y cantidad de DIF

La RL es una técnica que permite detectar ambos tipos de DIF, aunque bajo ciertas condiciones parece ser más potente para la detección del DIF uniforme (Hidalgo

y López-Pina, 2004; Finch y French, 2008), donde la ventaja del grupo de referencia frente al focal se mantiene a lo largo del nivel de habilidad. De los 50 artículos analizados, 40 analizan DIF uniforme (32 en ítems dicotómicos y 8 en ítems politómicos), y 29 estudian el DIF no uniforme (22 en ítems dicotómicos y 7 en politómicos).

En cuanto a la cantidad de DIF en ítems dicotómicos, los estudios incluidos en la revisión sistemática manipulan un amplio abanico de cantidades de DIF, desde ítems (y por lo tanto tests) libres de DIF (por ejemplo, Pei y Li, 2010), que típicamente se utilizan para analizar únicamente las tasas de error Tipo I, hasta áreas entre las curvas características de los ítems de .8 o 1, lo que se considera una cantidad de DIF elevada. En el caso de ítems politómicos, se observa esta misma situación, aunque el número de estudios es sustancialmente inferior.

4.2.9. Porcentaje de ítems con DIF

La presencia de contaminación en la variable de equiparación implica que las puntuaciones del test no indiquen el nivel real de habilidad de los sujetos y, en consecuencia, afecte al funcionamiento de la técnica de detección de DIF (Clauser, Mazor, y Hambleton, 1993; French y Maller, 2007; Hidalgo y Gómez, 2003).

Los estudios incluidos en la revisión simulan desde una variable de equiparación totalmente libre de DIF (en ítems dicotómicos, p.e., Mazor, Hambleton, y Clauser, 1998; en ítems politómicos, p.e., Elosua y Wells, 2013) hasta un 50% de ítems con DIF (en ítems dicotómicos, i.e., Welkenhuysen-Gybels, 2004; en ítems politómicos, Scott, et

al., 2009), lo que indicaría esta última situación una variable de equiparación sustancialmente contaminada.

4.2.10. Procedimiento de purificación

La purificación consiste en eliminar de la variable de equiparación los ítems que en un primer análisis son identificados con DIF, con la intención de que la técnica sea más eficiente en la detección de ítems con DIF. Este procedimiento representa los primeros pasos hacia una estimación del nivel de habilidad de los sujetos, sin estar distorsionada por los ítems con DIF, y una manera de diferenciar el DIF real y el artificial (falso positivo). Diferentes procedimientos de purificación han sido propuestos. Si este proceso se detiene tras la eliminación del primer bloque de ítems, se habla de purificación bietápica, mientras que cuando el proceso se repite hasta no detectar ningún ítem con DIF, se denomina purificación iterativa (Hidalgo y Gómez-Benito, 2010).

Un total de 37 estudios de simulación con ítems dicotómicos han optado por no emplear ningún procedimiento de purificación, y unos pocos por la purificación bietápica (por ejemplo, Navas-Ara y Gómez-Benito, 2002) o iterativa (por ejemplo, French y Maller, 2007), mientras que en el caso de ítems politómicos son siete los estudios que tampoco emplean técnicas de purificación y dos los que utilizan la purificación bietápica (Hidalgo y Gómez-Benito, 2003, 2006).

4.3. Conclusiones

Los años de publicación oscilan entre el año 1990 y 2016, situándose el mayor número de artículos científicos concentrados en el año 2009 con un 12% del total.

Son los ítems dicotómicos los más estudiados, el 82% de los artículos que se incluyeron en la revisión sistemática. En este tipo de ítems, la técnica más comparada con la RL es el estadístico Mantel-Haenszel (Holland y Thayer, 1988; n = 24), mientras que en el caso de ítems politómicos, la técnica de comparación más utilizada es IRT likelihood-ratio test (Thissen, 2001; Thissen, Steinberg, y Wainer, 1988; n = 2).

El 80% de los estudios analizan simultáneamente la tasa de error y la potencia (32 en ítems dicotómicos y 8 en politómicos), el 16% analizan únicamente la tasa de error Tipo I (todos ellos en ítems dicotómicos) y el 4% solamente la potencia tanto en ítems dicotómicos como politómicos.

En el caso de ítems dicotómicos es el modelo logístico de 3 parámetros el más empleado para simular las respuestas y el modelo de respuesta graduada en el caso de ítems politómicos.

Los estudios incluidos, en la presente revisión, analizan diversos valores de parámetros de discriminación y dificultad de los ítems con DIF (baja, moderada y alta), para ítems dicotómicos, igualmente para ítems politómicos, además se observa la misma tendencia para los parámetros de pendiente y de localización. El valor del parámetro de pseudoazar más ampliamente estudiado es de c = .20.

Cuando hablamos de la longitud del test, en la presente revisión sistemática, para ítems dicotómicos la longitud del test más ampliamente estudiada es de 20 ítems, siendo de 10 ítems la longitud mínima analizada y de 100 la máxima. En el caso de ítems politómicos, el número de ítems del test oscila entre 2 ítems y 30 ítems, siendo de alrededor de 25 ítems la longitud del test más escogida. Son escasos los estudios que analizan el funcionamiento de la RL en tests muy cortos o largos.

Respecto al tamaño muestral del grupo de referencia y focal, en ítems dicotómicos, se constata que mayoritariamente se emplean grupos de 1,000 sujetos, con un rango que oscila entre 50 y 10,000. En el caso de los estudios que emplean ítems politómicos, el tamaño muestral del grupo de referencia y focal más estudiado es de 500 sujetos, con valores que van desde 100 sujetos hasta 3,200 sujetos en el grupo de referencia y hasta 2,000 en el grupo focal.

En la mayoría de estudios con ítems dicotómicos, se analiza la igualdad de tamaños muestrales entre el grupo focal y de referencia (ratio 1:1), pero diferencias tan dispares como que el grupo focal presente un tamaño muestral 10 veces inferior (ratio 10:1) al del grupo de referencia. En el caso de los estudios con ítems politómicos, casi todos contemplan la igualdad de tamaños muestrales en sus condiciones de simulación, únicamente la diferencia más dispar multiplica por 4 el tamaño del grupo de referencia respecto al focal.

En la presente revisión sistemática, los estudios de simulación mayoritariamente analizan la LR en condiciones de inexistencia de impacto. En presencia de diferencias

en las distribuciones, se estudian impactos con valores desde .1 hasta 1.5 en ítems dicotómicos, y desde .1 hasta 1 en ítems politómicos.

De los 50 artículos analizados, en 40 se estudia el DIF uniforme (32 en ítems dicotómicos y 8 en ítems politómicos), y en 29 se analiza el DIF no uniforme (22 en ítems dicotómicos y 7 en politómicos).

En cuanto a la cantidad de DIF simulada en los ítems, se manipula un amplio abanico de cantidades de DIF, desde ítems libres de DIF, hasta áreas entre las curvas características de los ítems de .8 o 1, lo que se considera una cantidad de DIF elevada.

Los estudios incluidos en la revisión simulan desde una variable de equiparación totalmente libre de DIF, hasta un 50% de ítems con DIF (variable de equiparación sustancialmente contaminada).

Respecto al empleo de procedimientos de purificación, son muy pocos los estudios que incluyen este procedimiento. Son solo unos pocos estudios, los que han empleado alguna técnica de purificación, siendo la elegida la bietápica o iterativa.

CAPÍTULO 5

REGRESIÓN LOGÍSTICA: UN ESTUDIO DE SIMULACIÓN

De la revisión sistemática se extraen diversas conclusiones que llevan a realizar un estudio más pormenorizado de la eficacia de RL, en comparación con IRTLRDIF, en tests cortos.

Tal y como Gelin y Zumbo (2007) indican, el estudio de la eficiencia y efectividad de la RL se ha realizado, principalmente, con tests largos. De la citada revisión, se extraen conclusiones similares, observándose una escasez de trabajos con tests muy cortos. Se puede ver, en la revisión sistemática previa, cómo, para ítems dicotómicos, la longitud del test más ampliamente estudiada es de 20 ítems, siendo de 10 ítems la longitud mínima analizada y de 100 la máxima. Y cómo, en el caso de ítems politómicos, el número de ítems del test oscila entre 2 ítems y 30 ítems, siendo de alrededor de 25 ítems la longitud del test más escogida.

A pesar de la escasez de trabajos que analizan el funcionamiento de la RL en test muy cortos, en las ciencias de la salud, las escalas tienden a tener sólo un pequeño

número (3-6) de ítems (Scott, et al., 2010; Teresi, 2006). Las escalas cortas y las versiones breves de tests, de este tipo, son cada vez más populares ya que son rápidas de aplicar y fáciles de puntuar, por lo que son adecuadas para su uso en procesos de cribado, evaluación clínica, investigación de encuestas y otros contextos de evaluación.

Existen diversos factores que pueden producir resultados engañosos, con respecto a la detección de DIF, entre ellos, la longitud del test, que es el principal problema, y la contaminación del criterio de equiparación. Cuando la variable de equiparación usada para detectar DIF es la puntuación total del test o una estimación de la habilidad o rasgo medido por el instrumento, y uno o más ítems están sesgados, esto puede conducir a una estimación de la habilidad inexacta y, por lo tanto, a una falsa identificación de DIF. Por ello, bajo esta condición, si el test es también corto, esto sólo exacerba el problema, ya que aumenta el riesgo de detectar ítems con pseudo-DIF (Scott, et al., 2009). Ésto se suma a la dificultad de interpretar el DIF y de explicar sus causas.

No obstante, la efectividad de los métodos de detección de DIF, en relación con la longitud del test, no ha sido suficientemente explorada. Scott, et al. (2009), realizaron un estudio de simulación utilizando la regresión logística ordinal, empleando diferentes longitudes de tests (2, 3, 4, 5, 10 y 20 ítems) y diferentes tamaños muestrales (500 o menos), concluyendo que el DIF se detectaba correctamente. Donoghue, Holland y Thayer (1993), utilizaron el procedimiento de Mantel-Haenszel (MH) y el procedimiento de estandarización, empleando diferentes longitudes de los tests (4, 9, 19 y 39 ítems), concluyendo que los resultados eran demasiado dependientes de algunos factores, por lo que no recomendaron el análisis DIF en tests con 4 y 9 ítems. Paek y

Wilson (2011), compararon los métodos MH y IRT-DIF bajo el modelo de Rasch, utilizando las mismas longitudes de tests que Donoghue, Holland, y Thayer (1993) y pequeños tamaños muestrales (100/100, 200/200 y 300/300), y encontraron que los métodos de IRT-DIF funcionaban bien, alcanzando un poder estadístico más alto que con el procedimiento de MH.

Aunque se ha demostrado que la RL es fácil de implementar y más flexible y eficiente que otros procedimientos para detectar el DIF, en el estudio de escalas cortas parece no ser adecuada, si se emplea como variable de equiparación la puntuación observada del test, otros estudios sugieren que, puede ser más apropiada la estimación de rasgos latentes utilizando IRT (Bolt, 2002; Wang y Yeh 2003).

Aunque que el análisis del DIF con escalas cortas puede ser problemático, numerosos estudios aplicados han tratado de analizar el DIF en tales escalas (Scott y otros, 2010). En consecuencia, es importante determinar la eficacia relativa de los métodos paramétricos y no paramétricos para la detección del DIF.

Con este fin, el presente trabajo compara la eficacia para detectar ítems con DIF, de IRTLRDIF, una técnica basada en la teoría de respuesta a ítems, y de DLR, una técnica que emplea la puntuación observada del test como criterio de equiparación, en tests cortos con ítems politómicos.

A pesar de que ambos tipos de DIF son importantes (Sireci y Rios, 2013), el presente estudio se centra en el estudio del DIF uniforme en ítems politómicos con un

patrón de DIF constante, ya que el DIF no uniforme se produce sustancialmente menos a menudo que el DIF uniforme (Camilli y Shepard, 1994).

5.1. Método

5.1.1. Generación de datos

Las respuestas de los ítems fueron generadas usando el modelo de respuesta graduada (Samejima, 1969). Las curvas características de límites entre categorías fueron definidas para representar la probabilidad acumulada (*P*jk(θ)*) de una respuesta por encima de la categoría k. Éstas se dan como:

$$P_{jk}^{*}(\theta) = \frac{\exp(a_j(\theta - b_{jk}))}{1 + \exp(a_j(\theta - b_{jk}))}$$

donde a_j es el parámetro de discriminación del ítem *j*, b_{jk}es el parámetro de dificultad del ítem *j* en el límite de la categoría *k*, y θ es el parámetro de habilidad.

5.1.2. Condiciones experimentales

Se manipularon cuatro variables:

- El tamaño muestral para los grupos focales y de referencia: Se establecieron diferentes tamaños muestrales (250/250, 500/500 y 1000/1000) reflejando las situaciones que son más probables en la

práctica y seleccionado tamaños muestrales pequeños y grandes para los grupos de comparación. No obstante, también es frecuente encontrar tamaños muestrales desequilibrados entre grupos, con tamaños muestrales más pequeños para los grupos focales. En estos casos, se puede alcanzar el mismo tamaño muestral extrayendo una submuestra aleatoria del grupo mayoritario.

- La cantidad de DIF: Se manipularon dos diferencias entre parámetros de dificultad (.4 y .8), lo que indicó que las magnitudes de DIF fueron simuladas para ser moderadas y grandes para cada ítem. Los parámetros de la TRI, se muestran en la Tabla 5.1.

- Longitud del test: Los tests simulados constaban de 4, 5, 8 o 10 ítems. Una escala de 10 ítems es representativa de numerosas escalas cortas que suelen encontrarse en las ciencias del comportamiento, por lo que se utilizaron longitudes del test más cortas, para considerar otras escalas usadas frecuentemente en estudios de cribado. También se consideró el porcentaje de ítems con DIF en un test (0%, 10%, 12,5%, 20% y 25%). La condición del 10% refleja una situación general en el uso de tests, aunque en las adaptaciones, el porcentaje de los ítems con DIF es frecuentemente mayor al 20%. Esta condición se manipuló en combinación con la longitud del test, de tal manera que los tests simulados tenían sólo un ítem con DIF.

- Número de categorías de respuesta por ítem: Los ítems fueron simulados para representar ítems con cuatro o tres categorías de respuesta, como se encuentra en las ciencias del comportamiento y de la salud.

Se manipularon un total de 48, más 36, condiciones. Bajo cada condición se realizaron 100 repeticiones.

Tabla 5.1

Parámetros del ítem para el grupo de referencia

	k=4				k=3		
Ítem	a_R	b_{1R}	b_{2R}	b_{3R}	a_R	b_{1R}	b_{2R}
1	.99	-1.95	-.19	2.57	.99	-1.95	-.19
2	.37	-.64	.77	1.66	.37	-.64	.77
3	.90	-.91	.21	.98	.90	-.91	.21
4	.88	-2.25	-1.80	1.66	.88	-2.25	-1.80
5	.63	-2.11	-.54	.74	.63	-2.11	-.54
6	.99	-1.95	-.19	2.57	.99	-1.95	-.19
7	.37	-.64	.77	1.66	.37	-.64	.77
8	.90	-.91	.21	.98	.90	-.91	.21
9	.88	-2.25	-1.80	1.66	.88	-2.25	-1.80
10	.63	-2.11	-.54	.74	.63	-2.11	-.54

Nota: Ítems 1-4 para longitud de test = 4; Ítems 1-5 para longitud de test = 5; Ítems 1-8 para longitud de test = 8; Ítems 1-10 para longitud de test = 10. En todos los casos, el ítem manipulado para mostrar DIF fue el ítem 1; *a*: parámetro de discriminación, *b*: parámetro de umbral; R: grupo de referencia; *k*: número de categorías de respuesta del ítem.

"Parcialmente tomada de Hidalgo, López-Martínez, Gómez-Benito y Guilera (2016)"

5.2. Análisis de los datos

La detección de DIF por DLR se estimó utilizando un programa de software creado por M.D. Hidalgo y J. Gómez-Benito. La puntuación total se utilizó como criterio de equiparación. Se indicó que los ítems tenían DIF significativo cuando la comparación de G^2 del Modelo 2 con respecto al Modelo 1 era significativa a $p \leq .05$.

La detección de DIF por IRTLR se estimó utilizando el software IRTLRDIF (Thissen, 2001). Se aplicó una estrategia en una etapa y todos los ítems excepto el ítem bajo estudio se usaron como anclaje. Se identificó un ítem con DIF cuando el estadístico de comparación G^2 fue significativo al .05.

Por último, la tasa de error Tipo I para cada ítem sin DIF fue evaluada mediante la proporción de falsos positivos y la potencia de cada ítem con DIF fue evaluada mediante la proporción de verdaderos positivos. De acuerdo al criterio liberal de robustez de Bradley (1978), una prueba estadística puede considerarse robusta si su tasa de error de Tipo I, $\hat{\alpha}$, está dentro del intervalo $.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$. Por lo tanto, para el nivel nominal $\alpha - .05$, la tasa de error Tipo I debe estar dentro del intervalo $.025 \leq \hat{\alpha} \leq .075$.

5.2.1. Error Tipo I

La Tabla 5.2 muestra las tasas de error de Tipo I de cada procedimiento en cada cada condición manipulada. En cuanto a la condición de no DIF, las tasas de error Tipo I fueron, como se esperaba, cercanas al nivel nominal en todos los tamaños muestrales,

longitudes de tests y en ambas técnicas de detección de DIF. Específicamente, las tasas de error Tipo I oscilaron entre .023 y .068 y se encontraron entre .075 y .025 en todas las condiciones, excepto cuando se utilizó IRTLRDIF con tamaños muestrales pequeños y longitudes de tests cortos. En general, el número de categorías de respuesta de los ítems no tuvo ningún efecto: cuando k = 4 las tasas de error Tipo I para DLR oscilaron entre .035 y .068, mientras que cuando k = 3 oscilaron entre .034 y .068. Un patrón similar se observó cuando se utilizó IRTLRDIF: cuando k = 4 las tasas de error Tipo I oscilaron entre .023 y .049, mientras que cuando k = 3 oscilaron entre .025 y .065.

Tabla 5.2

Tasa de error Tipo I al 5% para todas las condiciones

			k=4		**k=3**	
Tamaño de la muestra Referencia / Focal	n	**Cantidad de DIF**	**DLR**	**IRTLRDIF**	**DLR**	**IRTLRDIF**
250/250	4	.0	.065	.023	.068	.040
		.4	**.120**	.033	**.307**	.060
		.8	**.190**	.057	**.430**	.067
	5	.0	.056	.046	.046	.040
		.4	**.078**	.030	.053	.038
		.8	**.103**	.058	**.083**	.048
	8	.0	.061	.054	.056	.059
		.4	.050	.053	.056	.046
		.8	.070	.069	.069	.064
	10	.0	.047	.045	.049	.056
		.4	.059	.063	.056	.050
		.8	.066	.060	.049	.062

500/500	4	.0	.068	.033	.058	.030
		.4	**.120**	.037	**.077**	.030
		.8	**.327**	**.110**	**.197**	.053
	5	.0	.046	.030	.034	.030
		.4	**.113**	.060	.070	.033
		.8	**.240**	**.078**	**.135**	.055
	8	.0	.054	.058	.041	.035
		.4	.066	.069	.053	.039
		.8	**.097**	.063	**.080**	.059
	10	.0	.050	.042	.048	.056
		.4	.041	.047	.059	.050
		.8	.069	.054	.058	.041
1000/1000	4	.0	.035	.025	.045	.025
		.4	**.220**	.070	.150	.063
		.8	**.557**	.210	**.380**	**.077**
	5	.0	.060	.040	.040	.040
		.4	**.135**	.045	**.085**	.025
		.8	**.340**	.155	.205	.053
	8	.0	.045	.045	.050	.065
		.4	**.076**	.047	.060	.044
		.8	**.166**	.086	**.110**	.051
	10	.0	.059	.048	.047	.054
		.4	**.077**	.052	.063	.047
		.8	**.118**	**.104**	**.076**	.039

Nota: En negrita, las tasas de error Tipo I error que excedieron el criterio liberal; *n*: tamaño de test; *k*: número de categorías de respuesta del ítem; DLR: regresión logística discriminante.

"Parcialmente tomada de Hidalgo, López-Martínez, Gómez-Benito y Guilera (2016)"

En las condiciones con DIF, la tasa de error de Tipo I se vio afectada no sólo por la longitud del test, sino también por la cantidad de DIF, el tamaño muestral y el número de categorías de respuesta de los ítems. Las tasas de error de Tipo I difieren

según la técnica de detección de DIF utilizada. Cuando se utilizó DLR para detectar DIF, las tasas de error de Tipo I estaban por encima de .075 en todas las condiciones, cuando se manipulaba el DIF en tests muy cortos. Las tasas de error de tipo I oscilaron entre .19 a .56 cuando la cantidad de DIF era .8 y de .08 a .31 cuando era .4. Con respecto al número de categorías por ítem, la tasa de error Tipo I varió de .12 a .56 para k = 4 y de .08 a .43 para k = 3. En general, las tasas de error Tipo I fueron mayores cuando el tamaño muestral era mayor.

Cuando se utilizó IRTLRDIF para detectar DIF en tests muy cortos, las tasas de error de Tipo I fueron mayores que el criterio liberal, cuando la cantidad de DIF fue de .8, NR = NF = 1000 y k = 4 y también cuando la cantidad de DIF fue .8 y la muestra total fue de 1000. Las tasas de error Tipo I fueron mayores cuando el tamaño muestral fue mayor.

Las tasas de error de Tipo I disminuyeron a medida que aumentaba el tamaño del test. Este efecto fue más notable para DLR que para IRTLRDIF. Cuando fue manipulado el DIF en tests con n = 8, las tasas de error de Tipo I variaron de .07 a .17 cuando la cantidad de DIF fue de .8, y entre .05 a .08 cuando fue de .4. En los tests con n = 8 y k = 4, las tasas de error Tipo I oscilaron entre .05 y .17, mientras que en los tests con k = 3 oscilaron entre .05 y .11. En tests muy cortos, las tasas de error Tipo I fueron mayores cuando el tamaño muestral era mayor. Cuando se utilizó IRTLRDIF en tests con ocho ítems, las tasas de error de Tipo I fueron mayores que el criterio liberal, en la condición de una mayor cantidad de DIF (.8), N = 2000 y k = 4. Más específicamente, las tasas de error de Tipo I oscilaron entre .05 y .09 cuando la cantidad de DIF fue .8, y entre .04 y .07 cuando fue .4. Estos resultados fueron similares independientemente del

número de categorías por ítem, ya que en ambas condiciones la tasa promedio de error Tipo I fue menor que el criterio liberal. En general, la tasa de error Tipo I fue inferior a .07 cuando el tamaño total de la muestra fue de 500 o 1000, y varió de .04 a .09 cuando el tamaño muestral total fue de 2000. Este patrón fue similar cuando se consideró un test más largo. En tests muy cortos, las tasas de error Tipo I fueron más bajas para IRTLRDIF que para DLR, mientras que los dos procedimientos se comportaron de forma similar con tests más largos.

5.2.2. Potencia

La Tabla 5.3 muestra que la potencia se vio afectada por la cantidad de DIF, el tamaño muestral y la técnica de DIF utilizada, pero no por la longitud del test. Cuando se consideraron tests muy cortos (n = 4) y DLR fue usado, la potencia estaba por encima de .80, cuando la cantidad de DIF fue de .8, independientemente del tamaño muestral y el número de categorías de respuesta del ítem. Sin embargo, cuando la cantidad de DIF era .4, la potencia oscilaba entre .59 y 1, y estaban por debajo de .80, cuando el tamaño total muestral era de 500 o 1000, independientemente del número de categorías de respuesta. Cuando se utilizó IRTLRDIF y la cantidad de DIF fue de 0,8, las tasas de identificaciones correctas fueron inferiores a .80 para tests muy cortos y tamaños muestrales pequeños. Cuando la cantidad de DIF era de .4, las tasas de identificaciones correctas estuvieron por encima de .80 con tamaños muestrales mayores.

Tabla 5.3

Tasa de potencia al 5% para todas las condiciones manipuladas

			k=4		k=3	
Sample Size Reference/Focal	**n**	**Amount of DIF**	**DLR**	**IRTLRDIF**	**DLR**	**IRTLRDIF**
250/250	**4**	.4	.59*	.23	.60*	.03
		.8	.99*	.64	1.00*	.00
	5	.4	.63*	.38	.40	.34
		.8	1.00*	.79	.89*	.77
	8	.4	.65	.44	.39	.20
		.8	.99	.87	.97	.74
	10	.4	.55	.32	.36	.32
		.8	1.00	.86	.96	.90
500/500	**4**	.4	.76*	.60	.60*	.36
		.8	1.00*	.83*	1.00*	.83
	5	.4	.84*	.50	.64	.42
		.8	1.00*	.87*	1.00*	.95
	8	.4	.87	.62	.68	.54
		.8	1.00*	.97	1.00*	.97
	10	.4	.87	.64	.75	.61
		.8	1.00	.98	1.00	.99
1000/1000	**4**	.4	1.00*	.93	.93*	.80
		.8	1.00*	.97*	1.00*	.93*
	5	.4	.99*	.93	.94*	.86
		.8	1.00*	.93*	1.00*	.99
	8	.4	.98*	.96	.96	.90
		.8	1.00*	1.00*	1.00*	.99
	10	.4	.99*	.96	.93	.87
		.8	1.00*	1.00*	1.00*	1.00

Nota: Un asterisco indica que la potencia no tenía sentido porque su correspondiente error Tipo I estaba inflado*; n:* tamaño de test*; k:* número de categorías de respuesta del ítem; DLR: regresión logística discriminante.

"Parcialmente tomada de Hidalgo, López-Martínez, Gómez-Benito y Guilera (2016)"

Cuando se consideraron tests más largos, tanto DLR como IRTLRDIF alcanzaron tasas de potencia inferiores a .80, cuando el tamaño total de la muestra era inferior al 2000, independientemente de la cantidad de DIF y del número de categorías de respuesta del ítem. Sin embargo, en las condiciones anteriores, la potencia para ambos procedimientos era mayor cuando el número de categorías de respuesta de ítem era k = 4 en lugar de k = 3. Por lo tanto, cuando se manipuló el DIF en tests con ocho ítems y la técnica utilizada fue DLR, las tasas de potencia variaron de .36 a .65 cuando la cantidad de DIF fue .4 y de .68 a .87 cuando fue de .8. En los tests con ocho ítems y con cuatro categorías por ítem, las tasas de identificaciones correctas fueron .65 (tamaños muestrales 250/250) y .87 (tamaños muestrales 500/500), mientras que en los tests con tres categorías por ítem fueron .39 (tamaños muestrales 250/250) y .75 (tamaño muestrales 500/500). Cuando se utilizó IRTLRDIF, variaron de .20 a .44 cuando la cantidad de DIF fue de .4 y de .54 a .64 cuando fue de .8. En tests con ocho ítems y cuatro categorías por ítem, la potencia fue de .44 (tamaños muestrales 250/250) y .62 (tamaños muestrales 500/500), mientras que en los tests con tres categorías por ítem fue de .20 (tamaños muestrales 250 / 250) y .54 (tamaños muestrales 500/500).

DLR mostró una mayor potencia que IRTLRDIF en tests cortos, pero este resultado no puede tenerse en cuenta debido a la alta tasa de error Tipo. Como era de esperar, en tests con 8 o 10 ítems, DLR mostró una mayor potencia que IRTLRDIF con tamaños muestrales más pequeños, independientemente de la magnitud del DIF manipulado.

5.3. Conclusiones

A cerca de los resultados en este estudio, en tests cortos (4 o 5 ítems), debe tenerse en cuenta el grado de contaminación del criterio de equiparación, a la hora de interpretar los resultados; ya que, en este caso, fue notablemente alto, con un 25% o 20% de los ítems con DIF en el test, debido a que sólo se simuló un ítem con DIF en cada test.

Respecto a la potencia, los principales resultados muestran que ésta se ve afectada por la cantidad de DIF y el tamaño muestral, pero no por la longitud del test. Se observa como DLR muestra mayor potencia que IRTLRDIF, en test cortos de 4 ó 5 ítems, teniendo en cuenta la alta tasa de error Tipo I implicada y en tests con 8 ó 10 ítems y tamaños muestrales pequeños. Esto se puede deber, a la gran proporción de ítems con DIF que muestran los tests; al igual que en otros estudios, este alto nivel de contaminación en los ítems de anclaje, se asocia con una disminución en la potencia de IRTLRDIF para identificar correctamente ítems con DIF (Wang y Yeh, 2003).

Respecto a las tasas de error Tipo I, los principales resultados son consistentes con la investigación anterior. Las tasas de error de Tipo I fueron menores para IRTLRDIF, mostrando un mejor control de la tasa de error Tipo I que DLR. Para ambos procedimientos, las tasas de error de Tipo I fueron superiores al nivel nominal, en tests cortos, con cuatro categorías de respuesta en cada ítem y cuando el tamaño muestral y la cantidad de DIF fueron mayores.

No obstante, se demostró que IRTLRDIF no era adecuado si el test de anclaje incluía una gran cantidad de DIF. Esto es consistente con los hallazgos de Finch y

French (2008), González-Betanzos y Abad (2012), y Wang y Yeh (2003), todos los cuales encontraron que la Tasa de Error Tipo I está muy influenciada por el nivel de contaminación en los ítems de anclaje.

La tasa de error Tipo I fue especialmente elevada cuando la cantidad de DIF fue grande. Cuando se utilizó DLR para detectar el DIF, las tasas de error Tipo I se vieron afectadas por el tamaño muestral, la longitud del test y la cantidad de DIF. Además, se dio interacción entre:

- La longitud del test y la cantidad de DIF (mayor tasa de error de Tipo I con una mayor cantidad de DIF en el test y tests más cortos).

- El tamaño muestral y la cantidad de DIF (alta tasa de error Tipo I con tamaños muestrales más grandes y una mayor cantidad de DIF).

DISCUSIÓN

Nos encontramos frente a una sociedad donde las personas difieren en multitud de variables como la cultura, el idioma o la etnia. Es por ello que los profesionales deben garantizar que sus evaluaciones sean justas e independientes de las diferencias interpersonales.

En este contexto, la correcta detección de los ítems que funcionan diferencialmente para ciertos grupos de personas adquiere un papel fundamental. De entre la diversidad de técnicas estadísticas para detectar ítems con DIF, el presente estudio se ha centrado en la RL. Sireci y Rios (2013) destacan múltiples ventajas de este procedimiento frente a otras técnicas de detección del DIF, por lo que recomiendan su uso.

El primer estudio, que ha sido realizado, analiza el comportamiento de la productividad científica en el estudio y aplicación de la RL, como técnica de análisis y detección del DIF.

Conocer el estado de la investigación en un campo determinado, el curso de su desarrollo y su tendencia de crecimiento, ofrece una información importante a la comunidad científica; que a su vez, puede verse beneficiada de este conocimiento, para dirigir u orientar su tarea investigadora hacia una mayor eficacia y aprovechamiento de los recursos.

Los análisis llevados a cabo, se han centrado en conocer la evolución de la producción científica de los trabajos que han empleado datos empíricos o simulados, en el estudio de la RL, como técnica de análisis y detección del DIF.

Se ha obtenido 352 artículos, publicados en revistas científicas, en inglés, entre 1990 y 2016. De los artículos obtenidos, en 319 se estudia el DIF, siendo en 223 de éstos, en los que emplea la RL como técnica de detección. Entre los 223 trabajos, parte de ellos son trabajos teóricos, por lo que, finalmente, son incluidos en esta revisión 210.

Parece claro que la RL es una técnica popular, que se aplica frecuentemente en la práctica para el análisis del DIF, ya que, algo más del 75% de los trabajos encontrados son estudios de este tipo (estudios de tipo empírico). Igualmente, es un procedimiento que suscita interés entre los investigadores psicómetras, ya que un 25% de los trabajos evalúan la eficacia de la RL mediante estudios de simulación. En el caso del estadístico MH, en el trabajo de Guilera, Gómez-Benitez e Hidalgo (2009), se puede observar que la mayoría de los artículos están centrados en el estudio de la técnica mediante datos simulados (57%), siendo el resto estudios de tipo empírico (32%), salvo un 7% de revisiones teóricas y un 5% que desarrollan un nuevo software para implementar el procedimiento MH. Esto muestra como la RL es frecuentemente

empleada en la práctica para la detección del DIF y cómo aún le queda mucho camino por recorrer, respecto a su estudio, para estar a la altura de MH, en número de trabajos con datos simulados.

El número de trabajos aplicados (que emplean datos empíricos) ha ido aumentando progresivamente a lo largo de los últimos años. Parece resultar evidente, no solo el interés actual, sino el crecimiento progresivo del uso aplicado de la RL para la detección del DIF. Su mayor crecimiento se observa en la primera década del siglo XXI, mostrando un desarrollo más tardío que el estadístico MH que alcanzó su pico en 1995, según el trabajo de Guilera, Gómez-Benito e Hidalgo (2009).

La preferencia a la hora de liderar trabajos en solitario o de colaborar con más autores es bastante distinta, según el tipo de estudio que se haya llevado a cabo. Si se trata de estudios con datos empíricos, lo más frecuente es colaborar entre dos, tres y cuatro autores, con un 62.1%; si se tratara de estudios con datos simulados, lo más común es encontrar artículos firmados por dos investigadores (57.7%). Podría dar explicación a este suceso el hecho de necesitar para los trabajos, en los que se manejan datos empíricos, la aplicación de una prueba de medición a una muestra o población, siendo más beneficiosas, para el resultado, las colaboraciones múltiples. En cambio, esta cifra se ve reducida en los trabajos con datos simulados, ya que no es necesaria la recogida de datos empíricos y el trabajo está centrado en el manejo y análisis de datos mediante un soporte informático.

El número de colaboraciones internacionales entre autores se incrementa a medida que aumenta el número de autores y autoras firmantes, y lo es,

sistemáticamente, en el caso de artículos entre dos y cinco firmantes. Aunque lo más frecuente son las colaboraciones nacionales.

La revista más productiva es *Educational and Psychological Measurement*, seguida por *International Journal of Testing* y *Journal of Educational Measurement*. Siendo, *Journal of Educational Measurement* la revista más productiva respecto al estadístico MH, según el trabajo de Guilera, Gómez-Benito e Hidalgo (2009). Podría afirmarse, que la citada revista, muestra cierto interés por los estudios de DIF, que emplean RL y MH.

El país con mayor producción de investigación es Estados Unidos, tanto en la RL como en MH (Guilera, Gómez-Benito e Hidalgo, 2009), ocupando España el tercer lugar de la lista de países. En el caso de MH (Guilera, Gómez-Benito e Hidalgo, 2009) España queda en segundo lugar. España puede considerarse un país productivo, a nivel mundial, en el estudio de DIF mediante RL y MH, por detrás de Estados Unidos, ocupando uno de los tres primeros puestos en el ranking mundial.

Las instituciones que contribuyen a la investigación son en su mayoría universidades, tanto en la RL como en MH (Guilera, Gómez-Benito e Hidalgo, 2009), liderando la lista la Universidad de Washington, seguida de la Universidad de British Columbia, la Universidad de Massachusetts y la Universidad de Barcelona. Se observa que el estudio del DIF, mediante RL y MH, no es un tema de gran interés en el ámbito privado, pero sí en el ámbito académico.

Después de aplicar la ley de Bradford de dispersión con respecto a la variable "revista", se definieron tres zonas concéntricas. El núcleo contuvo 42 artículos de los 210 totales, lo que corresponde al 20% de los documentos, que fueron publicados en dos revistas, *Quality of Life Research* (con 25 artículos publicados) y *Educational and Psychological Measurement* (con 17 artículos publicados).

Dados los resultados encontrados y con el fin de encontrar posibles brechas en el estudio de la RL se realiza un segundo estudio, una revisión sistemática centrada en los trabajos que emplean datos simulados, analizando en profundidad las condiciones de simulación en las que se ha estudiado el funcionamiento de la LR hasta la actualidad.

Para esta revisión se ha incluido un total de 50 artículos, publicados en revistas científicas, que analizan el funcionamiento de la RL, mediante la simulación de datos, cuando ésta se utiliza en la detección del DIF.

Se observa cómo la RL se compara principalmente con MH (en ítems dicotómicos) y con IRT likelihood-ratio (en ítems politómicos); y cómo, con mayor frecuencia, se estudian tests de 20 ítems (de tipo dicotómicos) y de 25 ítems (de tipo politómicos).

Algunos trabajos recientes, analizan el DIF en datos reales en el campo de la salud, utilizando tests de longitud reducida (Scott et al., 2009; Scott et al., 2010). Así por ejemplo, Asmundson, LeBouthillier, Parkerson, y Horswill (2015) estudiaron el DIF basado en el género en la escala Dimensiones de las Reacciones de Cólera-5 (DAR-5) [Dimensions of Anger Reactions–5 (dar-5) Scale] (Forbes, Hawthorne, Elliott,

McHugh, Biddle, Creamer y Novaco, 2004) en una muestra comunitaria de adultos que habían estado expuestos a un trauma. Aunque en menor medida, en el ámbito educativo también se encuentran algunos instrumentos de longitud corta, como por ejemplo las escalas del PISA, Cuestionario de Estudiantes, que tienen entre 4 y 8 ítems (OECD, 2006), e incluso algunos estudios han analizado el DIF en este tipo de pruebas (Benítez, Padilla, Hidalgo y Sireci, 2016; Balluerka, Plewis, Gorostiaga y Padilla, 2014). Ejemplos de estudios de DIF en tests de longitud reducida en el contexto educativo los encontramos en French, Finch y Valdivia (2016) que evalúan el DIF en el Inventario Comprensivo de Habilidades Básicas II [Brigance: Comprehensive Inventory of Basic Skills (CIBS) II] formado por tres subtests con 6, 7 y 15 ítems respectivamente.

Los resultados muestran que existe una tendencia a utilizar tests cada vez más cortos, con el fin de buscar un equilibrio razonable entre el tiempo de administración y la validez, y la precisión de las puntuaciones obtenidas con estos instrumentos.

Igualmente, tal y como Gelin y Zumbo (2007) indican, el estudio de la eficiencia y efectividad de la RL se ha realizado, principalmente, con tests largos. De la citada revisión, se extraen conclusiones similares, observándose una escasez de trabajos con tests muy cortos. Respecto a la longitud del test más ampliamente estudiada, para ítems dicotómicos es de 20 ítems (siendo de 10 ítems la longitud mínima analizada) y para ítems politómicos, la longitud el test más escogida es de 25 ítems (siendo de 2 ítems la longitud mínima analizada).

Así, el tercer estudio realizado en esta tesis doctoral pretende comprar la eficacia de IRTLRDIF y de la regresión logística discriminante (DLR) en la detección de DIF en

pruebas muy cortas (≤10 ítems), analizando las tasas de error Tipo I y la potencia de ambos procedimientos bajo determinadas condiciones.

Respecto a las tasas de error Tipo I, los principales resultados son consistentes con la investigación anterior. Las tasas de error de Tipo I fueron menores para IRTLRDIF, mostrando un mejor control de la tasa de error Tipo I que DLR. Respecto a la potencia, se observa que ésta se ve afectada por la cantidad de DIF y el tamaño muestral, pero no por la longitud del test.

No obstante, se demostró que IRTLRDIF no era adecuado si el test de anclaje incluía una gran cantidad de DIF. Esto es consistente con los hallazgos de Finch y French (2008), González-Betanzos y Abad (2012), y Wang y Yeh (2003), todos los cuales encontraron que la Tasa de Error Tipo I está muy influenciada por el nivel de contaminación en los ítems de anclaje.

Por último, como dice la teoría, tanto la fiabilidad de la variable de equiparación como la variabilidad en las puntuaciones totales de los tests será menor con tests cortos, de modo que la tasa de error Tipo I pueda incrementarse. Sin embargo, cuando se utiliza el procedimiento MH, la longitud del test tiene un efecto mínimo sobre la tasa de error Tipo I y la potencia, al detectar DIF en tests entre 20 y 40 ítems, con resultados peores con menos de 20 ítems (Guilera, Gómez-Benito, Hidalgo y Sánchez-Meca, 2013). Scott, et al. (2009), utilizando la RL ordinal, también encontraron que la longitud del test no era relevante para los tests de entre 5 y 20 ítems. Por el contrario, la cantidad de contaminación del criterio de equiparación tiene un efecto importante en la detección de DIF (Guilera, et al., 2013).

Se esperaba cierto desacuerdo entre los dos métodos de detección. En cuanto a los enfoques paramétricos para la detección de DIF, como IRTLRDIF, el problema aquí, como señala Bolt (2002), se refiere a la especificación del modelo y la necesidad de emplear tamaños muestrales mayores para evitar tasas de error Tipo I infladas debido al desajuste del modelo. En los métodos IRT-DIF, la potencia para detectar DIF incrementa con el aumento del tamaño de la muestra, mientras que los métodos como LR (enfoque no paramétrico) son suficientemente potentes con tamaños de muestra relativamente pequeños (Scott et al., 2009). En general, la elección entre IRTLRDIF y DLR debe guiarse no sólo por el tamaño de la muestra sino también por la disponibilidad de software y experiencia estadística. Es importante notar que el DLR está disponible en la mayoría de los paquetes de software estadístico y requiere tamaños de muestra relativamente pequeños, sin embargo los métodos de razón de verosimilitud basados en IRT requieren tamaños de muestra relativamente mayores y supuestos de modelo más restrictivos (Sireci y Rios, 2013).

En todo caso, e independientemente del método de detección DIF utilizado, el principal problema con escalas cortas es la dificultad de identificar qué ítem está causando DIF, ya que un ítem con DIF puede contaminar los otros ítems a través de su contribución a variable de equiparación (Scott, et al., 2009). Aunque pueden aplicarse procedimientos de purificación, pueden ser menos adecuados para las escalas con sólo un pequeño número de ítems, ya que la eliminación de los ítems puede afectar la precisión de la variable correspondiente (Scott, et al., 2010). Para las escalas de este tipo recomendamos el uso de métodos mixtos (Benítez, Padilla, Hidalgo y Sireci, 2016) acompañados de medidas de tamaño de efecto para tomar decisiones sobre la

eliminación cambio de los ítems DIF en una prueba (Gómez-Benito, Hidalgo, y Zumbo, 2013). También se puede considerar la RL multinivel (Balluerka, Gorostiaga, Gómez-Benito e Hidalgo, 2010).

Aunque los hallazgos del presente estudio proporcionan algunos consejos prácticos sobre la detección de DIF en test y escalas cortas, tiene varias limitaciones relacionadas principalmente con las condiciones de simulación consideradas. Por lo tanto, se necesitan más investigaciones para determinar el efecto de los tamaños de muestra desequilibrados para los grupos de referencia y focal, la detección de DIF no uniforme o la extensión de resultados a otros patrones de DIF. En este último caso, es importante señalar que el DIF en los ítems politómicos es mucho más complejo que en los ítems dicotómicos, y se pueden encontrar varios patrones de DIF dependiendo del modelo de respuesta (Penfield, 2007; Penfield, Alvarez y Lee, 2009). Sería interesante extender este estudio manipulando distintos patrones de DIF (Penfield, 2007).

Referencias Bibliográficas

Los documentos marcados con un asterisco han sido incluidos en la revisión sistemática.

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

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley.

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

*Aguerri, M. E., Galibert, M. S., Attorresi, H. F. y Prieto Marañón, P. (2009). Erroneous detection of nonuniform DIF using the Breslow-Day test in a short test. *Quality y Quantity, 43*(1), 35-44.

*Allahyari, E., Jafari, P. y Bagheri, Z. (2016). A simulation study to assess the effect of the number of response categories on the power of ordinal logistic regression for differential item functioning analysis in rating scales. *Computational and Mathematical Methods in Medicine*, volumen 2016.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrés, A. (2009). *Measuring academic research: How to undertake a bibliometric study*. Oxford, UK: Chandos Publishing.

Angoff, W. H. (1993). Perspectives on Differ-ential Item Functioning Methodology. En P. W. Holland y H. Wainer (Eds), *Differential Iten Functioning* (pp. 3-23). New Jersey: Lawrence Erlbaum Associates, Inc.

*Atar, B. y Kamata, A. (2011). Comparison of IRT Likelihood Ratio Test and Logistic Regression DIF Detection Procedures. *Hacettepe University Journal of Education, 41*, 36-47.

Balluerka, N., Gorostiaga, A., Gómez-Benito, J. e Hidalgo, M. D. (2010). Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema, 22*(4), 1018-1025.

Balluerka, N., Plewis, I., Gorostiaga, A. y Padilla, J. L. (2014). Examining sources of DIF in psychological and educational assessment using multilevel logistic regression. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *10*(2), 71.

*Bastug, Ö. Y. Ö. (2016). A Comparison of Four Differential Item Functioning Procedures in the Presence of Multidimensionality. *Educational Research and Reviews*, *11*(13), 1251-1261.

Benítez, I., Padilla, J. L., Hidalgo, M. D. y Sireci, S. G. (2016). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education*, *29*(1), 1-16.

*Berger, M. y Tutz, G. (2016). Detection of Uniform and Nonuniform Differential Item Functioning by Item-Focused Trees. *Journal of Educational and Behavioral Statistics*, *41*(6), 559-592.

Binet, A. y Simon, T. (1908). Le développement de l''intelligence chez les enfants. *L'Année Psychologique*, *14*, 1–94.

Binet, A. y Simon, T. (1911). Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants des écoles. *L'Année Psychologique*, *17*, 145–201.

Bolt, D. M. (2002). A Monte-Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Psychological Measurement*, *15*, 113-141.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, *23*(3), 85-88.

Bradford, S. C. (1948). Documentation. *Lockwood Sons Ltd. London.*

Bradley, J. V. (1978) Robustness?. *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Buela-Casal, G. (2003). Evaluación de la calidad de los artículos y de las revistas científicas: Propuesta del factor de impacto ponderado y de un índice de calidad. *Psicothema*, *15*(1), 23-35.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*(2), 129-147.

Camilli, G. y Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. United States: Sage Publications.

Chang, H. H., Mazzeo, J. y Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of educational Measurement, 33,* 333-353.

Clauser, B. E. y Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: issues and practice, 17*(1), 31-44.

Clauser, B. E., Mazor, K. M. y 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.

Clauser, B. E., Nungester, R. J., Mazor, K. y Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*(2), 202-214.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*(2), 115-124.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

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

Cole, N. (1993). History and Development of DIF. En P. W. Holland y H. Wainer (Eds), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates, Inc.

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach*, 5th Edition. Thousand Oaks, CA: Sage Publications.

De Ayala, R. J., Kim, S. H., Stapleton, L. M. y Dayton, C. (1999). *A Reconceptualization of Differential Item Functioning*. Documento presentado en la reunión anual de American Educational Research Association, Montreal, Canadá.

*DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*(2), 149-170.

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education, 24*, 189-209.

Donoghue, J. R., Holland, P. W. y Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P.W. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J. y Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale NJ: Lawrence Erlbaum.

Dorans, N. J. y Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach. *ETS Research Report Series, 1983(1).*

Dorans, N. J. y Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of educational measurement, 23*(4), 355-368.

Dorans, N. J., Schmitt, A. P. y Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*(4), 309-319.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E. y Tyler, R. W. (1951). *Intelligence and cultural differences: A study of cultural learning and problem-solving* (pp. 3-9). Chicago: University of Chicago Press.

Egghe, L. (1986). The dual of Bradford's law. *Journal of the American Society for Information Science*, *37*(4), 246.

Egghe, L. (1990). A note on different Bradford multipliers. *Journal of the Association for Information Science and Technology*, *41*(3), 204-209.

Egghe, L. y Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. msterdam: Elsevier Science Publishers.

*Elosua, P. y Wells, C. (2013). Detecting DIF in Polytomous Items Using MACS, IRT and Ordinal Logistic Regression. *Psicologica: International Journal of Methodology and Experimental Psychology*, *34*(2), 327-342.

Ferreres, D., Fidalgo, A. y Muñiz, J. (2000). Detección del funcionamiento diferencial de los ítems no uniforme: comparación de los métodos Mantel-Haenszel y regresión logística. *Psicothema*, *22*(Suplemento), 220-225.

Fidalgo, A. M. (1994). MHDIF – A computer-program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological Measurement*, *18*(3), 300-300.

Fidalgo, A. M. (1996). *Funcionamiento diferencial de los ítems*. En J. Muñiz (Coord.): Psicometría (pp. 371-455). Madrid: Universitas.

*Finch, W. H. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and psychological measurement*, *71*(4), 663-683.

*Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied Measurement in Education*, *29*(1), 30-45.

*Finch, W. H. y French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and psychological Measurement*, *67*(4), 565-582.

*Finch, W. H. y 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.

Forbes, D., Hawthorne, G., Elliott, P., McHugh, T., Biddle, D., Creamer, M., & Novaco, R. W. (2004). A concise measure of anger in combat-related posttraumatic stress disorder. *Journal of Traumatic Stress*, *17*(3), 249-256.

*French, A. W. y 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.

*French, B. F. y Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, *47*(3), 299-317.

French, B. F., Finch, W. H. y Valdivia, J. A. (2016). Differential Item Functioning on mathematics items using multilevel SIBTEST. *Psychological Test and Assessment Modeling*, *58*(3), 471-483.

*French, B. F. y Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*(3), 373-393.

Gelin, M. N. y Zumbo, B. D. (2007). Operating characteristics of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation for short scales. *Journal of Modern Applied Statistical Methods*, *6*, 573-588.

Gómez, J. e Hidalgo, M. D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de psicología*, *74*(3), 3-32.

Gómez, J., Hidalgo, M. D., Padilla, J. L. y González, A. (2005). Desarrollo informático para la utilización de la regresión logística como técnica de detección del DIF. *Demostración informática presentada al IX Congreso de Metodología de las Ciencias Sociales y de la Salud, Granada*.

Gómez-Benito, J. e Hidalgo, M. D. (2007). Comparación de varios índices del tamaño del efecto en regresión logística: Una aplicación en la detección del DIF.

Comunicación presentada en el X Congreso de Metodología de las Ciencias Sociales y de la Salud, Barcelona (pp. 6-9).

Gómez-Benito, J. y Navas, M. J. (1996). Detección del funcionamiento diferencial del ítem: Purificación paso a paso de la habilidad. *Psicológica*, *17*, 397-411.

Gomez-Benito, J. y Navas, J. (1998). Impacto y funcionamiento diferencial de los ítems respecto al género en una prueba de aptitud numérica. *Psicothema*, *10*(3), 685-696.

*Gómez-Benito, J. y Navas-Ara, M. J. (2000). A Comparison of $\chi 2$, RFA and IRT Based Procedures in the Detection of DIF. *Quality y Quantity*, *34*(1), 17-31.

Gómez-Benito, J., Hidalgo, M. y Guilera, G. (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles del Psicólogo*, *31*(1), 75-84.

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

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

González, A., Padilla, J. L., Hidalgo, M. D., Gómez-Benito, J. y Benítez, I. (2011). EASYDIF: Software for analysing differential item functioning using the Mantel-Haenszel and standardization procedures. *Applied Psychological Measurement*, *35*(6), 483-484.

González-Betanzos, F. y Abad, F. J. (2012). The effects of purification and the evaluation of Differential Item Functioning with the likelihood ratio test. *Methodology*, *8*, 134-145.

Guilera, G. (2009). *El funcionament diferencial del ítem: Una aproximació bibliomètrica i metaanalítica*. Universidad de Barcelona, Barcelona.

Guilera, G., Gómez-Benito, J. e 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. y 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.

*Güler, N. y Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement, 46*(3), 314-329.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*(3), 229-244.

Hambleton, R. K. (1996). Adaptación de tests para su uso en diferentes idiomas y culturas: fuentes de error, posibles soluciones y directrices prácticas. En J. Muñiz (Ed.), *Psicometría* (pp. 207-238). Madrid: Universitas.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hambleton, R. K. y Rodgers, J. (1995). *Item bias review*. ERIC Clearinghouse on Assessment and Evaluation, the Catholic University of America, Department of Education.

Hambleton, R. K. Yu, J. y Slater, S. C. (1999). Fieldtest of the ITC Guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment, 15*(3), 270.

Hambleton, R. K., Merenda, P. y Spielberger, C. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence Erlbaum Publishers.

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.

*Herrera, A. N. y Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality y Quantity, 42*(6), 739-755.

Hessen, D. J. (2003). *Differential item functioning: Types of DIF and observed score based detection methods*. Dissertation (supervisors: G.J. Mellenbergh y K. Sijtsma). Amsterdam: University of Amsterdam.

Hidalgo, M. D. y Gómez-Benito, J. (2000). Comparación de la eficacia de la regresión logística politómica y análisis discriminante logístico en la detección del DIF no uniforme. *Psicothema, 12* (Suplemento), 298-300.

*Hidalgo, M. D. y 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.

*Hidalgo, M. D. y Gómez-Benito, J. (2006). Nonuniform DIF detection using discriminant logistic analysis and multinomial logistic regression: a comparison for polytomous items. *Quality y Quantity, 40*(5), 805-823.

Hidalgo, M. D. y Gómez-Benito, J. (2010). *Education measurement: Differential item functioning*. En P. Peterson, E. Baker, y B. McGaw (Eds.), International Encyclopedia of Education (3rd edition), *4*, 36-44. USA: Elsevier - Science y Technology.

Hidalgo, M. D., Gómez-Benito, J. y Padilla, J. L. (2005). Regresión logística: alternativas de análisis en la detección del funcionamiento diferencial del ítem. Psicothema, 17(3), 509-515.

Hidalgo, M. D. y López-Pina, J. A. (2000). Funcionamiento diferencial de los ítems: Presente y perspectivas de futuro. *Metodología de las Ciencias del Comportamiento, 2*(2), 167-182.

*Hidalgo, M. D. y López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparision between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Hidalgo, M. D., Gómez-Benito, J. y Zumbo, B. D. (2008).Efficacy of R-square and Odds-Ratio effect size using Discriminant Logistic Regression for detecting DIF in polytomous items. *En la 6º Conferencia de International Test Commission*, 14- 16 July, Liverpool, UK.

*Hidalgo, M. D., Gómez-Benito, J. y Zumbo, B. D. (2014). Binary Logistic Regression Analysis for Detecting Differential Item Functioning: Effectiveness of R 2 and Delta Log Odds Ratio Effect Size Measures. *Educational and Psychological Measurement, 74*(6), 927-949.

Hidalgo, M. D., López-Martínez, M. D., Gómez-Benito, J. y Guilera, G. (2016). A comparison of discriminant logistic regression and Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (IRTLRDIF) in polytomous short tests. *Psicothema, 28*(1), 83-88.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the military testing association, 1*, 282-287.

Holland, P. W. y Thayer, D. T. (1988). *Differential item performance and the MantelHaenszel procedure*. En H. Wainer y H. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Holland, P. W. y Wainer, H. (1993). Preface En P. W. Holland y H. Wainer (Eds), *Differential ítem Functioning*. New Jersey: Lawrence Erlbaum Associates, Inc.

Horswill, S. C., Desgagné, G., Parkerson, H. A., Carleton, R. N. y Asmundson, G. J. (2016). A psychometric evaluation of hierarchical and oblique versions of five variants of the Posttraumatic Growth Inventory. *Psychiatry Research, 246*, 438-446.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley.

IBM Corp. Released (2013). *IBM SPSS Statistics for Windows, Version 22.0.* Armonk, NY: IBM Corp.

Jafari, P., Allahyari, E., Salarzadeh, M. y Bagheri, Z. (2016). Item-level informant discrepancies across obese–overweight children and their parents on the

PedsQL™ 4.0 instrument: an iterative hybrid ordinal logistic regression. *Quality of Life Research, 25*(1), 25-33.

Jafari, P., Stevanovic, D. y Bagheri, Z. (2016). Cross-cultural measurement equivalence of the KINDL questionnaire for quality of life assessment in children and adolescents. *Child Psychiatry & Human Development, 47*(2), 291-304.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achivement?. *Harvard Educational Review, 39*(1), 1-123.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

*Jin, Y., Myers, N. D. y Ahn, S. (2014). Complex Versus Simple Modeling for DIF Detection: When the Intraclass Correlation Coefficient (ρ) of the Studied Item Is Less Than the ρ of the Total Score. *Educational and Psychological Measurement, 74*(1), 163-190.

*Jodoin, M. G. y Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349.

Jöreskog, K. G. y Sörbom, D. (2006). *Lisrel 8 (version 8.8)*. Chicago, Illinois: Scientific Software International, Inc.

Kagan, J. (1975). *The Magical Aura of the IQ*. En A. Montagu (Ed.), *Race any IQ* (pp. 52-58). New York: Oxford.

*Kanjee, A. (2007). Using logistic regression to detect bias when multiple groups are tested. *South African Journal of Psychology, 37*(1), 47-61.

*Kim, J. y Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement, 73*(3), 458-470.

Kim, S. H. y Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied psychological measurement, 15*(3), 269-278.

Kim, S. H. y Cohen, A. S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis. *Applied Pyhologital Measurement*, *16*(2), 158-158.

Kirk, R. E. (1996). Practical Significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

Klieme, E. y Stumpf, H. (1991). DIF: a computer program for the analysis of differential item performance. *Educational and psychological measurement*, *51*(3), 669-671.

Kok, F. (1988). Item bias and test multidimensionality. *Latent trait and latent class models* (pp. 263-275). New York: Plenum.

*Kristjansson, E., Aylesworth, R., Mcdowell, I. y Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935-953.

*Lei, P. W., Chen, S. Y. y Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement*, *43*(3), 245-264.

Li, H. H. y Stout, W. (1994). *SIBTEST: A FORTRAN-V Program for Computing the Simultaneous Item Bias DIF Statistics*. Urbana-Champaign, IL: University of Illinois, Department of Statistics.

*Li, H. H. y Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647-677.

*Li, Y., Brooks, G. P. y 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.

*Li, Z. (2014). Power and sample size calculations for logistic regression tests for differential item functioning. *Journal of Educational Measurement*, *51*(4), 441-462.

Linn, R. L., Levine, M. V., Hastings, C. N. y Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*(2), 159-173.

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

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16* (12), 317-323.

Magis, D., Béland, S., Tuerlinckx, F. y 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.

*Magis, D., Tuerlinckx, F. y De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics, 40*(2), 111-135.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 690-700.

Mantel, N. y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719-748.

Martínez-Arias, R., Hernández-Lloreda, M. J. y Hernández-Lloreda, M. V. (2006). *Psicometría*. Madrid: Alianza Editorial.

Matsumoto, D. y van de Vijver, F. J. R. (Eds.) (2011). *Cross-cultural research methods in psychology*. Nueva York: Cambridge University Press.

Mazor, K. M., Clauser, B. E. y Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*(2), 284-291.

*Mazor, K. M., Hambleton, R. K. y Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*(4), 357-367.

McKelvey, R. D. y Zavoina, L. (1975). A statistical model for the analysis of ordinal dependent variables. *Journal of Mathematical Sociology, 4*, 103-120.

Mellenbergh, G. J. (1982). Contingency tables models for assesing item bias. *Journal of Educational Statistics*, *7*, 105-108.

Menard, S. (1995). *Applied Logistic Regression Analysis: Sage University Series on Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage.

Messick, S. (1989). Validity. En R. Linn (Ed.). *Educational measurement* (3rd edition, pp. 13-104). Washington, DC: American Council on Education.

Miller, T. R.; Spray, J. A. y Wilson, A. (1992). *A comparison of three methods for identifying nonuniform DIF in a polytomously scored item test.* Ponencia presentada en la Reunión de la Sociedad Psicométrica, Columbus, OH.

Miller, T. y Spray, J. (1993). Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scored Items. *Journal of Educational Measurement, 30*(2), 107-122.

Millsap, R. E. y Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement*, *16*(4), 389-402.

Monahan, P. O., McHorney, C. A., Stump, T. E. y Perkins, A. J. (2007). Odds-ratio, Delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Behavioral Statistics*, *32*(1), 92-109.

Muñiz, J. (1996) Fiabilidad. En J. Muñiz (Coord.), *Psicometría* (pp 1-48). Madrid: Editorial Universitas.

Muñiz, J. (1998) *Teoría clásica de los test*. Madrid: Ediciones Pirámide.

Muñiz, J. y Hambleton, R. K. (1996). Directrices para la traduccion y adaptacion de los tests. *Papeles del Psicólogo*, *66*(1), 63-70.

Muñiz, J., Elosua, P. y Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema*, *25*(2), 151-157.

Muthén, L. K., y Muthén, B. O. (1998, 2007). *MPLUS statistical analysis with latent variables. User's Guide.* Los Angeles, CA: Muthén and Muthén.

Narayanan, P. y Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328.

*Narayanan, P. y Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257-274.

Navas-Ara, M. J. (Coord.) (2001). *Métodos, diseños y técnicas de investigación psicológica.* Madrid: Editorial UNED.

*Navas-Ara, M. J. y Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment*, *18*(1), 9-15.

*Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, *27*(4), 286-300.

Ordóñez, X. G. y Romero, S. J. (2007). XS-DIF: Program for analysis of Differential Item Functioning in Excel. *Psicothema*, *19*(1), 171-172.

Osterlind, S. J. y Everson, H. T. (2009). *Differential item functioning* (2nd edition). Thousand Oaks, California: Sage Publications, Inc.

Paek, I. y Wilson, M. (2011). Formulating the Rasch Differential Item Functioning Model under the Marginal Maximum Likelihood Estimation Context and Its Comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*, 1023-1046.

Pao, M. L. (1985). Lotka's law: a testing procedure. *Information Processing and Management*, *21*(4), 305-320.

*Pei, L. K. y Li, J. (2010). Effects of Unequal Ability Variances on the Performance of Logistic Regression, Mantel-Haenszel, SIBTEST IRT, and IRT Likelihood Ratio for DIF Detection. *Applied Psychological Measurement*, *34*(6), 453-456.

Pelechano, V. (2000). *Psicología sistemática de la personalidad*. Barcelona: Ariel.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement*, *29*(2), 150-151.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education*, *20*, 335-355.

Penfield, R. D., Alvarez, K. y 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*, 61-78.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, *14*(3), 235-259.

Potenza, M. y Dorans, N. (1995). DIF assessment for politomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, *19*, 23-37.

Prieto, G. y Delgado, A. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, *31*(1), 67-74.

R Development Core Team (2008). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207.

Raju, N. S. (1995). *DFITPU: A FORTRAN program for calculating DIF/ DTF*. Atlanta: Georgia Institute of Technology.

Raju, N. S. (1999). *DFITP5: A Fortran program for calculating dichotomous DIF/DTF*. Chicago, IL: Illinois Institute of Technology.

Ramsay, J. O. (2000). *TestGraph: A program for the graphical analysis of multiple choice and test questionnaire data*. Manual no publicado.

Rey-Rocha, J., Martín-Sempere, M. J., Garzón, B. (2002). Research productivity of scientists in consolidated vs. non-consolidated teams: The case of Spanish university geologists. *Scientometrics*, *55*(1), 137-156.

Ribeiro, C. C., Gómez-Conesa, A. e Hidalgo, M. D. (2010). Metodología para la adaptación de instrumentos de evaluación. *Fisioterapia*, *32*(6), 264-270.

Robin, F. (2001). *STDIF: Standardization-DIF analysis program*. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

*Robitzsch, A. y Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning the case of Mantel-Haenszel and Logistic Regression analysis. *Educational and Psychological Measurement*, *69*(1), 18-34.

Rogers, H. J. y Hambleton, R. K. (1994). MH: A fortran 77 program to compute the Mantel-Haenszel statistic for detecting differential item functioning. *Educational and psychological measurement*, *54*(1), 101-104.

*Rogers, H. J. y Swaminathan, H. (1993). A comparison of Logistic Regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Rogers, H. J., Swaminathan, H. y Hambleton, R. K. (1993). *DICHODIF: A FORTRAN program for DIF analysis of dichotomously scored item response data*. Amherst, MA: University of Massachusetts.

Roussos, L. y Stout, W. (1996). A multidimensionality- based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355-371.

Rudner, L. M. (1977). *An approach to biased item identiJication using latent trait measurement theory.* Documento presentado en la reunión anual de American Educational Research Association, Nueva York.

Rudner, L. M., Getson, P. R. y Knight, D. L. (1980a). A montecarlo comparison of seven biased item detection techniques. *Journal of Educational Measurement*, *17*, 1-10.

Rudner, L. M., Getson, P. R. y Knight, D. L. (1980b). Biased item detection techniques. *Joumal of Educational Statistics*, *5*,213-233.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric*. Monograph Supplement 17.

*Schauberger, G. y Tutz, G. (2016). Detection of differential item functioning in Rasch models by boosting techniques. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 80-103.

Scheuneman, J. D. (1982). A posteriori analyses of biased items. En R. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.

Scheuneman, J. D. y Bleistein, C.A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education*, *2*, 255-275.

*Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., ... y EORTC Quality of Life Group. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of clinical epidemiology*, *62*(3), 288-295.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., Graeff, A., Groenvold, M., …, Sprangers, M. A. G. (2010). Differential Item Functioning (DIF) analyses of health related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes*, *8*, 81.

Shealy, R. y 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.

*Shih, C. L., Liu, T. H. y Wang, W. C. (2014). Controlling type I error rates in assessing DIF for logistic regression method combined with SIBTEST regression correction procedure and DIF-free-then-DIF strategy. *Educational and Psychological Measurement*, *74*(6), 1018-1048.

Sireci, S. G. y Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal of Theory and Practice*, *19*(2-3), 170-187.

Stern, W. (1914). *The psychological methods of testing intelligence*. Baltimore: Warwick y York.

Sternberg, R. J. (2001). Where was it published?. *APS Observer*, *14*(7).

Stout, W. F. y Roussos, L. A. (1995). *SIBTEST user manual.* Urbana-Champaign: University of Illinois, Department of Statistics.

Stout, W. y Roussos, L. (1999). *Dimensionality-based DIF/DBF package.* William Stout Institute for Measurement: University of Illinois, 11.

*Swaminathan, H. y Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Tian, F. (1999). *Detecting DIF in polytomous item responses*. University of Ottawa: Canada.

Teresi, J. A. (2006). Overview of quantitative measurement methods: equivalence, invariance, and differential item functioning in health applications. *Medical Care*, *44*, S39-49.

Thissen, D. (2001). *IRTLRDIF v. 2.0 b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning.* Chapel Hill, NC: LL Thurstone Psychometric Laboratory.

Thissen, D., (2003). *Multilog User's Guide: Multiple Categorical Item Analysis and Test Scoring Using Item Response Theory.* Scientific Software, Chicago.

Thissen, D., Steinberg, L. y Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118.

Thissen, D., Steinberg, L. y Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In Wainer, H. y Braun, H.I. (eds.) *Test Validity*. Pp. 147-169. Hillsdale, N.J.: Erlbaum.

Thomas, D. R. y Zumbo, B. D. (1998). *Variable importance in logistic regression based on partitioning an R-squared measure*. Presentado en las reuniones de la Sociedad Psicométrica, Urbana, IL.

*Tutz, G. y Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21-43.

Van de Vijver, F. J. R. y Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*(4), 263-279.

*Vaughn, B. K. y Wang, Q. (2010). DIF trees: Using classification trees to detect differential item functioning. *Educational and Psychological Measurement*, *70*(6), 941-952.

Waller, N. G. (1998a). EZDIF: Detection of uniform and nonuniform differential item functioning with the Mantel-Haenszel and logistic regression procedures. *Applied Psychological Measurement*, *22*(4), 391.

Waller, N. G. (1998b). LINKDIF: Linking item parameters and calculating IRT measures of differential functioning of items and tests. *Applied psychological measurement*, *22*(4), 392.

Wang, W. C, Shih, C. L. y Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713-731.

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

*Welkenhuysen-Gybels, J. (2004). The performance of some observed and unobserved conditional invariance techniques for the detection of differential item functioning. *Quality and quantity*, *38*(6), 681-702.

*Welkenhuysen-Gybels, J. y Billiet, J. (2002). A comparison of techniques for detecting cross-cultural inequivalence at the item level. *Quality y Quantity*, *36*(3), 197-218.

*Whitmore, M. L. y Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement*, *59*(6), 910-927.

*Woods, C. M. y Harpole, J. (2015). How item residual heterogeneity affects tests for differential item functioning. *Applied Psychological Measurement*, *39*(4), 251-263.

Zulueta, M. A. y Bordons, M. (1999). La producción científica española en el área cardiovascular a través del Science Citation Index (1990-1996). *Revista Española de Cardiología*, *52*(10), 751-764.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters.

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

Zumbo, B. D. y Thomas, D. R. (1997) A measure of effect size for a model-based approach for studying DIF. *Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B.C.*

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide?. *Journal of Educational Statistics*, *15*(3), 185-197.

Zwick, R., Donoghue, J. R. y Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233-251.