Efecto del tamaño de muestra y la razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems

Aura Nidia Herrera Rojas

EFECTO DEL TAMAÑO DE MUESTRA Y LA RAZON DE TAMAÑOS DE MUESTRA EN LA DETECCIÓN DE FUNCIONAMIENTO DIFERENCIAL DE LOS ITEMS.

Tesis doctoral

AURA NIDIA HERRERA ROJAS

Directora: JUANA GOMEZ BENITO

Universidad de Barcelona

Facultad de Psicología

Departamento de Metodología de las Ciencias del Comportamiento

Barcelona, 2005

INDICE DE CONTENIDOS

INDICE DE TABLAS

INDICE DE FIGURAS

INTRODUCCION

Desde comienzos del siglo pasado Stern (1914) había mostrado que el rendimiento en pruebas de inteligencia variaba según la clase social y Binet & Simon (1916) habían eliminado algunos ítems en la nueva versión de su prueba de inteligencia porque eran sensibles a efectos culturales (Camilli & Shepard, 1994), sin embargo, fue sólo después del importante auge en la construcción y uso de pruebas psicológicas, que encontró su punto máximo durante las dos guerras mundiales y la década subsiguiente, cuando empezó a popularizarse la idea de que éstas podían resultar injustas, sesgadas y favorecedoras de unas clases sociales o grupos étnicos y culturales particulares (Anastasi, 1974). Esa percepción se basaba en la observación de diferencias sistemáticas en los resultados arrojados por las pruebas, entre personas pertenecientes a distintos grupos según género, raza, clase socioeconómica o cultura. Sin embargo, una revisión de la literatura sobre el tema muestra que fueron los trabajos de Eelles, Havighurst, Herrick & Tyler (1951) y Jensen (1969) los que significaron un importante punto de partida para el avance de la investigación sobre lo que hoy se conoce como funcionamiento diferencial de los ítems (DIF, abreviatura de Differential Items Functioning)[1] y funcionamiento diferencial de los test (FDT).

El primero de estos dos trabajos es considerado por la mayoría de autores sobre el tema (Camilli y Shepard, 1994; Hidalgo Montesinos & Gómez Benito, 1996; Fidalgo, 1996a; Ferreres, 1998, entre otros), como la investigación pionera sobre sesgo de los ítems y de las pruebas psicológicas. Eelles, Havighurst, Herrick y Tyler (1951) mostraron empíricamente y con un gran número de ítems de pruebas de inteligencia, que algunos de ellos se dejaban afectar por diferencias culturales. A pesar de sus hallazgos, la discusión sobre el sesgo de las pruebas alcanzó su punto más alto solamente en la década de los sesenta; dentro del contexto del movimiento por la defensa de los derechos civiles y por consiguiente, de la igualdad de oportunidades educativas y laborales; ámbitos en los que las pruebas psicológicas eran ampliamente usadas y sus resultados constituían uno de los argumentos más fuertes para la toma de decisiones relacionadas con asignación de cupos (Cole, 1993).

Hasta aquí la noción de sesgo se asociaba a cualquier diferencia sistemática en los resultados de las pruebas entre grupos diferentes; en consecuencia, el término tenía una connotación negativa equiparable con injusticia, parcialidad e inequidad contra los grupos minoritarios o menos favorecidos. Angoff (1993), Cole (1993), Holland & Wainer (1993) y Fidalgo (1996a), entre otros, están de acuerdo en que esta asociación, dominante durante la década de los sesenta y que tuvo am-

[1] Por costumbre y facilidad en este documento se utilizará la abreviatura en inglés, DIF, en vez de FDI, expresión correcta para un texto escrito en castellano. En todos los demás casos se usarán las abreviaturas o siglas en castellano

plias implicaciones de tipo social, se debió en gran parte a un conflicto semántico y a una confusión en el uso del lenguaje común y del lenguaje técnico: público y psicólogos estaban usando el mismo término –sesgo– pero para los primeros estaba cargado de contenido social y político con una connotación claramente negativa, mientras que para los segundos estaba cargado de contenido técnico, y aunque la connotación no era buena, hacía referencia básicamente a 'características técnicas no óptimas' (Cole, 1993, p.27) y no a injusticia social.

Para Fidalgo, (1996a) esta discusión se basa en la falacia igualitaria (p.376) según la cual los hombres somos iguales de manera que las pruebas o cualquier instrumento que pusiera en evidencia diferencias entre grupos humanos, resultaba discriminatorio y sesgado. Sorprendentemente, un claro ejemplo de esta confusión se presentó en Estados Unidos casi dos décadas después –en 1984 cuando se había avanzado en la claridad conceptual y en el desarrollo de técnicas de detección del DIF- en el tristemente famoso caso de *Golden Rule Insurance Company* contra el Departamento de Seguros de Illinois y el *Educational Testing Service* (ETS), que derivó en lo que se conoció como la regla de oro (*Golden rule*). De acuerdo con ésta, el ETS no incluiría en sus pruebas, ítems que mostraran diferencias de proporción de acierto superiores a 0,15 entre blancos y negros. La decisión provocó una serie de reacciones entre las que pueden citarse Bond (1987), Faggen (1987), Lim & Drasgow (1987), y la misma Asociación Americana de Psicología (American Psychological Association - APA), debido a las posibles implicaciones de tipo legal y político, y a algunos intentos por generalizar la regla en otros contextos. Lim & Drasgow (1990) y Fidalgo (1996a) citan algunos de estos intentos.

El segundo trabajo antes mencionado (Jensen, 1969) apareció publicado en medio de la acalorada discusión de finales de los sesenta y su importancia radica en la enorme polémica que desató y su efecto sobre el desarrollo de técnicas para la detección de DIF. El autor afirmaba que la inteligencia era heredada y que en consecuencia las diferencias observadas en las pruebas entre grupos raciales podían explicarse genéticamente, argumentos que avivaron la discusión entre las explicaciones genéticas y las de determinantes ambientales y sociales de las diferencias en cociente intelectual. Los partidarios del segundo tipo de explicaciones atribuían en gran medida las diferencias entre grupos, al sesgo de las pruebas. En lo que tiene que ver con psicometría propiamente, sin embargo, la importancia del trabajo estuvo en que 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 artificios generados por el instrumento mismo, lo cual implicaba además de esclarecer conceptos, generar técnicas que permitieran evaluar el posible efecto de las pruebas, en las diferencias observadas entre grupos. Muñiz (1997) hace notar cómo las publicaciones psicométricas especializadas de las décadas de los cincuenta y los sesenta del pasado siglo, y la edición de 1966 de los *Standards for Educational and Psychological Test and Manuals* ignoraron por completo el tema, y es sólo a partir de los 70 cuando la comunidad psicométrica se apropia de la discusión que hasta el momento se había mantenido en las esferas legal, política, social y de la teoría psicológica.

En las últimas tres décadas se ha observado un importante incremento en las publicaciones de textos y artículos científicos sobre el tema y se ha llegado a conceptualizaciones mucho más precisas. Otro trabajo de Jensen (1980) contribuyó a esclarecer el significado del término “sesgo” buscando despejarlo de las connotaciones éticas, sociales y políticas para entenderlo como un problema técnico; pero sin lugar a dudas, una propuesta que contribuyó a superar el conflicto semántico fue la de Holland & Thayer (1988) quienes sugirieron cambiar el término “sesgo” por el de Funcionamiento Diferencial de los Ítems (DIF). Así, el primer término se usaría para referirse a un ‘juicio informado’ (Holland & Wainer, 1993, p. xiv) que además de los datos estadísticos sobre el ítem, tomara en cuenta el objetivo de la prueba y la información de tipo histórico, social o cultural que posiblemente explicara su funcionamiento. Aunque la nueva expresión resultaba en opinión de Angoff, (1993) más larga y menos comprensible, rápidamente se fue generalizando en las publicaciones especializadas a través del uso de su abreviatura DIF, para referirse a la observación desapasionada del hecho de que algunos ítems pueden mostrar propiedades psicométricas diferentes para grupos diferentes. Debe anotarse, sin embargo, que el relativo acuerdo sobre los términos no supuso superar la preocupación inicial sobre las diferencias entre grupos raciales, étnicos o de género, en cuanto al rendimiento en las pruebas y, por ende, en las oportunidades educativas y laborales; prueba de esto puede encontrarse en los más recientes escritos de Jensen (1998, 2000) o el número especial de la revista *Inteligencia* editado por Gottfredson (1997).

Además de la claridad conceptual, en el mismo periodo de tiempo se han identificado categorías diferentes de DIF y se han propuesto múltiples procedimientos estadísticos y estrategias metodológicas de estimación (Anastasi y Urbina, 1998; Ferreres, 1998; Muñiz, 1997, 1998). La literatura especializada de las últimas décadas muestra un número considerable de esfuerzos por identificar las ventajas y limitaciones de las diferentes técnicas de estimación, así como por encontrar las condiciones en las cuales resultan más o menos adecuadas; en esta línea se ha evaluado el efecto de múltiples factores sobre la potencia y el error tipo I de los diferentes estadísticos propuestos para la detección de DIF, en condiciones más o menos específicas. Entre los factores evaluados se pueden mencionar el tamaño de los grupos de examinados, la longitud de la prueba, la proporción de ítem con DIF en la misma, la magnitud del DIF, la distribución de la magnitud de atributo de los examinados y las características (parámetros) de los ítems, entre otros muchos.

El tamaño de muestra y su efecto sobre la potencia y el error tipo I de los estadísticos no es un asunto trivial, por el contrario puede tener profundas implicaciones prácticas; en general se sabe que la potencia aumenta con el tamaño de muestra pero generalmente el error tipo I también lo hace. En los estudios de detección de DIF un estadístico con bajo poder permitiría que algunos ítems con DIF, y sobre todo los de magnitud moderada, quedaran incluidos en las pruebas con las implicaciones técnicas, sociales y éticas que ello pueda tener; pero un estadístico con un alto error tipo I conduciría al rechazo de elementos que pueden tener propiedades adecuadas, elevando innecesariamente los costos en la construcción de instrumentos, costos que pueden ser importantes dependiendo del tipo de instrumento del que se trate. Jodoin & Huff (2001) hacen notar que

este asunto cobra mayor importancia con la introducción de tecnologías como video, efectos de sonido y otras que buscan simular situaciones reales en las pruebas. Resulta evidente entonces la necesidad de usar procedimientos que tengan elevada potencia en la detección de DIF y mantengan controlado el error tipo I, y para el usuario de los procedimientos o constructor de instrumentos, resulta valioso saber en qué condiciones prácticas el procedimiento mantiene estas características o empieza a perderlas.

Además, la estimación de DIF supone la comparación de dos grupos de examinados que suelen llamarse grupos "de Referencia" y "Focal", denominando con el último nombre al grupo de interés y con el primero, al grupo de comparación. En la práctica generalmente se considera que el grupo focal es desfavorecido por las pruebas y frecuentemente es minoritario; así, cuando se busca identificar los ítems con DIF en instrumentos utilizados para toma de decisiones, puede ocurrir que la diferencia de tamaños de los dos grupos sea importante. No obstante, hasta el momento no se encuentran estudios que evalúen el efecto de tal diferencia sobre el funcionamiento de los procedimientos; como se mostrará en los capítulos siguientes, las investigaciones sobre el tema trabajan con grupos de tamaño similar o, en todo caso, no evalúan de manera sistemática el posible efecto de este factor.

Este trabajo centra su atención en la evaluación del efecto del tamaño de muestra y de la razón de tamaños de los grupos, definida como $r = nr/nf$, donde nr es el tamaño del grupo de referencia y nf es el tamaño del grupo focal, sobre el funcionamiento de los estadísticos más frecuentemente utilizados hoy en la detección de DIF de ítems dicótomos. En concreto, este trabajo se ocupa de evaluar el efecto de estos factores sobre el error tipo I y la potencia de dos procedimientos basados en el análisis de tablas de contingencia (Mante-Haenszel y Regresión Logística) y uno basado en la teoría de Respuesta al Item (χ^2 de Lord) para la detección de DIF. Se espera, de esta manera, contribuir en la identificación de las condiciones prácticas en las cuales los procedimientos tienen su mejor desempeño y aquellas en las cuales su uso no resulta recomendable.

La estrategia metodológica elegida para el cumplimiento de este propósito general está compuesta por tres experimentos de Monte Carlo De acuerdo con Gentle (2003), Ross (1999) y Spence (1983), entre otros, una de las principales ventajas de este tipo de aproximación es que permite abordar problemas para los cuales no resulta factible o es muy difícil, encontrar la solución de forma analítica y puede además, ilustrar este tipo de solución. Dentro de la investigación psicométrica en general y sobre procedimientos para identificar DIF, en particular, este tipo de procedimientos se ha empleado frecuentemente para estudiar las distribuciones muestrales de nuevos estadísticos, para evaluar el comportamiento de los mismos cuando no se cumplen algunos de los supuestos que los sustentan (conocidos como estudios de robustez) o en condiciones prácticas en las cuales no se conoce su comportamiento.

En la categoría denominada “métodos de Monte Carlo" se incluyen una buena cantidad de técnicas empleadas para simular muchos experimentos, con el fin de obtener algunas inferencias sobre un proceso descrito por medio de un modelo estadístico. De acuerdo con Dorn & Greenberg (1970a), Nerlove (1997) y Ross, (1999), un experimento de Monte Carlo exige dos procedimientos fundamentales, de los cuales depende en gran parte, el éxito del experimento: generar números aleatorios ajustados a una ciertas características distribucionales y desarrollar un modelo que haga que el valor esperado de las distribuciones estadísticas establecidas por éste a lo largo de las réplicas sea igual al parámetro que se desea estudiar en el experimento. La aproximación más práctica para la generación de grandes cantidades de números aleatorios son los algoritmos recursivos que generan series de números que no son estrictamente aleatorias pero tienen la apariencia de serlo y puede comportarse de manera suficientemente similar al de variables verdaderamente aleatorias; a estas series de números se las conoce como pseudoaleatorios. Para generarlos, en la actualidad se dispone de múltiples algoritmos implementados en diversos lenguajes de programación, hoy casi todo paquete estadístico o manejador de datos, tiene incorporado un generador de número aleatorios. La mayoría de tales generadores implementa una clase de algoritmos conocidos como congruenciales lineales mixtos[2] que se basan en la operación *módulo* (operación matemática que devuelve el residuo de una división) partiendo de un valor inicial llamado “*semilla*”. Siguiendo un algoritmo de este tipo se generan series de números aleatorios con distribución uniforme en el rango [0 -1]; sin embargo, con frecuencia no es esta serie de datos la que interesa dentro del experimento sino que es necesario simular algunas características distribucionales específicas; entonces es necesario el segundo procedimiento antes mencionado. Una vez generada una serie de números distribuidos uniformemente, estos se trasforman a una distribución cualquiera de acuerdo con los objetivos y requerimientos del experimento y las especificaciones del diseño (Dorn & Greenberg, 1970; Nerlove, 1997; Ross, 1999; Sobol & Myshetskaya, 2003).

Uno de los principales usos de los experimentos de Monte Carlo en psicometría es la evaluación de la robustez de los estadísticos, es decir, la evaluación del comportamiento del estadístico cuando no se cumplen los supuestos en los que se sustenta. En la investigación sobre métodos de detección de DIF, además de los estudios de robustez y el análisis de las propiedades distribucionales de los estadísticos en condiciones de violación de supuestos, los estudios de Monte Carlo se han utilizado para estudiar el efecto de algunos factores específicos que pueden presentarse en la práctica profesional, sobre el funcionamiento del estadístico en términos de su potencia para detectar ítem DIF y su error tipo I. Este es el caso del presente trabajo que busca evaluar el posible efecto del tamaño de muestra y la razón de tamaños sobre tres estadísticos frecuentemente utili-

[2] Algunas descripciones de este tipo de procedimientos se pueden encontrar en Dorn & Greenberg (1970a), Press, Flannery, Teukolsky & Vetterling (1992), Ross, (1999) y Tejada (2002)

zados para detectar DIF. A través de tres estudios de Monte Carlo se pretende lograr objetivos parciales que conduzcan al cumplimiento de este propósito global. En los dos primeros estudios se evalúa el efecto del tamaño de muestra y la razón de tamaños de los grupos sobre la potencia y el error tipo I del Mantel-Haenszel (MH) y la Regresión Logística (RL), respectivamente, tomando en consideración el tipo de DIF y los parámetros de los ítems. El tercer estudio evalúa el efecto de estos mismos factores y de la longitud de prueba sobre el χ^2 de Lord (1980). En todos los casos la variable respuesta es la tasa de detección de ítems con DIF (potencia) y de ítems sin DIF (error tipo I) con diferentes niveles de significación.

Este documento, compuesto por dos partes, presenta la discusión teórica previa a la realización de tales estudios y la descripción de los mismos, su procedimiento, resultado y conclusiones. La primera parte del documento presenta la revisión bibliográfica sobre los aspectos teóricos y metodológicos básicos que apoyan el presente trabajo; esta parte incluye tres capítulos: Conceptos y métodos, procedimientos basados en análisis de tablas de contingencia y procedimientos basados en la Teoría de Respuesta al Ítem (TRI). La segunda parte incluye tres capítulos en cada uno de los cuales se presenta uno de los estudios empíricos antes mencionados.

PRIMERA PARTE:

REVISION TEORICA: CONCEPTOS Y METODOS

Capítulo 1:
PRINCIPIOS CONCEPTUALES Y METODOLOGICOS

Como se mencionó anteriormente después de las acaloradas discusiones de las décadas de los 60 y 70 se ha avanzado de manera importante en la claridad conceptual sobre términos como Sesgo, Impacto, DIF y FDT; y en el diseño, adaptación y evaluación de técnicas para detectarlos. Este capítulo presenta una revisión conceptual sobre estos temas y las características generales de algunas técnicas de detección de DIF. En primer lugar se resumen las principales discusiones y acuerdos en torno a los términos ya mencionados haciendo especial énfasis en el concepto de DIF, se presentan y describen los tipos de DIF y se resume la propuesta explicativa del mismo -teoría de la multidimensionalidad- y las críticas que ha recibido recientemente. Posteriormente se introduce al tema de las técnicas de detección mediante una breve descripción de algunas de ellas que no se tratarán en capítulos posteriores, y la presentación general de los principios, ventajas y limitaciones de aquellas que se describirán con más detalle en los dos capítulos siguientes.

Sesgo, impacto y funcionamiento diferencial

Una vez acogida la propuesta de Holland & Thayer (1988), la expresión DIF se popularizó entre la comunidad académica para referirse a la evidencia empírica, soportada estadísticamente, de que un ítem funciona de manera diferente para grupos diferentes. Hoy el DIF hace referencia al hecho objetivo de que la probabilidad de acertar[3] en un ítem cambia en función del grupo de pertenencia, es decir, se centra en los procedimientos de tipo matemático y estadístico para identificar aquellos ítems que tienen diferente funcionamiento entre personas con igual magnitud del atributo pero que pertenecen a grupos diferentes. Así las cosas, el DIF se puede distinguir del sesgo en dos direcciones. En primer lugar, el sesgo incluye además de la detección de ítems con DIF, la identificación de las razones para que eso ocurra, lo cual trasciende los alcances de los procedimientos estadísticos (Gómez Benito e Hidalgo Montesinos, 1997b; Camilli y Shepard, 1994;). El análisis de sesgo, que constituye la meta en la mayoría de trabajos aplicados, implica además, la identificación del factor o factores que lo producen y la discusión teórica acerca de la relevancia de tales factores en el constructo que pretende medir la prueba o en los propósitos de la medida. En segundo lugar, que un ítem tenga DIF no necesariamente implica que el mismo esté sesgado; una causa frecuente de DIF es que el ítem mida algo más de lo que originariamente pretende medir; si se sustentan la relación de ese 'otro factor' que mide el ítem con el atributo objeto de la medida, y

[3] La expresión "acertar en el ítem" puede suponer que éste tiene una respuesta correcta, lo cual no es cierto en todas los casos, en particular en las denominadas pruebas de ejecución típica. Aunque la expresión "puntuar en el ítem" resulta más incluyente, en la literatura se usa frecuentemente la primera debido a que hasta hace unos pocos años, la mayoría de estudios sobre el tema se han realizado en el contexto de la medición de habilidades, rendimiento y aptitudes.

su relevancia para los propósitos de la prueba, el ítem podría declararse no sesgado (Camilli y Shepard, 1994; Muñiz, 1997). El efecto será que quienes tengan la misma magnitud de atributo en lo que mide el ítem (combinación de factores relevantes) tendrán la misma probabilidad de puntuar y éste no provocará diferencias entre grupos o debidas a factores irrelevantes. Asunto diferente ocurre cuando el "otro factor" que mide el ítem resulta irrelevante para los propósitos de la medida, este tema se tratará con más detalle posteriormente en este mismo capítulo.

De lo expuesto hasta ahora resultan dos conclusiones que, aunque parezcan obvias vale la pena explicitar: En primer lugar, resulta evidente que el uso de los procedimientos para detectar ítems con DIF constituye una tarea necesaria pero no suficiente para el análisis de sesgo de un instrumento de medida. En segundo lugar, el éxito en el uso del término "DIF" y su diferenciación del "sesgo" no invalida las afirmaciones de Jensen (1980): al hablar de sesgo no se está fuera del terreno de lo técnico o académico para entrar en el campo de la ética o la política. La diferencia entre DIF y sesgo, al menos como la entienden la mayoría de los autores hoy, está en que en el primer caso los resultados y decisiones se sustentan en términos de la estadística y la probabilidad, mientras que en el segundo, los argumentos son teóricos (Camilli, 1992, 1993; Shealy & Stout, 1993b; Camilli y Shepard, 1994; Fidalgo, 1996a).

Un asunto diferente pero no de menor importancia es la distinción entre DIF e impacto. Cuando se observan diferencias en la probabilidad de acertar en un ítem de una prueba, entre personas pertenecientes a grupos diferentes, cabe preguntarse ¿tales resultados se deben a diferencias reales en las magnitudes del atributo que pretende medir la prueba, o a otro tipo de variables relacionadas con la ejecución en el ítem pero diferentes al objeto de medida?. Según Millsap & Everson (1993) la identificación de la posible fuente de variación entre grupos permite distinguir el *Funcionamiento Diferencial* del *Impacto*, también denominado *Impacto adverso* por Camilli & Shepard (1994) o *diferencias válidas*, según van de Vijver & Leung (1997). Cuando las diferencias entre los grupos se deben a diferencias en la magnitud de atributo medido, se habla de impacto y cuando se deben a otras variables asociadas a los grupos, se habla de funcionamiento diferencial (Ackerman, 1992). Esta es una precisión conceptual importante por cuanto tiene amplias implicaciones de diferente orden; el caso antes mencionado que dio origen a la regla de oro (*Golden rule*) es tal vez el más famoso ejemplo de confusión entre DIF e impacto. La decisión de excluir los ítems de diferente proporción de acierto entre grupos puede conducir a prescindir de aquellos que son más sensibles a las diferencias reales existentes entre los mismos, esto es, los ítems con alto poder de discriminación. El efecto de tal práctica no necesariamente es la eliminación de los ítems sesgados; en cambio, como lo mostraron Linn y Drasgow (1987) y Marco (1988), se puede estar obrando en detrimento de la confiabilidad y la validez del instrumento.

Si existen diferencias entre grupos en la magnitud del atributo medido por una prueba, y ésta lo mide con precisión, sus resultados deben reflejar tales diferencias; es de esperarse entonces que la probabilidad de acierto sea diferente para personas pertenecientes a grupos diferentes, en los

ítems que tengan la mayor capacidad selectiva y no necesariamente en los ítems sesgados. Angoff, (1993) comentando el caso *Golden rule*, y utilizando el término "sesgo" en un sentido diferente al expuesto antes, afirma que existen las diferencias entre blancos y negros y que allí hay un claro sesgo pero en el tratamiento que históricamente habían tenido los negros en la sociedad norteamericana, el cual producía condiciones educativas diferentes que conducían a desempeño más pobre en comparación con los blancos; las pruebas entonces recogen esa información que se manifiesta en diferencias de puntajes.

Uno de los ejemplos más comprehensivos e ilustrativos para entender la noción de sesgo o funcionamiento diferencial, es el que presenta Muñiz (1998): "*un metro estará sistemáticamente sesgado si no proporciona la misma medida para dos objetos o clases de objetos que de hecho miden lo mismo, sino que sistemáticamente perjudica a uno de ellos*" (p.236). Esta afirmación tiene varias implicaciones que permiten aclarar la noción de funcionamiento diferencial. En primer lugar, para hablar de funcionamiento diferencial es necesario tener por lo menos dos objetos o clases de objetos, en la práctica las técnicas de detección de DIF trabajan con dos grupos de comparación conocidos como grupo de referencia y grupo focal. En segundo lugar, para afirmar que un instrumento o un ítem funcionan diferencialmente es necesario tener alguna evidencia de que los grupos que se comparan tienen el mismo nivel del atributo medido, para Fidalgo (1996a) esta exigencia constituye la paradoja de los procedimientos para evaluar DIF, puesto que conduciría a que "sólo es posible evaluar correctamente el DIF cuando es innecesario" (p. 435); ya que se trata de un asunto nada fácil, se discutirá más adelante. Finalmente, lo que define el funcionamiento diferencial es el hecho de que los resultados que arroja el instrumento son sistemáticamente diferentes entre los grupos; gran parte de la investigación de los últimos treinta años se ha dedicado a proponer y evaluar técnicas para identificar estas diferencias cumpliendo la exigencia anterior. Sin embargo, el *metro* de Muñiz puede hacer referencia a una prueba como un todo o a los elementos que la componen, los ítems. La discusión original acerca del *sesgo* hacía referencia básicamente a las diferencias observadas en los resultados de las pruebas como un todo más que en lo que hoy se conoce como funcionamiento diferencial de los ítems; sin embargo, los desarrollos técnicos y metodológicos posteriores han puesto especial énfasis en el segundo.

Funcionamiento diferencial de las pruebas y de los ítems

Green (1975), Reynolds (1982b), Shepard (1982), Kok (1988) y Shealy & Stout (1993b) entre otros, coinciden en entender el sesgo en las pruebas como una clase de *invalidez*, de manera que su comprensión conceptual requiere hacer un par de precisiones en torno a la noción de validez: Por una parte, son los errores sistemáticos y no los aleatorios, los que afectan la validez de la medida; el desarrollo de esta idea ayudará a precisar el concepto de funcionamiento diferencial. De otra parte, la validez no sólo hace referencia a aspectos intrínsecos de la prueba sino que involucra elementos del contexto y su relación con los propósitos de la evaluación; el desarrollo de

esta idea permitirá diferenciar FDT de DIF y entender algunos aspectos metodológicos para la detección de cada uno.

Desde la teoría Clásica de los Test (TCT) Ghiselli (1964), Horst (1966), Lord & Novick (1966) y Brown (1980), entre otros, han entendido por error aleatorio toda aquella varianza irrelevante que puede afectar los resultados de la medida en un momento determinado y que pueden provenir del instrumento, del examinado, del contexto o de sus complejas interacciones; este tipo de error produce inconsistencia en las medidas y afecta por igual los resultados de personas de diferentes grupos, en consecuencia, no hace que el instrumento –prueba o ítem- funcionen diferencialmente. Los errores sistemáticos, como su nombre lo indica, afectan sistemáticamente una porción de resultados produciendo una subestimación o sobrestimación de la magnitud del atributo medido para esa porción (Ghiselli, 1964; Gulliksen, 1950). El más claro ejemplo de error sistemático, que además permite introducir a la teoría de la multidimensionalidad iniciada por Kok, (1988) y Shealy & Stout (1989), presentada formalmente por Ackerman (1992) y que ha encontrado sustento empírico en trabajos como el de McCauley & Mendoza (1985), es el hecho de que un instrumento mida, además del atributo para el que fue diseñado, otro irrelevante. Si una prueba diseñada para evaluar un atributo principal[4] mide en algún grado, alguna dimensión diferente, este hecho afectará su validez de constructo, pero si los grupos de interés tienen la misma distribución en esa dimensión irrelevante, la prueba no presentará funcionamiento diferencial. Por el contrario, si los grupos de interés difieren en la distribución de la magnitud de la dimensión irrelevante, el efecto será una subestimación sistemática de la magnitud de atributo principal para el grupo de menor desempeño en la dimensión irrelevante, y entonces el instrumento presentará funcionamiento diferencial contra dicho grupo. Desde este punto de vista para que se configure funcionamiento diferencial debe existir el efecto de un error sistemático –variable o factor irrelevante que interviene en el proceso- y desigual distribución de los grupos de interés en dicha variable.

De otra parte, a partir de la década de los ochenta (Cronbach, 1980) y particularmente con la publicación de las ponencias de Angoff (1988), Cronbach (1988) y Messick (1988) se ha considerado que hablar de validez de una prueba implica recoger toda clase de evidencia empírica y argumentos teóricos que respalden su adecuación, sus usos potenciales, interpretaciones o inferencias en diferentes contextos y con propósitos particulares; desde este punto de vista la validación de una prueba más que un procedimiento sería un proceso continuo de investigación sobre el instrumento y su funcionamiento bajo condiciones particulares y con diferentes objetivos, incluyendo los estudios sobre sesgo de las pruebas (Paz, 1996). Dado que uno de los usos más frecuentes de las pruebas es la predicción del desempeño de los examinados en tareas particulares del

[4] Se emplea aquí la expresión “atributo principal” para hacer referencia a la dimensión o combinación de éstas que constituyen el objeto de la prueba. No se refiere entonces a una prueba necesariamente unidimensional en el sentido de que mida una única dimensión, factor o atributo.

ámbito laboral o educativo, se espera que una prueba no sesgada pueda predecir el desempeño en un criterio externo de manera igualmente precisa para diferentes grupos, y que para individuos que obtengan la misma puntuación en la prueba se prediga el mismo rendimiento en la medida de criterio externo, independientemente del grupo al que pertenezca (Reynold, 1982a, Shepard, 1982, Thorndike, 1995, Anastasi & Urbina, 1998). Así, se habla de FDT cuando los parámetros de las ecuaciones de regresión (pendiente e intercepto) de las puntuaciones de la prueba sobre un criterio predictivo externo, no son iguales para los dos grupos de interés.

En términos de Reynolds (1982a) las situaciones descritas permiten distinguir dos tipos de sesgo: *sesgo* en la validez de constructo y *sesgo* en la validez predictiva, Shepard (1982) habla de sesgo en los ítems de la prueba y sesgo en selección; mientras que Camilli & Shepard (1994), Fidalgo (1996a), Millsap (1997) y Hong (2001) distinguen procedimientos para detectar sesgo interno y sesgo externo. En cualquier caso hoy se acepta el término FDT para referirse a la evidencia empírica de que la prueba tiene diferente funcionamiento para dos grupos tomando un criterio externo a la misma, lo cual implica necesariamente examinar la relación entre los puntajes de la prueba y una medida de dicho criterio externo y comparar la ecuación de regresión obtenida para los dos grupos separadamente. Desde esta perspectiva, el FDT o sesgo externo se manifiesta en diferencias entre los dos grupos en la pendiente o el intercepto de la recta de regresión (Millsap, 1997; Hong, 2001). Este trabajo no se ocupará de la revisión de estos métodos que pueden encontrarse en Reynolds (1982a), Thorndike (1995) o Anastasi & Urbina (1998). Por su parte, DIF hace referencia al funcionamiento diferencial de los elementos de la prueba en términos de la probabilidad de acierto de personas con la misma magnitud de atributo pero pertenecientes a grupos diferentes, y la teoría explicativa aceptada en la actualidad, es la de *la multidimensionalidad* (Ackerman, 1992) que se revisará más adelante. A diferencia de lo que ocurre con el FDT, las técnicas para detectar DIF utilizan criterios internos, generalmente las puntuaciones en la prueba (o subprueba) o una estimación de la magnitud de atributo con base en las respuestas de la misma. La presencia de DIF en algunos ítems de la prueba no necesariamente conduce a la presencia de FDT. En esta relación se han distinguido dos fenómenos; el primero conocido como *amplificación*, ocurre cuando varios ítems con DIF actúan conjuntamente en la misma dirección haciendo que la prueba total favorezca a algún grupo, es decir, presente FDT; pero como lo mostraron Drasgow (1987) y Shealy & Stout (1993b), también puede ocurrir que algunos ítems con DIF contrarresten el efecto de otros que también tengan DIF, haciendo que la prueba no presente FDT, este efecto se conoce como *cancelación*. De otra parte, en un estudio más reciente Hong, (2001) encontró que los ítems con DIF causan diferencias en la pendiente de regresión entre los grupos pero las diferencias en la distribución de atributo entre los grupos (impacto) también se asocia con tales diferencias de pendiente; de esta manera es posible que una prueba compuesta por ítems libres de DIF, presente diferencias en la ecuación de regresión, si hay impacto entre los grupos.

Lo plateado hasta ahora deja abierto un interrogante en torno a la selección y la calidad de la medida de criterio –interno o externo- utilizado para la detección de FDT o de DIF; este constituye

un problema metodológico importante que es común a las diferentes técnicas. Por una parte, si se define el FDT en términos de la relación entre los puntajes en la prueba y un criterio externo, resulta evidente que la validez de los resultados dependerá en gran parte de que la medida del criterio sea insesgada. Este problema, que ha sido reconocido y discutido desde hace décadas por autores como Petersen & Novick (1976) y Shepard (1982) no tiene hoy una solución única; se reconoce la necesidad de contar con alguna evidencia de la validez de la medida de criterio como condición necesaria para su uso y se recurre con frecuencia a su *relevancia* y su relación con el desempeño que se quiere predecir, como argumentos para respaldarla (Shepard, 1982).

Por otra parte, si se define el DIF en términos de comparación de grupos con igual magnitud de atributo medido, es evidente que se requiere de un criterio de igualación de los grupos, del cual pueda afirmarse que arroja medidas no sesgadas del atributo de interés. Las técnicas de detección de DIF utilizan con frecuencia como parámetro de igualación los puntajes en la prueba total o una estimación de la magnitud de atributo a partir de las respuestas en la totalidad de ítems de la prueba, con la gran dificultad de que ese criterio incluye el posible sesgo de los ítems que la componen. La solución más aceptada hoy es el uso de técnicas iterativas, generalmente de dos fases: en una primera fase se equiparan los grupos tomando la totalidad de los ítems de la prueba, y en una segunda fase, se repite el procedimiento excluyendo del parámetro de equiparación aquellos ítems identificados como DIF, en la primera. Este procedimiento general, ha sido utilizado, adaptado y evaluado con diferentes técnicas en trabajos como los de Lord, (1980), Van Der Flier, Mellenbergh, Adèr & Wijn (1984), Holland & Thayer (1988), Park & Lautenschlager (1990), Miller & Oshima (1992), Kim & Cohen (1992), Clauser, Mazor & Hambleton (1993), Lautenschlager, Flaherty & Park (1994), Gómez Benito & Navas-Ara (1996), Hidalgo Montesinos & Gómez Benito (1996, 2003), Gómez Benito & Hidalgo Montesinos (1997a), Fidalgo, Mellenbergh & Muñiz (1999), Navas-Ara & Gómez Benito (2002b), Hidalgo Montesinos & López Pina (2002), entre otros. Algunos de ellos se revisarán más detenidamente en el capítulo dedicado al método correspondiente.

Funcionamiento diferencial de los ítems

Ya resulta reiterativo afirmar que hasta ahora, el DIF se ha entendido como la diferencia en la probabilidad de acierto en un ítem, entre individuos que tienen la misma magnitud del atributo medido por el mismo, pero que pertenecen a grupos diferentes. Partiendo de esta noción fundamental, Ackerman, (1992) propuso la teoría explicativa más aceptada hasta ahora, sobre las razones para que tal comportamiento ocurra, conocida como teoría de la multidimensionalidad. Sin embargo, partiendo de la idea de que el DIF es la contraparte de la invarianza de la medida y de que es posible distinguir dos tipos de invarianza (absoluta y relativa), recientemente Borsboom, Mellenbergh & van Heerden (2002) plantearon que la noción de DIF que hasta ahora se ha aceptado, y en consecuencia la teoría de la multidimensionalidad, parecen ser adecuadas en el contexto de la evaluación de las habilidades, aptitudes y rendimiento, pero necesitan una reformulación que per-

mita explicar la relación entre sesgo y validez de constructo en las medidas de personalidad, intereses o actitudes.

DIF y multidimensionalidad

Desde la perspectiva que se ha expuesto hasta ahora, las técnicas de la TRI[5] y en particular, la observación de la Curva Característica del Ítem (CCI) ha resultado muy útil para la comprensión del DIF. En la figura 1 se representa las CCI de un ítem, esto es, la probabilidad de acertar en el mismo (eje y) en función de la magnitud de atributo medido (eje x) para dos grupos diferentes. Sin mayores elaboraciones teóricas o metodológicas puede observarse que las curvas de este ítem son muy similares para los dos grupos, de manera que no existe una diferencia en la probabilidad de acierto en el ítem entre los dos grupos comparados sino que ésta cambia únicamente en función de la magnitud de atributo medido, se trata de un ítem que no presenta DIF. En la figura se representa además, la distribución de la magnitud del atributo (θ) que busca evaluar el ítem; puede observarse que ambos grupos tienen la misma distribución, luego tampoco hay impacto. Si este ítem mide una habilidad de dos dimensiones, (θ, η) con η, una dimensión irrelevante; de acuerdo con la teoría de la multidimensionalidad de Ackerman (1992) la distribución condicional de η dado θ sería igual para los dos grupos ($f(\eta_1 \mid \theta) = f(\eta_2 \mid \theta)$) y en consecuencia, como las distribuciones de θ también son iguales, el ítem no presentaría DIF. En términos de los parámetros del modelo de la TRI, este ítem tiene igual dificultad y discriminación para los dos grupos de interés ($a_1 = a_2 = 1; b_1 = b_2 = 1$).

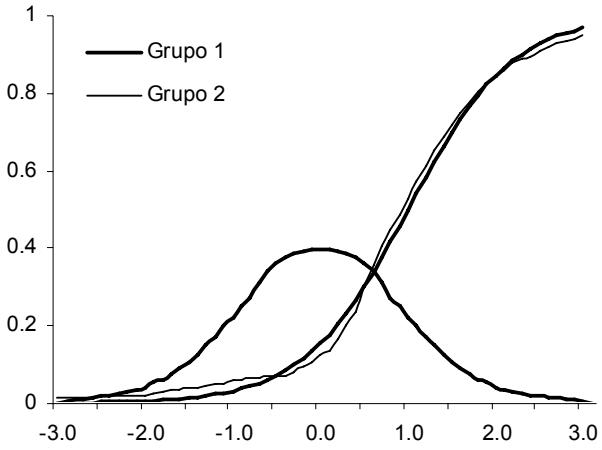


Figura 1: Curva característica y distribución de la magnitud del atributo de un ítem bidimensional sin DIF y sin Impacto.

[5] Posteriormente en el capítulo 3 se presenta una breve exposición sobre los principios de la TRI. En Gruijter & Van der Kamp (1984), Hambleton, Swaminathan & Rogers (1991) o en Muñiz, (1997) se encuentran presentaciones bastante completas y didácticas sobre el tema.

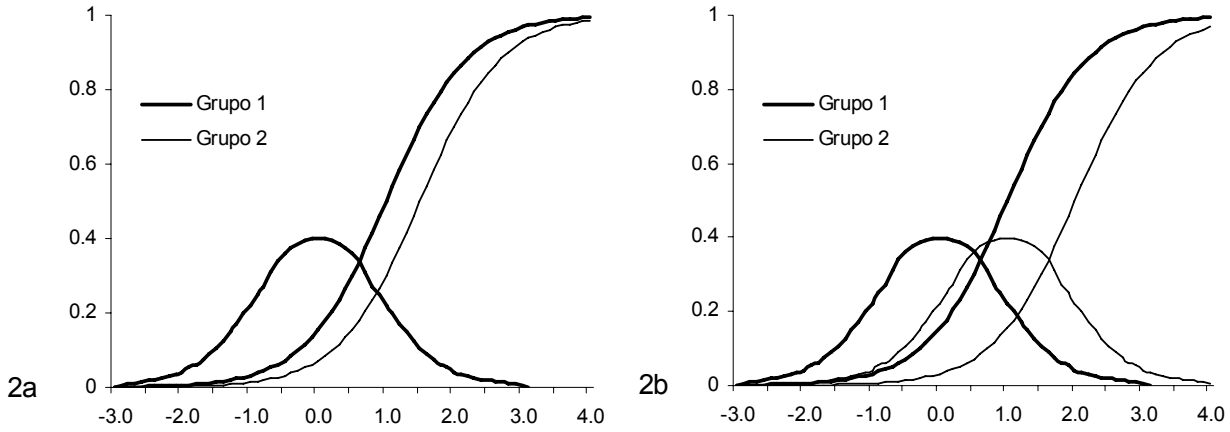


Figura 2: Curva característica de dos ítems bidimensionales con DIF uniforme. 2a: igual distribución de θ y diferente distribución condicional de η, 2b: diferente distribución de θ y correlación entre θ y η.

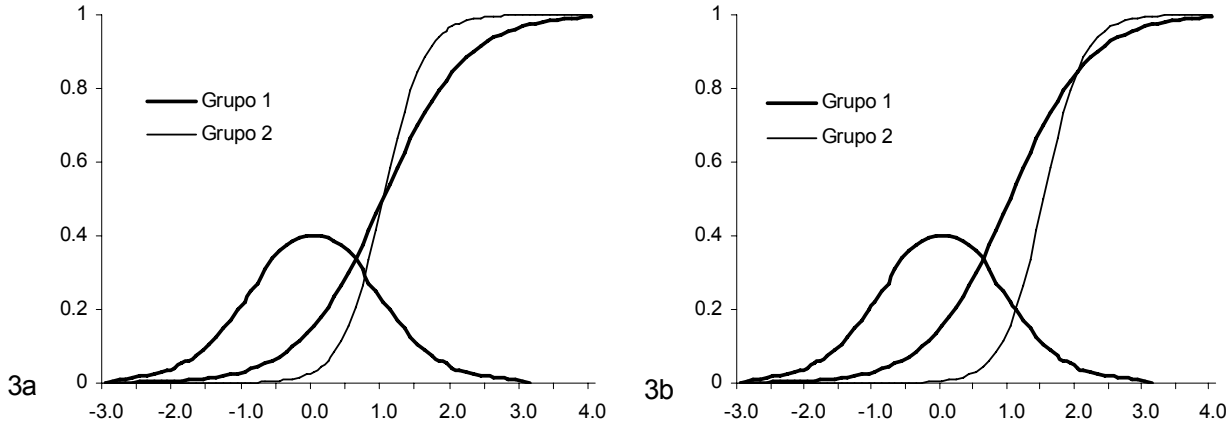


Figura 3: Curva característica de dos ítems bidimensionales con DIF no uniforme. 3a: igual dificultad y diferente discriminación entre los grupos; 3b: diferentes dificultad y discriminación.

Asunto diferente ocurre con los cuatro ítems que se representan en las figuras 2 y 3: En todos los ítems se observan diferencias en la probabilidad de acierto para personas con igual θ pero pertenecientes a grupos diferentes, son ítems con DIF. Sin embargo, el comportamiento es diferente en cada caso. En la figura 2 se representan dos ítems en los que la probabilidad de acierto es mayor para el grupo 1 en todos los niveles de magnitud de atributo pero para el primero (figura 2a) la distribución de θ es la misma, con medias iguales entre los dos grupos ($\mu_1 = \mu_2 = 0$) mientras que en el segundo (figura 2b) éstas son diferentes con media menor para el primero ($\mu_1 = 0; \mu_2 = 1$). Ackerman (1992) mostró que en el primer caso el DIF se presenta porque la distribución condicional de η es diferente entre los grupos ($f(\eta_1 \mid \theta) \neq f(\eta_2 \mid \theta)$) y en el segundo caso, además de impacto, existe una correlación no nula entre las dos dimensiones medidas por el ítem ($\rho_{\theta\eta} \neq 0$). Los dos ítems tienen igual discriminación para los dos grupos pero

dificultad mayor para el grupo 2, los valores de los parámetros son $a_1 = a_2 = 1$, $b_1 = 1$ y $b_2 = 1{,}5$ para el primer ítem, y $a_1 = a_2 = 1$, $b_1 = 1$ y $b_2 = 2$ para el segundo.

Los ítems de la figura 3 difieren de los de la 2 en que presentan una interacción entre la magnitud de atributo y el grupo, lo que hace que las curvas se crucen en algún punto. Mellenbergh (1982a) propuso distinguir el DIF uniforme del no uniforme, dependiendo de si existe tal interacción o no; cuando ésta no existe el ítem presenta DIF uniforme; por el contrario, si hay interacción la diferencia en probabilidad de acertar al ítem cambia para diferentes niveles de atributo y existe DIF no uniforme (Gómez Benito & Hidalgo Montesinos, 1997b; Ferreres, 1998; Prieto, Barbero, y San Luis Costas, 1997). Las CCI de los ítems de la figura 3 se cruzan en algún punto del eje x, haciendo que la probabilidad de acierto sea mayor para un grupo en niveles bajos de magnitud de atributo pero tal diferencia se invierta para niveles altos del mismo; estos dos ítems presentan DIF no uniforme. De acuerdo con Ackerman (1992) si tanto la distribución de θ como las medias de las distribuciones condicionales de η son iguales para los dos grupos, este tipo de DIF se puede presentar cuando: a) las varianzas de las distribuciones de la dimensión irrelevante son diferentes o, b) la correlación entre las dos dimensiones es diferente para los grupos ($\rho_{\theta_1\eta_1} \neq \rho_{\theta_2\eta_2}$).

Pero los dos ítems de la figura 3 también son diferentes entre sí; en el primero (figura 3a) las CCI se cruzan aproximadamente en la mitad de la escala de magnitud de atributo ($\theta = 1$) mientras que en el segundo se cruzan en uno de los extremos ($\theta = 2$); en ambos casos los dos grupos tienen idéntica distribución de θ con media igual a 0, pero el primer ítem tiene igual dificultad y diferente discriminación para los dos grupos ($a_1 = 1$, $a_2 = 2$ y $b_1 = b_2 = 1$), mientras que en el segundo hay diferencias en ambos parámetros ($a_1 = 1$, $b_1 = 1$, $a_2 = 2$ y $b_2 = 1{,}5$). Rogers & Swaminathan (1993), han distinguido el DIF no uniforme propiamente dicho del DIF no uniforme mixto o sencillamente, DIF no uniforme y DIF mixto. El primer caso ocurre cuando el ítem tiene igual dificultad para los dos grupos y el segundo, cuando ambos parámetros son diferentes entre los grupos. Esta distinción tiene importancia metodológica puesto que, como se verá en los estudios empíricos, cuando las curvas se cruzan hacia el centro de la escala de θ los efectos se anulan y puede ocurrir que algunas técnicas no los detecten como ítems con DIF.

En síntesis, si un ítem es unidimensional o la distribución de la dimensión irrelevante es igual para los grupos comparados, no tendrá DIF y los parámetros TRI serán iguales para los dos grupos. En el DIF uniforme no hay interacción entre la magnitud de atributo y el grupo, este ocurre cuando la distribución de la magnitud de atributo es diferente entre los dos grupos y hay correlación entre las dimensiones evaluadas por el ítem, o cuando los grupos difieren en la distribución de la dimensión irrelevante, en cualquier caso el parámetros de dificultad cambia entre los grupos pero el de discriminación es el mismo. Finalmente, en el DIF no uniforme se presenta interacción

entre la magnitud de atributo y el grupo, éste ocurre cuando la diferencia entre los dos grupos está en la varianza de la distribución de la dimensión irrelevante o en la correlación entre las dos dimensiones; puede ocurrir que el parámetro de dificultad sea el mismo para ambos grupos pero el de discriminación no (DIF no uniforme) o que ambos parámetros sean diferentes para los grupos (DIF mixto).

Invarianza de la medida y DIF absoluto y relativo

De acuerdo con Meredith (1993) se entiende que una medida de un atributo es invariante con respecto a una variable cualquiera, si el resultado de la medida es función de la magnitud del atributo y esta relación se mantiene cuando se considera dicha variable. En términos más formales si se considera que un ítem es una medida de un determinado atributo, p es la probabilidad de puntuar en dicho ítem, θ es la magnitud de atributo y X es una variable cualquiera, el ítem es una medida invariante de θ con respecto a X si se cumple que $F(p \mid \theta, X) = F(p \mid \theta)$. Desde este punto de vista el DIF puede entenderse como carencia de invarianza de la medida con respecto a la variable que define los grupos, por ejemplo sexo, raza, etnia o nacionalidad; los ítems con DIF serían medidas no invariantes con respecto al grupo.

Si es verdadero que 3X = 2A y que 2A < 15, entonces es falso afirmar que:	
A. X<15	*B. A<X*
C. 3X>A	*D. No sé*

Soy más divertido que la mayoría de las personas que conozco.	
A. Siempre	*B. Algunas veces*
C. Nunca	

Figura 4[6]: Ejemplos de ítems de una prueba de razonamiento matemático (izquierda) y de personalidad (derecha)

Partiendo de esta noción básica, recientemente Borsboom, Mellenbergh & van Heerden (2002) han distinguido dos tipos de invarianza: absoluta y relativa. Ellos basan su propuesta en las diferencias que pueden observarse en diversos ámbitos de aplicación de las pruebas y por tanto, de los análisis de sesgo. Según los autores puede esperarse que el proceso cognitivo para responder un ítem de una prueba de rendimiento, aptitud o habilidad sea diferente al de un ítem de una prueba de personalidad, intereses o actitudes (ver figura 4). Si se trata de resolver una situación como la planteada en el primer ítem de la figura 4 que está construido para evaluar razonamiento matemático, y si efectivamente lo mide, cabría esperarse que dos personas con igual razonamiento matemático tengan la misma probabilidad de acierto, independientemente del grupo cultural; de esta manera, si se encuentra que personas (orientales y occidentales) con el mismo nivel de razonamiento matemático tienen diferente probabilidad de acertar, se podría concluir que el ítem tiene

[6] Ejemplos utilizados con autorización del Laboratorio de Psicometría del Departamento de Psicología de la Universidad Nacional de Colombia

DIF y buscar las razones de acuerdo con la teoría de la multidimensionalidad. En términos de invarianza de la medida, se concluiría que el ítem es una medida no invariante de razonamiento matemático con respecto al grupo cultural. Dado que la igualación por magnitud de atributo se hace con la misma escala para los dos grupos, se hablaría de invarianza absoluta y este ítem presentaría DIF absoluto. En síntesis, la noción de DIF que se ha expuesto hasta ahora, correspondería a lo que los autores denominan DIF absoluto.

En el segundo caso (pregunta de la derecha en la figura 4) las personas responden la pregunta comparándose con su grupo cultural y entonces puede ocurrir que los occidentales sean más extrovertidos que los orientales, y que un oriental y un occidental igualmente extrovertidos (en una escala de medida común para los dos grupos) tengan diferente probabilidad de puntuar, se afirmaría entonces que hay impacto y DIF absoluto. Sin embargo, los autores sostienen que desde la perspectiva de la evaluación de la personalidad, este DIF absoluto no necesariamente resulta relevante, ya que lo que puede resultar de interés, es que dos examinados con la misma posición relativa dentro de su grupo (por ejemplo, con una puntuación de media desviación estándar por encima de la media de su grupo) tengan la misma probabilidad de puntuar. Si esto no ocurre, el ítem carece de invarianza relativa y se afirmaría que tiene DIF relativo. Se pueden distinguir entonces dos formas de medida: "La *medición absoluta* se refiere al procedimiento de medida de un rasgo en una escala absoluta …, y la *medición relativa* se refiere al procedimiento para medir el rasgo en una escala relativa …, donde las unidades de medida se expresan en términos de la posición relativa dentro del grupo al que pertenece el participante" (Borsboom, Mellenbergh & van Heerden ,2002 p.436).

Esta concepción cambia la mirada que actualmente se tiene sobre la relación entre DIF (o sesgo) y validez de constructo. En el ejemplo expuesto antes, si el ítem de razonamiento matemático presentara DIF (absoluto) parece razonable suponer que puede estar midiendo, además del atributo de interés, una dimensión no relevante y que este hecho afecta su validez de constructo. En el segundo caso sin embargo, el DIF absoluto no representaría sesgo en el ítem ni iría en detrimento de la validez de constructo ya que se trata de una medida relativa y válida del atributo de interés. Desde esta perspectiva los autores sostienen que el DIF absoluto no implica multidimensionalidad, un ítem con invarianza relativa y DIF absoluto no necesariamente es multidimensional, puede ser un indicador válido del constructo que se pretende medir (unidimensional) dentro de un grupo. En consecuencia, la relación entre invarianza y validez de constructo parece más compleja de lo que se ha planteado hasta ahora y la teoría de la multidimensionalidad debe reformularse o completarse para dar cuenta de este fenómeno.

Además de las implicaciones teóricas, aceptar la existencia de medidas absolutas y relativas y en consecuencia, sesgo relativo y absoluto; tiene implicaciones metodológicas que no han sido estudiadas hasta el momento. Según los autores, técnicas como el Mantel Haenszel, la regresión logística y las basadas en la TRI pueden adaptarse para detectar los dos tipos de DIF. En las figuras 5 y 6 se representan dos situaciones diferentes que ilustran los tipos de DIF y las posibles

estrategias de detección desde esta última aproximación. En la figura 5 se ilustra un ítem que presenta impacto, DIF absoluto e invarianza relativa; en el lado izquierdo (5a) se muestran las distribuciones de magnitud de atributo de los dos grupos y las CCI de los mismos tal como se han mostrado hasta el momento, es decir, utilizando una escala de medida de la magnitud de atributo común para los dos grupos (absoluta); puede observarse que tanto las distribuciones de magnitud de atributo como las CCI son ampliamente diferentes, lo que ilustra la presencia de impacto y DIF absoluto, respectivamente. En el lado derecho (5b) se muestran las distribuciones de magnitud de atributo en puntuación *Z* calculada con la media y desviación típica de cada grupo[7] (escala relativa) y las CCI relativas para cada uno de los grupos; obviamente las distribuciones coinciden dada la transformación que se ha realizado y el hecho de que, en este caso, las CCI también coincidan, indica que personas con la misma puntuación relativa dentro de su grupo, tienen la misma probabilidad de puntuar en el ítem, el ítem presenta invarianza relativa o no presenta DIF relativo.

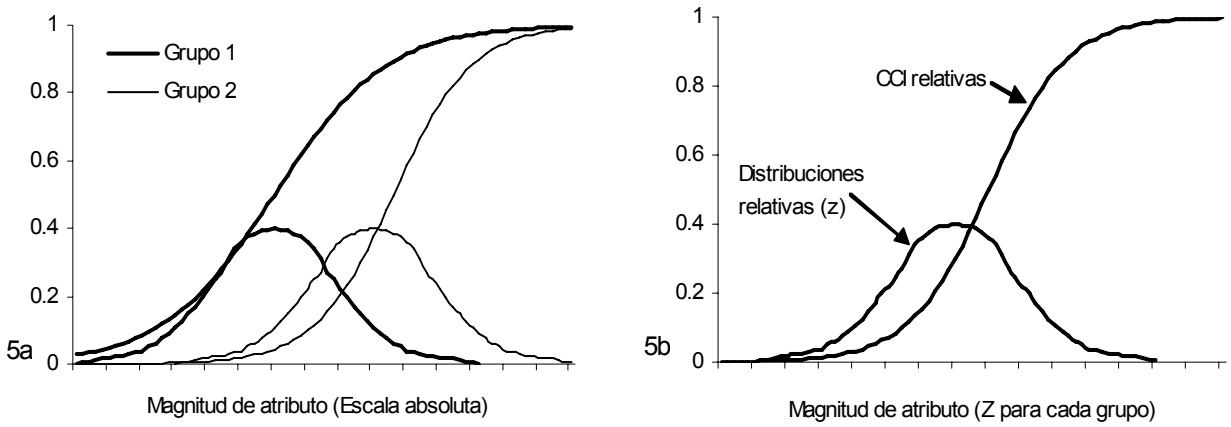


Figura 5: Ilustración de un ítem hipotético que presenta impacto, DIF absoluto e invarianza relativa. 5a: distribuciones de magnitud de atributo y CCI absolutas; 5b: distribuciones expresadas en puntuación z y CCI relativas

En la figura 6 se muestra una situación diferente, este ítem presenta impacto, invarianza absoluta y DIF relativo; puede observarse que las distribuciones de magnitud de atributo son diferentes entre los grupos (impacto) y la CCI absolutas coincide para los mismos (invarianza absoluta); sin embargo, en la figura 6b se observa que cuando se transforma la escala de magnitud de atributo y se construyen las CCI relativas, éstas difieren para los grupos ya que el ítem resulta más fácil y menos discriminativo para el grupo 1 que para el 2; de esta forma, personas con la misma posición relativa dentro de su grupo, tienen mayor probabilidad de puntuar si pertenecen al grupo 1 para casi todos los niveles de magnitud de atributo. Este ítem presenta DIF relativo. De manera similar se pueden ilustrar situaciones en las que el ítem presente impacto, DIF absoluto y DIF relativo; o

[7] Se utiliza esta transformación porque este es un tipo de escala relativa que resulta muy sencilla, fácil de manejar e ilustrativa para los efectos expositivos del tema que se está tratando, sin embargo, puede pensarse en cualquier otro tipo de escala relativa.

no presente ninguno de ellos. Borsboom, Mellenbergh & van Heerden (2002) muestran cómo la invarianza relativa y absoluta no pueden presentarse juntas a menos que las distribuciones de la magnitud de atributo sean iguales para los dos grupos, ya que las CCI absolutas y relativas no pueden coincidir, si las distribuciones de magnitud de atributo difieren. Además, los autores muestran las equivalencias entre los parámetros absolutos y relativos de dificultad y discriminación de la TRI, proponen una modificación al modelo de regresión logística para detectar DIF relativo e ilustran una aplicación de este modelo y de un modelo de ecuaciones estructurales, con ítems de dos pruebas diferentes.

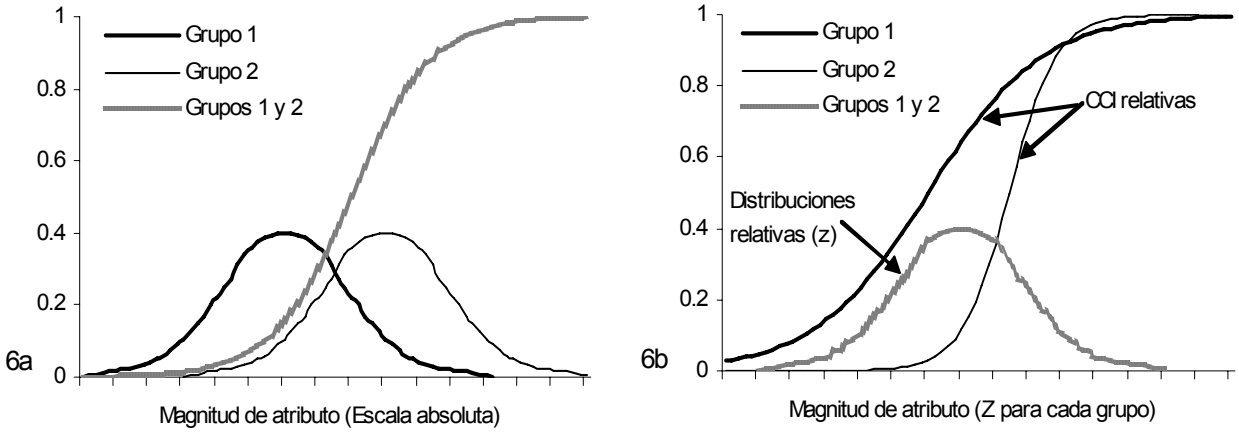


Figura 6: Ilustración de un ítem hipotético que presenta impacto, invarianza absoluta y DIF relativo. 6a: distribuciones de magnitud de atributo y CCI absolutas; 6b: distribuciones expresadas en puntuación *z* y CCI relativas

Según los autores las diferencias entre invarianza y DIF relativo y absoluto no se habían reconocido hasta el momento debido a que la mayoría de estudios aplicados de sesgo y, en consecuencia, de investigación metodológica sobre su detección, ha hecho énfasis en medidas de inteligencia, aptitud, rendimiento o habilidad; frecuentemente utilizadas para la toma de decisiones sobre asignación de cupos en el ámbito laboral y educativo, con todas sus implicaciones sociales éticas y políticas. El reconocimiento de la "*naturaleza relativa de la medida*" ocurre inicialmente en los estudios de evaluación de la personalidad y aún se requiere investigación empírica en otros ámbitos de la evaluación psicológica. Mientras esa investigación avanza y se encuentran argumentos para generalizar, modificar o ignorar esta propuesta, sigue siendo de interés evaluar y mejorar las técnicas existentes en la actualidad. Este trabajo versa sobre técnicas de detección de DIF absoluto, o simplemente DIF, en los términos en que se ha entendido en la exposición anterior y en consecuencia, se seguirán utilizando dichos términos.

Técnicas de detección del DIF

La revisión de la literatura muestra una diversidad de procedimientos para la detección de DIF y de clasificaciones de los mismos. Camilli & Shepard (1994) los clasifican en tres categorías: a) basados en el análisis de varianza y la TCT, b) basados en la TRI y c) los que se basan en el

análisis de tablas de contingencia. Siguiendo a Millsap & Everson (1993), Gómez Benito & Hidalgo Montesinos (1997b) los clasifican en métodos no condicionales y condicionales distinguiendo, dentro de la segunda categoría, los métodos de invarianza condicional observada y no observada. De otra parte, Fidalgo (1996a) presenta dos grandes categorías: los procedimientos que no especifican un modelo de medida y los que se basan en la TRI. En Ferreres (1998) se puede encontrar una revisión de algunas otras clasificaciones que obedecen a criterios diferentes.

Siguiendo la clasificación de Camilli & Shepard (1994) ilustrada en la tabla 1, en este apartado, bajo el rótulo de procedimientos pioneros (Gómez Benito & Hidalgo Montesinos, 1997b), se presenta una breve descripción de las principales técnicas de la primera categoría, las cuales tienen más interés histórico que aplicado, puesto que, dadas las limitaciones que presentan, actualmente no se utilizan. Posteriormente se hace una breve descripción del procedimiento de sesgo simultáneo SIBTEST, que si bien no puede ubicarse en una categoría de esta clasificación, merecen algunos comentarios en una revisión como ésta. Finalmente, se introduce a las características de las otras dos categorías de técnicas para detección de DIF en ítems dicótomos: basadas en análisis de tablas de contingencias (TC) y basadas en la TRI. Estas últimas se tratarán con algo más de detalle en los dos capítulos siguientes.

Tabla 1*

Una clasificación de técnicas para la detección del DIF

Categoría	Técnicas
Técnicas basadas en Teoría Clásica de los Test (TCT) y Análisis de varianza (ANOVA)	Análisis de varianza Índice transformado de dificultad (ITD) ITD ajustado Técnicas basadas en correlaciones
Técnicas basabas en el análisis de tablas de contingencias	Aplicaciones de χ^2 Mantel-Haenszel Método de estandarización Modelos logit y log-lineales Regresión logística
Técnicas basadas en la Teoría de Respuesta al Ítem	Medidas de área Comparación de parámetros: χ^2 de Lord Comparación de modelos
Prueba de sesgo simultáneo SIBTEST	

* Tomada de Herrera, Sánchez Pedraza & Gómez Benito (2001)

Procedimientos pioneros

Las técnicas basadas en la TCT y en ANOVA también puede ubicarse en la primera categoría de Gómez Benito & Hidalgo Montesinos (1997b) ya que no efectúan ajustes en los grupos comparados en relación con la magnitud de atributo medido. Estos métodos, que constituyeron las primeras propuestas metodológicas, presentan la gran limitación de que pueden confundir el DIF con el impacto, precisamente porque no equiparan los grupos por la magnitud de atributo. Así, una dife-

rencia significativa entre los grupos analizados en cuanto a la probabilidad de acertar en un ítem, puede conducir al rechazo de algunos elementos que tienen alta capacidad de discriminación entre personas con alto y bajo nivel de atributo. Esta razón ha hecho que estas técnicas de análisis hayan caído rápidamente en desuso, sólo se mencionarán aquí brevemente las cuatro que aparecen en la tabla 1.

El ***Análisis de varianza*** (ANOVA) fue la primera técnica de elección para detección de DIF hasta comienzos de los años 80. Consiste en un ANOVA de medidas repetidas de dos factores representados por el grupo (factor intrasujeto) y por el ítem, y su interacción; la diferencia entre grupos se observa en el efecto principal de grupo y la diferencia de dificultad, en la interacción de los dos factores (Cleary & Hilton ,1968 en Camilli y Shepard, 1994). Dado que el análisis de varianza depende de la proporción de respuestas correctas, confunde DIF con discriminación del ítem y además, arroja altos valores de error tipo I y tipo II, en la actualidad no se usa como técnica de detección de DIF (Camilli & Shepard, 1987, 1994, Gómez Benito e Hidalgo Montesinos, 1997b)

El ***índice transformado de dificultad*** (ITD) también conocido como *delta plot* fue propuesto por Angoff (1972); Angoff & Ford (1973) partiendo del supuesto de que el DIF se manifiesta a través de una dificultad diferencial entre dos grupos y en consecuencia, identificando los ítems con las mayores diferencias en la proporción de aciertos, se detectarán los ítems con DIF. El método consiste en calcular las proporciones de acierto, p, de cada ítem para cada grupo de interés, estos valores se transforman linealmente para expresarlos en una escala de media 13 y desviación estándar 4 y se representan en un diagrama de dispersión donde cada eje representa la dificultad del ítem para uno de los grupos. Si todos los ítems tienen la misma dificultad para los dos grupos, los puntos caerán sobre una línea recta, en realidad caen en una nube elíptica de la cual puede calcularse un eje principal. De esta manera puede calcularse un valor de ITD que es la distancia perpendicular desde el punto de un ítem hasta dicho eje principal; valores ITD altos indican presencia de DIF. En las publicaciones de Angoff (1982); Angoff, (1993) se pueden encontrar las descripciones detalladas del procedimiento y su presentación formal. A pesar de la sencillez y economía que favorecía el uso generalizado del método, éste dejó de usarse muy pronto debido a la dificultad antes mencionada: no iguala los grupos por magnitud de atributo y en consecuencia confunde impacto y DIF, señalando como ítems con DIF, aquellos con mayor discriminación (Shepard, Camilli & Averill, 1981). El ***ITD ajustado*** busca superar esta limitación, Angoff (1982) propuso emparejar los grupos con base en un criterio externo y corregirlo teniendo en cuenta las correlaciones ítem-prueba; sin embargo, además de no resolver por completo las limitaciones del método, en la práctica no se tiene disponible un criterio externo. Aunque Jensen, (1980) y Shepard, Camilli & Williams (1985) hicieron otras propuestas de modificación del ITD como el ajuste mediante el cálculo de índices ITDI residualizados, en la práctica ninguno de dichos métodos ha resultado más eficiente para detectar DIF que los basados en tablas de contingencia que se presentarán más adelante.

Finalmente, se propusieron varias **técnicas basadas en *correlaciones***. Una de éstas (Jensen, 1974, en Camilli y Shepard, 1994) consiste en ordenar los ítems de acuerdo con el nivel de dificul-

tad, dentro de cada grupo y posteriormente correlacionar los rangos. Valores de correlación cercanos a 1 suponen que los ítems están midiendo un mismo atributo en ambos grupos; en caso contrario debe considerarse DIF. Otro método de correlación busca detectar cómo se comportan los índices de discriminación de los ítems en grupos diferentes, aplicando la correlación biserial-puntual (Green & Draper, 1972 en Gómez Benito & Hidalgo Montesinos, 1997b); Sin embargo, ésta no es una técnica recomendada para la evaluación de DIF ya que la dificultad del ítem afecta este tipo de correlación.

El SIBTEST

La prueba de sesgo simultáneo de ítems, conocida como SIBTEST se basa en la teoría de la multidimensionalidad antes expuesta y, como su nombre lo indica, permite identificar el DIF en uno o varios ítems de manera simultánea, es decir, identifica tanto DIF como FDT. Este procedimiento fue propuesto por Shealy & Stout (1989, 1993a, 1993b) a partir de un modelo TRI multidimensional no paramétrico, suponiendo que la puntuación observada en una prueba puede verse como el compuesto de una puntuación en una subprueba válida y en una que se somete a estudio (subprueba estudiada). Consecuente con la teoría de la multidimensionalidad, la subprueba válida está compuesta por ítems que miden únicamente el atributo de interés, es decir, es esencialmente unidimensional, mientras que la subprueba estudiada tiene ítem sospechosos de DIF y por tanto, no unidimensionales; estos ítems estarían midiendo además del atributo de interés alguna(s) dimensión irrelevante que los autores denominan 'ruido'. Habiendo identificado estas dos subpruebas, las puntuaciones en la primera de ellas se constituyen en un criterio insesgado de igualación de los grupos por magnitud de atributo.

De acuerdo con lo expuesto hasta ahora, si p es la probabilidad de acierto en un ítem estudiado, X es la puntuación observada en la subprueba válida, Y es la puntuación observada en la subprueba estudiada y f y r representan los grupos focal y de referencia, respectivamente, intuitivamente puede entenderse que la ausencia de FDT se dará cuando $F_f(Y \mid X) = F_r(Y \mid X)$ es decir, cuando la distribución condicional de los puntajes en la subprueba estudiada sea igual para los grupos focal y de referencia. De manera similar, la ausencia de DIF en el ítem estudiado ocurre cuando $F_f(p \mid X) = F_r(p \mid X)$ es decir, cuando la distribución de la probabilidad de acierto dada la magnitud de atributo (medida por la puntuación en la subprueba válida) sea igual para los dos grupos. Formalmente Shealy & Stout (1993b) definen el FDT mediante la comparación de los puntajes esperados para muestras aleatorias de los grupos comparados, igualados por magnitud de atributo. Así, si θ es la magnitud de atributo de interés, que toma valores θ_i, y Y_f y Y_r son las puntuaciones en la prueba estudiada para los grupos focal y de referencia, respectivamente, la prueba muestra FDT contra el grupo focal al nivel de magnitud de atributo θ_i, si $E(Y_f \mid \theta = \theta_i) < E(Y_r \mid \theta = \theta_i)$; esto es, cuando el valor esperado del puntaje en la prueba estu-

diada para una muestra aleatoria del grupo focal con magnitud de atributo θ_i es menor que el valor esperado para una muestra aleatoria del grupo de referencia, con la misma magnitud de atributo. De la misma manera, la prueba mostrará FDT contra el grupo de referencia si $E(Y_f \mid \theta = \theta_i) > E(Y_r \mid \theta = \theta_i)$ y no estará sesgada si se cumple que $E(Y_f \mid \theta = \theta_i) = E(Y_r \mid \theta = \theta_i)$. Con base en estos planteamientos, el parámetro β con estimador $\hat{\beta} = \sum_{x=1}^{n} P_{fx}(\overline{Y}_{rx} - \overline{Y}_{fx})$ indica la cantidad y dirección del sesgo. En esta estimación P_{fk} es la proporción de examinados del grupo focal que tienen puntuación x en la prueba válida y $\overline{Y}_{rk}$ y $\overline{Y}_{fk}$ son los promedios en la prueba estudiada para personas del grupo focal y de referencia con puntuación x en la prueba válida. Valores positivos indican que la prueba favorece al grupo de referencia y valores negativos tienen la interpretación contraria. Estas hipótesis ($H_0 : \beta = 0$ contra $H_1^1 : \beta \neq 0$ o $H_1^2 : \beta < 0$ o $H_1^3 : \beta > 0$) puede someterse a prueba mediante el estadístico $B = \frac{\hat{\beta}}{S.E(\hat{\beta})}$, con $S.E(\hat{\beta})$ el error estándar de $\hat{\beta}$, que sigue una distribución normal estandarizada. En Fidalgo, (1996a) se puede encontrar una explicación bastante sencilla e ilustrativa del procedimiento.

Este procedimiento tiene algunas ventajas que hicieron que en sus inicios fuera calificado como muy prometedor (Fidalgo, 1996a): permite detectar DIF o FDT, en consecuencia es especialmente útil para estudiar los fenómenos de amplificación y cancelación de DIF, y tiene propiedades matemáticas bastante sólidas, superiores a las de otras técnicas comúnmente utilizadas (Roussos & Stout ,1996; Jodoin & Gierl, 2001). Por estas razones se han generado una serie de trabajos como el de Nandakumar (1993) que estudia los fenómenos de amplificación y cancelación, los de Narayanan & Swaminathan (1994) y Roussos & Stout (1996) que evalúa su poder y error tipo I en comparación con el Mantel Haenszel, y el de Li & Stout (1996) que proponen una modificación que permite detectar DIF o FDT no uniforme. Aunque los resultados son bastante favorables, este procedimiento no ha tenido la popularidad que habría cabido esperar, probablemente debido a que su costo computacional es mayor que el de otros como el MH que también ha mostrado resultados muy satisfactorios.

Métodos basados en tablas de contingencia y en la TRI

Ya que estas dos categorías de métodos se tratarán con más detalle en los próximos capítulos, en este apartado solamente se presenta una introducción a la 'filosofía' de los mismos intentando una comparación entre ellos. En primer lugar, los métodos basados en el análisis de tablas de contingencia (TC) se fundamentan en el estudio de las relaciones que se dan entre dos variables categóricas: grupo y respuesta al ítem, cuando se considera una tercera variable que puede ser

categórica o numérica: la magnitud de atributo medida a partir del puntaje total en la prueba. Si se comparan solamente dos grupos, como generalmente ocurre en los estudios de DIF, y los ítems de la prueba son dicótomos, situación igualmente frecuente, entonces esta tabla tiene dimensión *2x2* y la estructura que se muestra en la tabla 2. Las estrategias específicas de análisis de dichas relaciones, las hipótesis que se prueban y los estadísticos que se utilizan, cambian según el procedimiento específico, para efectos de esta exposición se tomará como referencia el MH, el más popular de todos y uno de los estudiados en este trabajo.

Desde la perspectiva de la TRI, si la prueba es esencialmente unidimensional y, dada una magnitud de atributo la probabilidad de acertar en un ítem es independiente de la probabilidad de acertar en los demás; entonces la probabilidad condicional de acierto en el mismo ($P(U_i \mid \theta)$, que suele notarse sencillamente como $P_i(\theta)$) es una función monotónica creciente que puede expresarse como un modelo logístico de la forma $P_i(\theta) = \frac{e^{dx}}{1+e^{dx}}$, donde d es una constante que toma los valores 1 o $1{,}7$ dependiendo del tipo de modelo que se desee ajustar, y x es una expresión que involucra los parámetros de los ítems que se consideren en cada caso. Hasta el momento se han considerado tres parámetros: dificultad, discriminación y pseudoazar, notados para el ítem i, como b_i, a_i y c_i, respectivamente. Si se considera que todos son relevantes para el caso particular, entonces se tendrá un modelo logístico de tres parámetros expresado como $P_i(\theta) = c_i + (1-c_i)\frac{e^{da_i(\theta-b_i)}}{1+e^{da_i(\theta-b_i)}}$. Si se supone que el ítem no puede responderse acertadamente por azar o que esta probabilidad es despreciable ($c_i = 0$), entonces se tendrá un modelo de dos parámetros de la forma $P_i(\theta) = \frac{e^{da_i(\theta-b_i)}}{1+e^{da_i(\theta-b_i)}}$ y, finalmente, si se supone que todos los ítems de la prueba tienen la misma discriminación, entonces el modelo sería de un parámetro, definido como $P_i(\theta) = \frac{e^{d(\theta-b_i)}}{1+e^{d(\theta-b_i)}}$. Los modelos de dos y tres parámetros fueron desarrollados por Birnbaum (1958, 1968) y el último fue propuesto por Rasch (1960).

Como puede intuirse a partir de los ejemplos gráficos presentados hasta ahora (figuras 1 a 3, 5 y 6) los procedimientos basados en la TRI para la detección de DIF, se fundamentan en la comparación de los modelos ajustados separadamente para los dos grupos de interés. La idea de fondo es que si para un ítem la probabilidad de acierto puede expresarse mediante el mismo modelo para los dos grupos, éste no presenta DIF (Ver figura 1), mientras que las diferencias observadas entre ellos, constituyen evidencia de DIF. En general, tanto el parámetro de magnitud de atributo (θ) como los de los ítems (b_i, a_i y c_i) se estiman a partir de los vectores de respuestas de los n

individuos en los k ítems de la prueba y se representan mediante la CCI. En los procedimientos para la detección de DIF estas estimaciones se realizan separadamente para los dos grupos y se representan mediante las dos CCI como las que aparecen en las figuras 1 a 3. Los valores de los parámetros b_i, a_i y c_i del modelo pueden observarse en la CCI mediante la localización de la curva sobre la escala de magnitud de atributo (eje x), la inclinación o pendiente de la misma y la asíntota menor o punto de corte en el eje y, respectivamente. Los diferentes procedimientos propuestos dentro de esta categoría para la detección de DIF difieren en la estrategia usada para comparar las CCI de los dos grupos, en general se ha propuesto tres opciones: comparar los modelos como un todo, comparar los valores de los parámetros y medir el área que separa a las dos curvas. Algunos detalles de estos procedimientos se presentarán en el capítulo 3.

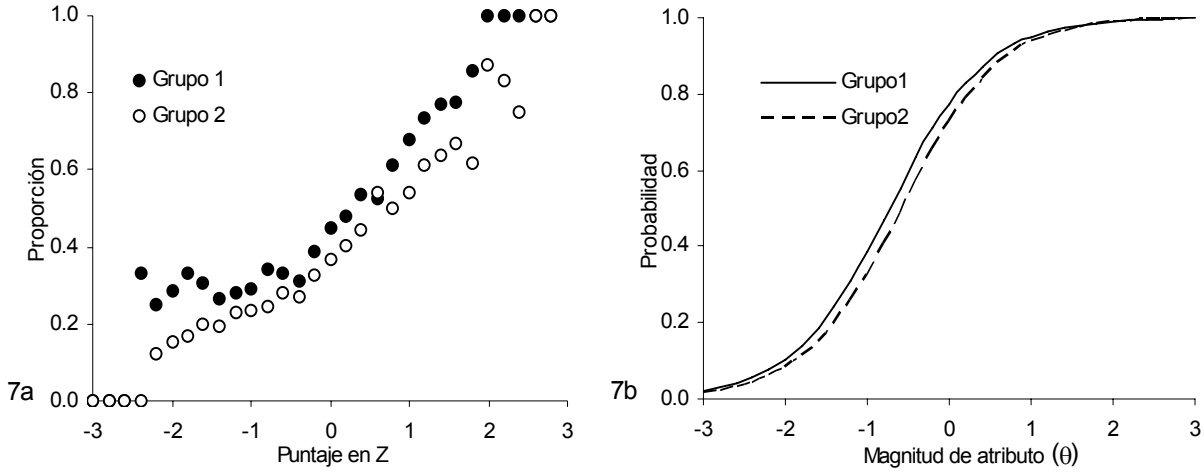


Figura 7: Representaciones del comportamiento de un ítem con DIF. 7a: diagramas de dispersión de las proporciones de acierto en función del puntaje en la prueba. 7b: CCI del ítem para los dos grupos ajustando un modelo logístico de un parámetro

En la figura 7 se muestran dos representaciones gráficas del comportamiento de un ítem con DIF modesto, aplicado a dos grupos de 1140 y 592 examinados. En la figura 7a aparecen los diagramas de dispersión de la proporción de acierto observada (eje y) en función del puntaje obtenido en la prueba total transformado a Z (eje x), para dos grupos. En la figura 7b se muestran las CCI obtenidas cuando se ajustó un modelo logístico de un parámetro y la magnitud de atributo se expresó en una escala de media 0 y desviación típica 1 para ambos grupos. Aunque puede resultar atrevido intentar una comparación de las técnicas a partir de las representaciones de esta figura, lo que resulta claro es que se trata de dos miradas del mismo fenómeno que tienen diferencias y similitudes entre sí. Una primera diferencia fundamental es que las técnicas basadas en TC, específicamente el MH, analiza los datos observados como los que se representan en la figura 7a, sin recurrir a un modelo de medida, utilizando los puntajes en la prueba total como criterio de igualación de los grupos por magnitud de atributo; por su parte, las técnicas basadas en la TRI proponen un modelo de medida y estiman tanto los parámetros de los ítems como el parámetro de magnitud

de atributo de los examinados; de esta manera, el criterio de igualación de los grupos, si bien toma como base las respuestas de los examinados en la prueba, es una estimación de su magnitud de atributo bajo los supuestos del modelo que se utilice.

El cumplimiento de los supuestos necesarios para el uso adecuado de los procedimientos es uno de los aspectos frecuentemente mencionados cuando se trata de decidir la aplicación de una u otra técnica, por parte de los usuarios de las mismas en los estudios aplicados. En general, se ha mencionado este aspecto como una desventaja de las técnicas basadas en la TRI ya que estos procedimientos paramétricos hacen una serie de supuestos sobre los datos, del cumplimiento de los mismos depende el ajuste del modelo y, de éste depende la precisión en la detección de DIF. Sin embargo, según Camilli & Shepard (1994), la idea generalizada de que los métodos basados en TC por su rótulo de "no paramétricos" hacen menos supuestos, no es del todo exacta. Los modelos TRI parten de los supuestos antes mencionados sobre la dimensionalidad de la prueba y la independencia entre la probabilidad de acierto de los ítems; además, el modelo de Rasch supone que la probabilidad de acierto por azar es despreciable y que la discriminación de los ítems es constante, supuestos fuertes que no siempre se cumplen en la práctica. En los métodos basados en TC, por su parte, se hacen implícitamente algunos supuestos que también se pueden considerar fuertes. En primer lugar, al estimar la magnitud de atributo mediante el puntaje observado están suponiendo que los puntajes en todos los ítems tienen el mismo peso; además al no considerar la probabilidad de acierto por azar están suponiendo que ésta es igual para los dos grupos comparados y, finalmente, al no considerar la discriminación de los ítems, están suponiendo que ésta también es igual para los dos grupos. "Así, parece que los métodos basados en TC implícitamente hacen supuestos bastante fuertes" (Camilli & Shepard, 1994 p. 104).

Otros aspectos relacionados con el anterior y de implicaciones importantes son el tamaño de los grupos y la longitud de la prueba. Mediante un estudio de simulación ajustando modelos logísticos de dos parámetros, Cohen, Kane & Kim (2001) mostraron que la precisión en la estimación de los parámetros de dificultad y discriminación de los ítems (b_i, a_i) depende de manera importante del número de examinados, mientras que la precisión de la estimación del parámetro de magnitud de atributo de los individuos, (θ) depende de la longitud de la prueba. En condiciones de grupos pequeños y/o pruebas cortas habrá dudas sobre la precisión de las estimaciones y por tanto, sobre los resultados de la técnica de detección de DIF basada en la TRI, además, bajos estas condiciones resulta más difícil evaluar el ajuste del modelo (Camilli & Shepard, 1994). Las técnicas basadas en TC, específicamente el MH, son menos exigentes en tamaño de grupo (Mazor, Clauser, & Hambleton, 1992; Hambleton, Clauser, Mazor y Jones, 1993) y la longitud de la prueba no ha sido muy estudiada hasta ahora. Este aspecto se retomará más adelante en la presentación de los trabajos empíricos.

Pero los costos en tamaño de muestra y longitud de la prueba, cuando las condiciones prácticas son favorables, se ven ampliamente recompensados por la precisión de los resultados –

cuando el modelo se ajusta a los datos-, el sustento teórico y la elegancia matemática de los modelos TRI. Las técnicas basadas en TC no tienen la sustentación teórica y matemática de la TRI pero, a cambio, resultan bastante sencillos, intuitivos y económicos. Por esta razón el MH se considera aún el método más popular en los análisis de sesgo en contextos aplicados; sin embargo, el resumen de Gómez Benito & Hidalgo Montesinos (1997b) no muestra una diferencia sustancial entre el número de trabajos publicados entre 1980 y 1995 que utilizaron métodos basados en TC y basados en la TRI; además, una revisión no sistemática de las publicaciones más recientes parecen mantener esa tendencia.

Finalmente, con base en una comparación formal entre el MH y el modelo de Rasch en la detección de DIF, Holland & Thayer (1988) concluyeron que "la percepción de que las aproximaciones basadas en la TRI son 'teoricamente preferidas' sobre los procedimientos basados en el χ^2 no es la forma más precisa de describir la situación." (Holland & Thayer, 1988, p. 142). De acuerdo con su análisis, estas dos aproximaciones están íntimamente relacionadas cuando los datos satisfacen los supuestos del modelo de Rasch, los grupos se equiparan según el número de aciertos en la prueba total incluyendo el ítem que se está estudiando, los ítem incluidos en el criterio de equiparación –con excepción del que se está analizando- no presentan DIF y los grupos que se están comparando se puede considerar muestras aleatorias de las poblaciones de referencia y focal. Dorans & Holland (1993) advierten sin embargo, que sólo bajo estos supuestos, bastante fuertes, puede sustentarse una estrecha relación entre los dos métodos; además, subrayan el hecho de que se están haciendo afirmaciones sobre la relación entre las categorías de procedimientos χ^2 y los basados en la TRI, a partir del análisis del MH y el modelo de Rasch que son solamente una de las técnicas de cada una de las categorías.

Capítulo 2:
PROCEDIMIENTOS BASADOS EN EL ANALISIS DE TABLAS DE CONTINGENCIA

Dentro de los métodos basados en el análisis de tablas de contingencia se pueden distinguir dos enfoques: los que se fundamentan en la prueba de hipótesis sobre la igualdad de proporciones analizando tablas de contingencia bidimensionales, y los que generan modelos para el análisis de variables categóricas trabajando con tablas de contingencia de más de dos dimensiones. Dentro de los primeros se encuentran algunas aplicaciones de la prueba χ^2, el método de estandarización y el Mantel-Haenszel; mientras que dentro de los últimos se pueden incluir los modelos logit y loglineales y la regresión logística (RL). En este capítulo se revisan estas técnicas haciendo especial énfasis en el Mantel-Haenszel (MH) y en la RL, técnicas utilizadas en los estudios empíricos que se exponen en la segunda parte del documento.

Comparación de proporciones

Los métodos que comparan proporciones se fundamentan en que el DIF se detecta probando una hipótesis sobre la igualdad de proporciones entre los grupos igualados por la magnitud de atributo medido. Esta última, tomando como medida el puntaje total en la prueba, se fracciona de manera que genere diferentes estratos ($K = 1,..k,...m$) y para cada estrato se construye una tabla de contingencias que cruza las variables grupo ($J = 1,..j,...g$) y respuesta en el ítem ($I = 1,...i...p$). Así, cuando se trata de ítems dicótomos ($I = 0,1$) y se comparan solamente dos grupos ($J = r, f$), la información completa sobre los aciertos y fallos en cada ítem tienen en tantas tablas bidimensionales de *2x2*, como estratos de la magnitud de atributo medido; cada una de ellas tiene la estructura que se muestra en la tabla 2,

Tabla 2

Estructura de una tabla de contingencia correspondiente a un nivel k de magnitud de atributo

Puntaje en el ítem	Grupo		Total
	Referencia	Focal	
1	f_{1rk}	f_{1fk}	f_{1k}
0	f_{0rk}	f_{0fk}	f_{0k}
Total	f_{rk}	f_{fk}	f_k

donde: f_{1rk} es el número de examinados de magnitud de atributo k y del grupo de referencia que acertaron en el ítem

f_{1fk} es el número de examinados de magnitud de atributo k y del grupo focal que acer-

taron en el ítem

f_{0rk} es el número de examinados de magnitud de atributo k y del grupo de referencia que fallaron en el ítem

f_{0fk} es el número de examinados de magnitud de atributo k y del grupo focal que fallaron en el ítem

f_{rk} y f_{fk} son el número total de examinados de los grupos de referencia y focal, respectivamente, con magnitud de atributo k.

f_{1k} y f_{0k} son el número total de examinados con magnitud de atributo k que acertaron y fallaron en el ítem, respectivamente

f_k es el número total de examinados de magnitud de atributo k.

Aplicaciones de χ^2

Un procedimiento propuesto por Scheuneman (1979) y conocido posteriormente como el **Ji cuadrado de los aciertos**, parte del supuesto de que si, dentro de cada estrato, la proporción de aciertos es igual en cada uno de los grupos, se puede descartar la presencia de DIF en el ítem. En este sentido propuso probar la hipótesis de igualdad de proporciones de los aciertos, mediante un estadístico χ^2 con $(m-1)(g-1)$ grados de libertad, siendo m el número de estratos y g el número de grupos. El estadístico, como la prueba Ji cuadrado tradicional, se basaba en la comparación de las frecuencias observadas y las esperadas en cada una de las celdas de tablas de contingencias como la de la tabla 2 pero considerando solamente las proporciones de acierto en cada ítem. Así, si f_{1jk} y e_{1jk} son la frecuencia observada y esperada, respectivamente, de personas del grupo j y del estrato k que aciertan al ítem, el estadístico se define como

$$\chi^2_{aciertos} = \sum_{j=1}^{g}\sum_{k=1}^{m}\frac{(e_{1jk} - f_{1jk})^2}{e_{1jk}}$$

Aunque la sencillez y economía del procedimiento habrían podido augurar un uso generalizado, Baker (1981) y el mismo Scheuneman (1981) discutieron cómo este procedimiento, al considerar solamente los aciertos, puede inestabilizarse cuando existe impacto, caso en el cual los resultados pueden depender de qué tan similares sean los tamaños de los dos grupos; en estas condiciones, no resulta claro que la distribución que siga sea realmente la χ^2. De acuerdo con Ironson (1982), Camilli propuso una modificación al procedimiento considerando tanto las proporciones de acierto como las de los fallos, esta suma seguiría una distribución χ^2 con $m(g-1)$ grados de libertad.

De esta manera, siguiendo la misma notación utilizada hasta ahora, **el** χ^2 **completo** se define como: $\chi^2_{completo} = \sum_{i=0}^{I}\sum_{j=1}^{g}\sum_{k=1}^{m}\frac{(e_{ijk} - f_{ijk})^2}{e_{ijk}}$

Tanto el procedimiento inicial como su modificación fueron criticados por Mellenbergh, (1982a) por su incapacidad para detectar DIF no uniforme; además, ambos procedimientos se inestabilizan con valores bajos dentro de las celdas, con diferencias en los tamaños de muestra o ante la presencia de impacto en los grupos. En la actualidad estos procedimientos no se utilizan en la práctica y no son recomendados por los autores sobre el tema.

Método de estandarización

La primera referencia que generalmente se reporta sobre el método de estandarización, también conocido como Diferencia de proporciones estandarizadas (DPE), son las publicaciones de Dorans & Kulick (1986) y Dorans (1989); sin embargo, según Dorans & Holland (1993) los primeros desarrollos del procedimiento se iniciaron tres años antes mediante trabajos de aplicación en diferentes pruebas del Educational Testing Service (ETS) cuyos resultados se encuentran en Kulick & Dorans (1983a); Kulick & Dorans (1983b) y Dorans & Kulick (1983), reportes de investigación no publicados. El procedimiento se basa en el supuesto de que un ítem presenta DIF cuando su desempeño difiere para individuos de grupos diferentes con el mismo nivel de magnitud de atributo. Inicialmente ese desempeño se estimó a partir de las proporciones condicionales de aciertos en el ítem para los grupos, lo que le dio al procedimiento la denominación DPE; sin embargo, Dorans y Holland (1993) lo estiman por medio de regresiones no paramétricas ítem test para los dos grupos separadamente. Si $e_f(I \mid K)$ y $e_r(I \mid K)$ son las regresiones empíricas ítem-prueba para los grupos focal y referencia respectivamente, donde I es puntaje en el ítem y K es el nivel de magnitud de atributo, y, e_{fk} y e_{rk} son estimaciones de dichas regresiones para el estrato k para los mismos grupos, se puede definir el DIF en el estrato k como $D_k = e_{fk} - e_{rk}$ Es decir, la diferencia en el desempeño en un ítem entre individuos del grupo focal y de referencia, habiendo ajustado por la magnitud del atributo medido.

Tales valores son sometidos inicialmente a una inspección visual, para lo cual se realizan gráficos como los que aparecen en las figuras 8 y 9; éstas muestran el comportamiento de un ítem sin DIF y uno con DIF, respectivamente, aplicados a 592 personas del grupo 1 y 1140 del grupo 2, con idéntica distribución en los puntajes totales en la prueba. En las figuras 8a y 9a aparecen los diagramas de dispersión de las proporciones de acierto en el ítem en función del rango (de longitud 0,2) de puntaje en la prueba total, expresado en puntuación Z; en las figuras 8b y 9b se muestran las diferencias entre los dos grupos expresadas en porcentaje, en función de los mismos rangos de puntaje total en la prueba. La comparación de las dos figuras permiten afirmar que en el

primer caso (figura 8) el comportamiento del ítem es bastante similar en los dos grupos y las diferencias, que se encuentran en un pequeño rango alrededor del 0, pueden considerarse despreciables; es un ítem sin DIF. Por el contrario, tanto el diagrama de dispersión como las diferencias entre grupos en el ítem de la figura 9 muestran discrepancias importantes y patrones que permiten identificar el tipo de DIF: las curvas para los dos grupos se cruzan alrededor $z = 0.5$ (ver figura 9a) y además, las diferencias son negativas para bajos puntajes en la prueba y positivas para puntajes altos (ver figura 9b); se trata de un ítem que muestra DIF no uniforme.

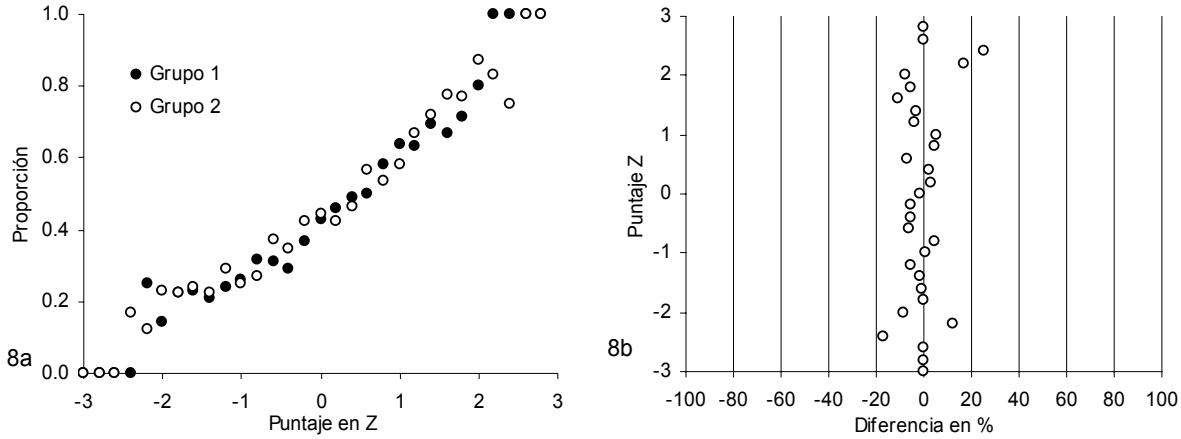


Figura 8: Representaciones gráficas del comportamiento de un ítem sin DIF: 8a: Diagrama de dispersión de la proporción de acierto en función del puntaje en la prueba; 8b: magnitud de las diferencias entre los grupos en función del puntaje en la prueba

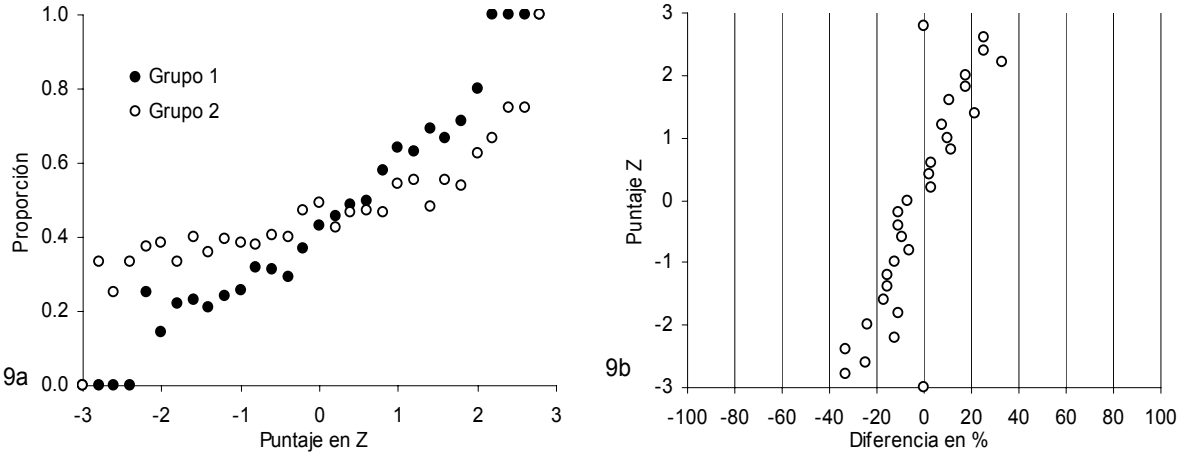


Figura 9: Representaciones gráficas del comportamiento de un ítem con DIF: 9a: Diagrama de dispersión de la proporción de acierto en función del puntaje en la prueba; 9b: magnitud de las diferencias entre los grupos en función del puntaje en la prueba

Aunque la anterior herramienta visual es muy importante en el procedimiento para la descripción del DIF, debe complementarse con el cálculo de un índice numérico que permita identificar los ítems sospechosos de DIF con algún criterio. Siguiendo la notación utilizada hasta ahora, este

índice se define como: $DPE = \frac{\sum_{k=1}^{m} w_k \left(e_{fk} - e_{rk} \right)}{\sum_{k=1}^{m} w_k} = \frac{\sum_{k=1}^{m} w_k D_k}{\sum_{k=1}^{m} w_k}$ o como $DPE = \frac{\sum_{k=1}^{m} w_k \left(p_{fk} - p_{rk} \right)}{\sum_{k=1}^{m} w_k}$,

según se trabaje con las estimaciones de regresión o con las proporciones condicionales de acierto. $p_{fk} = \frac{f_{1fk}}{f_{fk}}$ y $p_{rk} = \frac{f_{1rk}}{f_{rk}}$ son las proporciones de acierto en el ítem para el estrato k en los grupos focal y de referencia, respectivamente, y w_k es el factor de ponderación de la diferencia en dicho estrato. Este factor de ponderación es común para los dos grupos, lo cual constituye uno de los elementos esenciales del procedimiento ya que diferencia el cálculo del DIF del de impacto, definido como la diferencia de proporciones ponderadas cada una por la frecuencia relativa del grupo correspondiente. La elección del factor de ponderación depende de los objetivos de la investigación. Algunos valores que puede tomar w_k son: a) f_k, el número total de examinados en el estrato k, b), f_{rk}, el número de examinados del grupo de referencia en el estrato k, c) f_{fk}, el número de examinados del grupo focal en el estrato k, o d) la frecuencia relativa en el estrato k en algún grupo de referencia. Uno de los que se usa con mayor frecuencia es el número de examinados del grupo focal en el estrato correspondiente (f_{fk})

Los valores que puede tomar el DPE van entre –1 y +1, si los valores son positivos el ítem está favoreciendo al grupo focal. Además, Dorans & Holland (1993) sugieren rangos de interpretación: valores entre –0.05 y 0.05 se consideran despreciables, valores entre 0.05 y 0.1 en valor absoluto son dudosos y valores por fuera de estos últimos rangos obligan a una revisión cuidadosa de los ítems en busca de DIF. Estos valores, sin embargo, se transforman a una métrica delta que arroja correlaciones mayores con el estadístico Mantel-Haenszel. Tomando f_{fk} como el factor de ponderación, la transformación delta es $DPE \quad D - DIF = -2.35 \ln \left(\frac{P_f / 1 - P_f}{p_f / 1 - p_f} \right)$; donde $P_f = \frac{\sum_{k=1}^{m} f_{fk} p_{rk}}{\sum_{k=1}^{m} f_{fk}}$ es la proporción del grupo de referencia utilizando como factor de ponderación el número de examinados del grupo focal en el estrato respectivo, y $p_f = \frac{\sum_{k=1}^{m} f_{fk} p_{fk}}{\sum_{k=1}^{m} f_{fk}}$.

Aunque estos índices numéricos (el DPE y el DPE D-DIF) se venían utilizando en el ETS como uno de los procedimientos para la descripción del DIF, algunos años más tarde se desarrollaron las formas de calcular sus respectivos errores estándar (Dorans & Holland, 1993), lo que permite someter a prueba la hipótesis de la existencia de DIF con algún nivel de confianza, siguiendo una distribución normal estándar. Una explicación detallada de este procedimiento se encuentra en Fidalgo (1996a)

Mantel Haenszel

El estadístico Mantel-Haenszel (MH), propuesto originalmente por Mantel & Haenszel (1959) para el análisis de grupos pareados en el ámbito de los estudios epidemiológicos y utilizado posteriormente para detección de DIF por Holland & Thayer (1986, 1988), es hoy uno de los procedimientos más populares en el ámbito de los estudios de sesgo en ítems dicótomos. Su éxito se debe, además de su sencillez y economía computacional, a que maneja de manera eficiente los distintos niveles de habilidad como una variable de control, permitiendo evaluar y describir cómo la relación entre las variables grupo y respuesta en el ítem, se modifica por la presencia de otra variable categórica con m estratos (magnitud del atributo medido). El MH parte del supuesto de que si un ítem no tiene DIF, la razón entre el número de personas que aciertan y fallan el mismo debe ser igual en los grupos de comparación para los m estratos. Si, siguiendo la notación utilizada hasta ahora, f_{1rk} y f_{0rk} son el número de examinados del grupo de referencia y del estrato k que acertaron y fallaron en el ítem, respectivamente; y f_{1fk} y f_{0fk} son el número de examinados del grupo focal y del estrato k que acertaron y fallaron en el ítem, respectivamente; la supuesta igualdad de razones puede expresarse como $\frac{f_{1rk}}{f_{0rk}} = \alpha \frac{f_{1fk}}{f_{0fk}}$, lo que permite afirmar que $\frac{f_{1rk} f_{0fk}}{f_{0rk} f_{1fk}} = \alpha = 1$; razón conocida como *Odds ratio* para el estrato k.

El MH prueba la hipótesis de que en m estratos de magnitud del atributo medido, la razón (llamada *Odds ratio común* y notada generalmente como α_{MH}), es igual a 1, ($H_0 : \alpha_{MH} = 1$). Un estimador de este *Odds ratio* común es $\hat{\alpha}_{MH} = \frac{\sum_{k=1}^{m} f_{1rk} f_{0fk} / f_k}{\sum_{k=1}^{m} f_{0rk} f_{1fk} / f_k}$ y el estadístico de prueba para $H_0 : \alpha_{MH} = 1$ contra $H_1 : \alpha_{MH} \neq 1$, conocido como el Ji cuadrado de Mantel-Haenszel sigue una distribución χ^2 con un grado de libertad, y está dado por $\chi^2_{MH} = \frac{\left(\left| \sum_{k=1}^{m} f_{1rk} - \sum_{k=1}^{m} E(f_{1rk}) \right| - 0{,}5 \right)^2}{\sum_{k=1}^{m} Var(f_{1rk})}$;

donde: $E(f_{1rk}) = \frac{f_{rk} f_{1k}}{f_k}$ es el número de personas del grupo de referencia y del estrato k, que deben acertar en el ítem si éste no tiene DIF; es decir, la frecuencia esperada en la celda $1rk$ calculada, como en las aplicaciones tradicionales de Ji cuadrado, como el producto de las respectivas marginales, y

$Var(f_{r1k}) = \frac{f_{rk} f_{1k} f_{fk} f_{0k}}{f_k^2 (f_k - 1)}$ es la varianza de la celda $1rk$, también calculada a partir de las marginales de cada tabla de contingencias, asumiendo una distribución hipergeométrica.

Si se rechaza H_0, el ítem estudiado tiene DIF y el valor del estimador $\hat{\alpha}_{MH}$, que puede variar en el rango $(0,\infty)$, da información acerca de la magnitud y dirección del mismo: $\hat{\alpha}_{MH} > 1$ sugiere sesgo a favor del grupo de referencia, mientras que $\hat{\alpha}_{MH} < 1$ sugiere sesgo en contra del mismo grupo. Estos valores, sin embargo, pueden expresarse en métricas diferentes que pueden facilitar la interpretación o comparación de los resultados. Una de estas métricas propuesta por Holland & Thayer (1985) conocida como escala delta del ETS, análoga a la empleada en el procedimiento de estandarización, es $MH \quad D - DIF = -2.35 ln(\hat{\alpha}_{MH})$. Esta transformación expresa los valores de $\hat{\alpha}_{MH}$ en una escala simétrica donde 0 es el valor de hipótesis nula, es decir, indica ausencia de DIF, y los valores negativos o positivos informan la dirección y magnitud del DIF en términos de diferencia de la dificultad del ítem. Valores negativos indican que el ítem resulta más fácil para el grupo de referencia (DIF en contra del grupo focal) y valores positivos tienen la interpretación inversa.

Dorans & Holland (1993) muestran otras dos transformaciones que también permiten interpretaciones en términos de la dificultad del ítem. La primera de ellas consiste en una conversión de las proporciones de acierto a una escala *z* usando la inversa de la función normal acumulada, para posteriormente expresarlos en una nueva escala de media 12 y desviación estándar 4, mediante una transformación lineal. La segunda, conocida como métrica p, es $MH \quad P - DIF = p_f - P_r$, donde $P_r = \frac{\hat{\alpha}_{MH} p_f}{(1 - p_f) + \hat{\alpha}_{MH} p_f}$ puede verse como la proporción de aciertos predicha para el grupo de referencia con base en el *Odds ratio* de Mantel-Haenszel, y p_f es la proporción de aciertos en el grupo focal.

De otra parte, Robins, Breslow & Greenland (1986) y Phillips & Holland (1987) propusieron expresiones equivalentes para estimar el error estándar del log de Mantel-Haenszel, las cuales per-

miten obtener un estimador del error estándar para el $MH\ D\text{-}DIF$. Por su parte, Holland (1989) desarrolló una expresión para estimar el error estándar del $MH\ P\text{-}DIF$. Estos, sin embargo no han tenido uso tan generalizado como el estadístico χ^2_{MH} para identificar DIF y la estimación de α_{MH} para describirlo. La explicación de dichas expresiones pueden encontrarse en Dorans & Holland (1993); además, en Holland & Thayer (1988), Donoghue, Holland & Thayer (1993), Camilli (1993) y Fidalgo (1996a), entre otros, se encuentran explicaciones detalladas e ilustraciones del procedimiento completo.

Sin lugar a dudas este es uno de los procedimientos más utilizados en los estudios actuales sobre sesgo en ítems de calificación dicótoma. Entre las razones que justifican tal hecho se encuentran, además de su eficiencia en la detección de DIF, su sencillez, su bajo costo computacional y sus bajos requerimientos de tamaño de muestras o supuestos distribucionales. Sin embargo, se han identificado algunas limitaciones una de las cuales, que es común a todos los métodos que utilizan los puntajes observados en la prueba como criterio de pareamiento de los grupos, es la posible contaminación de los ítems con DIF en dicho criterio. Como respuesta a esta dificultad, Holland & Thayer (1988) propusieron un procedimiento de dos etapas que permite refinar el criterio de pareamiento. En la primera etapa se aplica el procedimiento tal cual se ha descrito, utilizando como criterio el puntaje en la totalidad de los ítems de la prueba y se identifican los ítems con DIF. En la segunda etapa se recalifica la prueba excluyendo los ítems identificados en la etapa anterior, exceptuando el que se está analizando, y se repite el procedimiento para todos los ítems; así, en el criterio de pareamiento se ha excluido la posible contaminación de los ítems con DIF, exceptuando el ítem que se está estudiando ya que "… los análisis sugieren que la inclusión del ítem estudiado en el criterio de pareamiento no oculta la existencia de DIF, más bien es la inclusión de otros ítems con DIF en el criterio, lo que puede conducir a que se identifique como no DIF el ítem estudiado cuando de hecho lo es." (Holland & Thayer, 1988 p 141). De otra parte, en su tesis doctoral no publicada Fidalgo (1996b) propuso un procedimiento iterativo que también se inicia con la aplicación del procedimiento para identificar los ítems con DIF y continúa iterativamente purificando el criterio de pareamiento mediante la eliminación de los ítems identificados con DIF; el proceso se detiene cuando en dos iteraciones consecutivas se identifiquen los mismos ítems con DIF. Diferentes estudios como los de Miller & Oshima (1992), Clauser, Mazor & Hambleton (1993), Navas-Ara & Gómez Benito (1994, 2002b), Fidalgo & Mellenbergh (1995) y Fidalgo (1996b) han mostrado que un procedimiento de purificación de la medida de la magnitud de atributo mejora eficientemente la potencia del estadístico MH y en consecuencia su uso se ha generalizado hoy.

Otra de las limitaciones que se ha reportado del MH es su incapacidad para detectar DIF no uniforme; los hallazgos de Narayanan & Swaminathan (1996); Rogers & Swaminathan (1993); Swaminathan & Rogers (1990), son bastante consistente al respecto. Esto ocurre especialmente cuando los ítems tienen dificultad media; sin embargo, su poder mejora cuando se trata de ítems

de alta o baja dificultad. Puesto que el DIF no uniforme se presenta cuando los ítems tienen igual dificultad y diferente discriminación para los dos grupos, si la dificultad es media las CCI se cruzan hacia el centro de la escala de magnitud de atributo y el MH, diseñado para detectar un efecto de grupo, no identifica este tipo de interacción; por el contrario, si el ítem tiene dificultad baja o alta con similar discriminación para los dos grupos, las CCI se cruzan hacia los extremos bajo o alto de la escala de magnitud de atributo haciendo que MH detecte un efecto de grupo (Rogers & Swaminathan, 1993; Uttaro & Millsap, 1994). Con el propósito de superar esta limitación Mazor, Clauser & Hambleton (1994) propusieron una modificación del MH que consiste en correr separadamente el procedimiento para los grupos con mayor y menor desempeño en la prueba y comparar los resultados; Fidalgo & Mellenbergh (1995) han reportado resultados satisfactorios utilizando esta modificación pero los resultados hallados por Hidalgo Montesinos & López Pina (2004) mostraron que al utilizar esta modificación no hay ganancia en la tasa de detección de DIF uniforme y el aumento en las detecciones correctas de DIF no uniforme no es muy importante.

Dada la rápida popularización del MH, las dos últimas décadas han sido testigos de la producción de un importante volumen de trabajos que buscaban analizar el efecto de diversos factores sobre su poder, su distribución y su error tipo I. Algunos de los factores estudiados han sido el tamaño de muestra, el número de estratos o rangos que se emplean para separar los estratos de magnitud de atributo, el tipo de ítem estudiado en términos de los valores de sus parámetros, las diferencias en la distribución de magnitud de atributo entre los grupos focal y de referencia, el porcentaje de ítems con DIF dentro de la prueba, y el tipo y magnitud de DIF del ítem estudiado. Hambleton, Clauser, Mazor & Jones (1993) presentan un completo resumen de los principales trabajos realizados entre los ochenta y primeros años de los noventa, y sus más relevantes implicaciones prácticas; más adelante, en el capítulo 4, se presentan y discuten algunos de éstos y otros más recientes.

Otro tipo de estudios han comparado el MH con otros procedimientos para evaluar su eficiencia relativa. En dos estudios diferentes Narayanan & Swaminathan (1994) y Roussos & Stout (1996) compararon el comportamiento del error tipo I del MH y SIBTEST de Shealy & Stout (1993a) y no observaron grandes diferencias entre ellos. Desde una perspectiva similar Narayanan & Swaminathan (1996) compararon el error tipo I y el poder del MH, la regresión logística (RL) y el SIBTEST y encontraron tasas de detección del DIF uniforme muy similar para los tres procedimientos pero muy inferiores para el MH cuando se trataba de detectar DIF no uniforme. De otra parte Rogers & Swaminathan (1993) evaluaron las propiedades distribucionales y el poder del MH y la RL y encontraron diferencias entre los dos procedimientos para detectar DIF uniforme y no uniforme, pero poder similar para detectar DIF mixto. En el trabajo de Miller & Oshima (1992) el MH mostró comportamiento más estable a través del tamaño de muestra, comparado con seis índices de DIF basados en la TRI. Finalmente, Gómez Benito & Navas-Ara (2000) compararon la potencia y el error tipo I del MH con el de otros seis procedimientos para detección de DIF, inclu-

yendo tres basados en la TRI, y encontraron que el MH mostró los mejores resultados en términos de tasas de detecciones correctas y falsos positivos.

Modelos para el análisis de tablas

Los métodos que se presentan a continuación se fundamentan en el ajuste de modelos que expliquen las relaciones existentes entre las variables consideradas en las tablas de contingencia descritas antes (magnitud de atributo medido tomando como medida el puntaje total en la prueba, $K = 1,...k,...m$, grupo $J = r, f$; y respuesta en el ítem $I = 1,0$), o la estimación de los parámetros de un modelo que explique la probabilidad de acierto en el ítem en función del grupo de pertenencia, de la magnitud de atributo medido y de las interacciones. Comparar el ajuste de los diferentes modelos o rechazar la hipótesis para cada uno de los parámetros, brinda información no solamente sobre la presencia sino sobre el tipo de DIF.

Modelos log-lineales y modelos logit

Los modelos loglineales, modelos lineales sobre el logaritmo natural de las frecuencias esperadas, permiten analizar la interacción entre todas las variables de una tabla de contingencia, simultáneamente, incluyendo las interacciones entre grupos de variables; es decir, los modelos loglineales analizan una tabla de contingencia multidimensional, en lugar de m tablas bidimensionales. Dicho brevemente, se trata de ajustar un modelo que busca predecir la frecuencia esperada en cada celda de la tabla como un producto de los efectos (efectos principales y sus interacciones); posteriormente se efectúa una transformación logarítmica para convertir dicho producto en un sistema lineal. Cuando no hay ningún efecto significativo, el valor esperado en cada celda es igual al total de observaciones dividido por el número de celdas; si los totales marginales son diferentes, esto se debe a la presencia de algún efecto. Si un modelo se ajusta a los datos, es decir, explica de manera satisfactoria las relaciones existentes entre las variables, el valor esperado en cada celda no se diferencia significativamente de la frecuencia observada en la misma; si son diferentes, debe probarse otro modelo. Además, dependiendo de los parámetros que resulten significativos en el modelo ajustado, este tipo de análisis permite describir las características de las variables y sus interacciones, por lo que Mellenbergh (1982b) los propuso como herramientas útiles para la detección de diferentes tipos de DIF. En Knoke & Burke (2000) o en Powers & Xie (2000) se encuentran explicaciones sencillas y completas de este tipo de modelos estadísticos, aquí se hace una breve aproximación de su aplicación a la detección de ítems con DIF.

En general, el valor esperado en una de las celdas puede expresarse como $E(X) = N \times p(X)$, donde N es el número de observaciones en ausencia de efecto y $p(X)$ es la probabilidad de los diferentes efectos; y, como es bien sabido, bajo la hipótesis nula de independencia de los efectos, la probabilidad de una celda se expresa en términos multiplicativos. Así,

siguiendo con la notación utilizada hasta ahora, puede plantearse que el valor esperado en la primera celda ($1rk$) de la tabla 2, es $e_{1rk} = N \times \beta_{I(1)} \times \beta_{J(r)} \times \beta_{K(k)}$;

donde: $\beta_{I(1)}$ es el efecto principal de la respuesta al ítem (variable I) en la celda 1 (acierto),

$\beta_{J(r)}$ es el efecto principal del grupo (variable J) en la celda r (grupo de referencia)

$\beta_{K(k)}$ es el efecto principal del nivel de magnitud de atributo (variable K) en la celda k (*k-ésimo* estrato).

Utilizando una transformación logarítmica el modelo anterior puede expresarse mediante un sistema lineal de la forma $ln(e_{1rk}) = \lambda + \lambda_{I(1)} + \lambda_{J(r)} + \lambda_{K(k)}$, donde los λ son los efectos respectivos (parámetros del modelo). Este sistema lineal expresa un modelo de efectos principales pero también pueden generarse modelos de interacciones de segundo, tercer o más orden, dependiendo del número de variables. En la detección de DIF un modelo con interacciones de segundo orden incluye las interacciones entre respuesta al ítem y grupo, respuesta al ítem con magnitud de atributo, y grupo con magnitud de atributo. Un modelo saturado es el que incluye todos los efectos principales e interacciones posibles; es decir, incluye además de las interacciones mencionadas, la de los tres efectos, y se expresa en general, como: $ln(e_{ijk}) = \lambda + \lambda_{I(i)} + \lambda_{J(j)} + \lambda_{K(k)} + \lambda_{IJ(ij)} + \lambda_{IK(ik)} + \lambda_{JK(jk)} + \lambda_{IJK(ijk)}$.

En los estudios de sesgo, interesa probar tres modelos, a partir de los cuales se pueden probar las hipótesis de interés sobre presencia y tipo de DIF (ver tabla 3). Si el único modelo que se ajusta es el saturado, que incluye todos los términos de interacción, se acepta que el ítem tiene DIF no uniforme; si el mejor modelo es el que excluye la interacción de tercer orden (ítem x grupo x magnitud de atributo), se acepta que el ítem presenta DIF uniforme; y si el mejor modelo excluye además la interacción entre grupo y respuesta al ítem, se afirmará que el ítem no presenta DIF. La tabla 3 sintetiza la relación entre el modelo elegido y la hipótesis sobre DIF.

Tabla 3

Relación entre el modelo log-lineal y logit ajustado y la hipótesis sobre existencia y tipo de DIF

Hipótesis	Modelo log-lineal	Modelo logit
DIF no uniforme	$ln(e_{ijk}) = \lambda + \lambda_{I(i)} + \lambda_{J(j)} + \lambda_{K(k)} + \lambda_{IJ(ij)} + \lambda_{IK(ik)} + \lambda_{JK(jk)} + \lambda_{IJK(ijk)}$	$ln\left(p_{ijk} / 1 - p_{ijk}\right) = \tau + \tau_{J(j)} + \tau_{K(k)} + \tau_{JK(jk)}$
DIF uniforme	$ln(e_{ijk}) = \lambda + \lambda_{I(i)} + \lambda_{J(j)} + \lambda_{K(k)} + \lambda_{IJ(ij)} + \lambda_{IK(ik)} + \lambda_{JK(jk)}$	$ln\left(p_{ijk} / 1 - p_{ijk}\right) = \tau + \tau_{J(j)} + \tau_{K(k)}$
Ausencia de DIF	$ln(e_{ijk}) = \lambda + \lambda_{I(i)} + \lambda_{J(j)} + \lambda_{K(k)} + \lambda_{IK(ik)} + \lambda_{JK(jk)}$	$ln\left(p_{ijk} / 1 - p_{ijk}\right) = \tau + \tau_{K(k)}$

Siguiendo un algoritmo elegido[8] se realizan las estimaciones de los parámetros con la restricción de que la suma de todos los valores λ de un mismo efecto sea nula (igual a cero). Se trata de un análisis jerárquico, que, gracias a que los modelos son anidados, empieza con el modelo saturado (que se ajusta perfectamente a los datos) y va excluyendo términos no significativos empezando por los de orden superior, hasta llegar al modelo de efectos principales. En cada etapa del análisis se evalúa el ajuste del modelo que se basa en las diferencias existentes entre las frecuencias observadas y las esperadas a partir del mismo; y continúa comparando el ajuste del modelo anterior con el actual, habiendo excluido algún término. Según Bishop, Fienberg & Holland (1975) los dos estadísticos más frecuentemente usados para evaluar el ajuste son el Ji cuadrado.de Pearson y la razón de verosimilitud, conocida como G^2; los cuales siguen una distribución χ^2 con los mismos grados de libertad del modelo en cuestión. Dependiendo de cuál modelo resulte con mejor ajuste a los datos, se puede identificar los ítems con DIF y el tipo de DIF, como se ilustra en la tabla 3

Los modelos logit pueden verse como un caso especial de los modelos log-lineales, cuando la variable respuesta al ítem se considera dependiente de las otras dos y se realiza una transformación logit $log\,it(\,p\,) = ln\left(\frac{p}{1-p}\right)$ que no es otra cosa que el logaritmo del odds de los aciertos en el ítem, donde p es la proporción de acierto en el mismo. Así, los tres modelos antes mencionados pueden expresarse en términos de la trasformación logit como parece en la tabla 3, donde los τ representan los diferentes efectos de acuerdo con la notación utilizada hasta ahora. El análisis sigue la misma lógica ya expuesta en los modelos log-lineales; es decir, se trata de encontrar el modelo que mejor se ajuste a los datos y de acuerdo con él, concluir sobre la presencia o tipo de DIF del ítem analizado.

La aplicación de estos modelos a los análisis de sesgo de los ítems tiene la misma limitación ya comentada de los métodos que utilizan la puntuación observada en la prueba como medida de la magnitud de atributo. Siguiendo los mismos principios del procedimiento de dos etapas, Van Der Flier, Mellenbergh, Adèr & Wijn (1984) propusieron un procedimiento logit iterativo para purificar la medida de magnitud de atributo. El estudio de Kok, Mellenbergh & Van Der Flier (1985) y, más recientemente, el de Navas-Ara & Gómez Benito (2002b) muestran que purificar la escala de medida de la magnitud de atributo mejora la precisión del método. Sin embargo, Mellenbergh (1989) encontró que este procedimiento no es muy preciso cuando el porcentaje de ítems con DIF en la prueba es grande ya que la estimación de la magnitud de atributo debe hacerse solamente con los

[8] Según Fidalgo (1996a) los algoritmos más utilizados en los estudios sobre DIF son el de Newton-Rapson y el de ajuste proporcional iterativo.

ítems sin DIF y no resulta muy precisa. De otra parte, Fidalgo & Paz Caballero (1995) en un estudio con datos simulados encontraron que los modelos log-lineales tienen una tasa de detección bastante alta cuando la magnitud de DIF es grande e incluso cuando ésta es moderada y los grupos tienen tamaño superior a 200; sin embargo, comparado con otros métodos como el MH, su error tipo I resulta demasiado alto; los autores concluyen que aunque estos modelos presentan ventajas demostradas teóricamente por la posibilidad de evaluar las relaciones entre todas las variables y de detectar tanto DIF uniforme y no uniforme, su uso no parece más aconsejable que el del MH en los estudios aplicados.

Regresión logística

La RL, procedimiento propuesto para la detección de DIF uniforme y no uniforme por Spray & Carlson (1986), Bennet, Rock & Kaplan (1987) y Swaminathan & Rogers (1990), puede verse como un caso especial de los análisis de regresión múltiple en el cual la variable dependiente es dicótoma; o como una extensión del MH donde la magnitud de atributo no se categoriza[9]. Un modelo de regresión múltiple para explicar la proporción de acierto en un ítem tendría la forma $p = \beta_0 + \beta_1\theta + \beta_2 J + \beta_3(\theta J)$, donde θ es la magnitud de atributo, J es el grupo de pertenencia y los β representan los parámetros de los respectivos efectos. En este modelo, sin embargo, la variable respuesta sólo toma valores entre 0 y 1 luego tendría distribución bernoulli o binomial (Christensen, 1997), de manera que no se cumplen los supuestos de la regresión lineal y además, se tendrían serias limitaciones en la estimación de los parámetros. Al realizar la transformación $logit(p) = ln\left(\frac{p}{1-p}\right)$, comentada en el apartado anterior, el modelo quedaría de la forma $ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\theta + \beta_2 J + \beta_3(\theta J)$, que suele expresarse en términos de la conocida función logística como $p(u_i = 1 \mid \theta, J) = \frac{e^z}{1+e^z}$, o sencillamente como $p_i = \frac{e^z}{1+e^z}$, con $z = \tau_0 + \tau_1\theta + \tau_2 J + \tau_3(\theta J)$,

donde: $p_i = P(u_i = 1 \mid \theta, J)$ es la probabilidad de acierto en el ítem, toma valores entre 0 y 1

θ es la magnitud de atributo medido y entra en el modelo como una variable numérica medida a partir del puntaje total en la prueba

[9] En Kleinbaum (1994), Bishop, Fienberg & Holland (1975) y Christensen (1997) se encuentran explicaciones detalladas sobre este tipo de modelos estadísticos y sus fundamentos. Aquí se presenta su aplicación a la detección de DIF.

$J = 0$ ó 1 dependiendo del grupo al que pertenezca el examinado, y

τ_0, τ_1, τ_2 y τ_3 son los parámetros constante, del efecto de habilidad, del efecto de grupo y del efecto de interacción, respectivamente.

El análisis parte entonces de un modelo logístico que explica la probabilidad de acierto en el ítem en función de la magnitud de atributo, el grupo de pertenencia del examinado y la interacción entre éstos. Las estimaciones de dichos parámetros ($\hat{\tau}_0, \hat{\tau}_1, \hat{\tau}_2$ y $\hat{\tau}_3$) se obtienen por máxima verosimilitud. Si $\tau_3 = 0$ con $\tau_2 \neq 0$, el ítem tiene DIF uniforme y si la variable grupo se codifica de manera que se asigna el valor 1 al grupo de referencia y 0 al grupo focal, $\tau_2 > 0$ indica DIF a favor del grupo de referencia y $\tau_2 < 0$ indica DIF a favor del grupo focal. El DIF no uniforme ocurre cuando $\tau_3 \neq 0$, si $\tau_3 > 0$ el ítem favorece a los miembros del grupo de referencia de alta magnitud de atributo y a los miembros del grupo focal con baja magnitud de atributo, mientras que $\tau_3 < 0$ indica DIF a favor de las personas de baja magnitud de atributo del grupo de referencia y alta magnitud de atributo del grupo focal. Hidalgo Montesinos & Gómez Benito (2003) distinguen además, DIF no uniforme simétrico y no simétrico; si siendo $\tau_3 \neq 0$, también $\tau_2 \neq 0$ el ítem muestra DIF no uniforme asimétrico y si $\tau_2 = 0$, entonces se trata de DIF no uniforme simétrico. Estas hipótesis pueden someterse a prueba comparando la razón de verosimilitud, de los modelos de regresión incluyendo y sin incluir el parámetro de interés (ver Bishop, Fienberg & Holland, 1975), la diferencia en el logaritmo de la función de verosimilitud obtenida incluyendo y sin incluir el parámetro τ_3 constituye la prueba de DIF no uniforme; así mismo tal diferencia entre el modelo que incluye y excluye τ_2 sin τ_3, es la prueba de DIF uniforme.

La hipótesis de ausencia de DIF tanto uniforme como no uniforme, formulada en términos de los parámetros del modelo, es $H_0 : \tau_2 = \tau_3 = 0$, lo que indicaría que la probabilidad de acierto depende solamente de la magnitud de atributo medido. La diferencia en el logaritmo de la función de verosimilitud entre el modelo compacto que incluye solamente el parámetro constante y el de magnitud de atributo $\left(p_i = \frac{e^{\tau_0 + \tau_1 \theta}}{1 + e^{\tau_0 + \tau_1 \theta}} \right)$ y el modelo saturado, que incluye todos los parámetros $\left(p_i = \frac{e^{\tau_0 + \tau_1 \theta + \tau_2 J + \tau_3 \theta J}}{1 + e^{\tau_0 + \tau_1 \theta + \tau_2 J + \tau_3 \theta J}} \right)$, sigue una distribución χ^2 con dos grados de libertad. De esta manera, se puede someter a prueba la hipótesis de ausencia de ambos tipos de DIF simultáneamente. Sin embargo, Swaminathan & Rogers (1990) propusieron que la hipótesis de ausencia de DIF

tanto uniforme como no uniforme, puede formularse matricialmente como $H_0 : C\hat{\tau}' = 0$ y someterse a prueba mediante el estadístico χ^2 con *2* grados de libertad, $\chi^2_{RL} = \hat{\tau}' C' \left(C \sum C\right)^{-1} C\hat{\tau}'$

donde: $\hat{\tau} = \left[\hat{\tau}_0 \quad \hat{\tau}_1 \quad \hat{\tau}_2 \quad \hat{\tau}_3\right]$ representa el vector de estimaciones de los parámetros,

Σ es la matriz de varianzas y covarianzas entre las estimaciones de los parámetros, y

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Resulta evidente la similitud entre este modelo y los modelos logit comentados antes; puede mostrarse que las dos aproximaciones son equivalentes cuando en el modelo de regresión las variables independientes se codifican como variables Dummy. En las aplicaciones frecuentes de detección de DIF, la principal diferencia entre los dos métodos es que en los modelos logit, como en los métodos presentados antes, la magnitud de atributo se maneja como una variable categórica, mientras que en la RL se considera la naturaleza continua de la magnitud de atributo, característica que se menciona con frecuencia como una de las ventajas de la RL. French & Miller (1996), Zumbo, (1999) e Hidalgo Montesinos & Gómez Benito (2003) destacan además, su facilidad y versatilidad para ajustarse al análisis de ítems politómicos o con más de dos grupos, mientras que Swaminathan & Rogers (1990), Millsap, & Everson (1993) y Jodoin & Huff (2001) enfatizan en la capacidad de la RL para detectar tanto DIF uniforme como no uniforme. Swaminathan & Rogers (1990) y Rogers & Swaminathan (1993) advierten sin embargo, que el grado de libertad que se pierde al incluir un parámetro para detectar DIF no uniforme, puede disminuir su poder para detectar DIF uniforme.

A pesar de sus bondades, la RL ha recibido algunas críticas por el efecto que pueden tener los ítems con DIF en el criterio de equiparación de los grupos, la inflación de su error tipo I sobre todo con altos tamaños de muestra y la carencia de alguna medida de la magnitud del DIF. Con respecto a la primera dificultad, compartida con todos los métodos que utilizan como criterio de equiparación la puntuación observada en la prueba, se han propuesto procedimientos iterativos como los ya descritos, que buscan corregir el efecto de los ítems con DIF sobre los puntajes de los examinados en la prueba, como medida de la magnitud de atributo. En general se trata de identificar los ítems con DIF y excluirlos del criterio de equiparación para romper la circularidad que se genera cuando éstos tienen algún peso en el criterio. Como se ha reportado con otros métodos, múltiples trabajos como los de Rogers & Swaminathan (1993), Gómez Benito & Navas-Ara, (1996) y Navas-Ara & Gómez Benito (2002b), entre otros, han mostrado que el procedimiento iterativo de purificación del criterio de equiparación mejora la potencia de la RL. Además, Gómez Benito & Hidalgo Montesinos (1997a) y Hidalgo Montesinos & Gómez Benito (1996, 2003) han utilizado también la purificación del criterio con regresión logística multinomial, con resultados satisfactorios.

Con respecto a la carencia de una medida de la magnitud de DIF, Zumbo & Thomas (1996) y Zumbo (1999) propusieron una medida de mínimos cuadrados ponderados que permite interpretar un valor en términos de la magnitud de DIF tanto uniforme como no uniforme. Si β_j es el coeficiente de regresión estandarizado para la variable j en el modelo, y r_j es la correlación de la misma variable y la variable respuesta, la contribución de cada variable en el modelo puede definirse como $R^2\Delta = R_1^2 - R_2^2$

donde: $R_1^2 = \sum_{j=1}^{p} \beta_j r_j$ para las p variables incluidas en el modelo saturado

$R_2^2 = \sum_{j=1}^{k} \beta_j r_j$ para las k variables del modelo compacto, (k<p) en el que se ha excluido la variable de interés

De esta manera puede evaluarse la contribución del efecto de grupo ($\tau_2 J$) y de la interacción entre magnitud de atributo y grupo ($\tau_3 \theta J$) e interpretarse como una medida de la magnitud de DIF uniforme y no uniforme, respectivamente. Mediante un estudio con datos simulados y tomando como criterio los parámetros de clasificación del MH (Zieky, 1993) y del SIBTEST (Roussos & Stout, 1996), Jodoin & Gierl (2001) propusieron una clasificación de la magnitud de DIF, con base en los valores del $R^2\Delta$, así: DIF despreciable si $R^2\Delta < .035$, DIF moderado si se rechaza la hipótesis y $.035 < R^2\Delta < .07$ y, DIF amplio si se rechaza la hipótesis y $R^2\Delta > .07$. De acuerdo con los autores, esta categorización concuerda con la de Zieky (1993), adoptada por el ETS para el MH, y la de Roussos & Stout (1996) desarrollada par el SIBTEST. Jodoin & Gierl (2001); Jodoin & Huff (2001) evaluaron el funcionamiento de esta medida del efecto bajo diferentes condiciones de tamaño de muestra y distribución de la magnitud de atributo y encontraron que este criterio de clasificación no solamente provee una medida satisfactoria de la magnitud de DIF, sino que mantiene controlado el error tipo I, incluso con muestras grandes. Sin embargo, Hidalgo Montesinos & López Pina (2004) encontraron que tanto el sistema de clasificación de Zumbo & Thomas (1996) como el de Jodoin & Gierl (2001) fueron poco sensitivos, puesto que sólo clasificaron un máximo del 20% de ítems con DIF, en las categorías de moderado o alto. Además, en el estudio de Jodoin & Gierl (2001) se encontraron tasas de detección bajas para DIF uniforme y moderadas para DIF no uniforme con grupos de 250 examinados, las primeras solamente mejoraron hasta un nivel aceptable con grupos iguales de 1000 examinados por grupo. El comportamiento de la RL en función de diferentes tamaños de grupo se tratará más adelante en el capítulo 5.

Capítulo 3:
PROCEDIMIENTOS BASADOS EN LA TEORIA DE RESPUESTA AL ITEM

Aunque ya se ha hecho uso de algunas nociones de la Teoría de Respuesta al Item (TRI) en los capítulos anteriores, se incluye aquí una breve presentación del antes de tratar los procedimientos que se han generado con base en esta aproximación. En primer lugar se hace una breve presentación de los principios conceptuales y metodológicos de la TRI y posteriormente se describen los tres enfoques utilizados hasta ahora para la detección de DIF y algunas de las técnicas derivadas de los mismos.

Introducción a la Teoría de Respuesta al Item[10]

La TRI, modelo de medición psicológica que surgió, según Muñiz (1997), Muñiz & Hambleton (1992) y Hambleton, Swaminathan & Rogers (1991), como una alternativa que supera algunas de las limitaciones de la popular TCT, se desarrolló a partir de dos ideas básicas: el concepto de rasgos latentes y la relación funcional entre éstos y las respuestas de los examinados.

En primer lugar se supone que las respuestas de los examinados en una prueba pueden ser satisfactoriamente explicadas si se definen claramente los rasgos latentes que intervienen en la misma, y se generan procedimientos que permitan ubicar al individuo en una escala que represente la magnitud de tales rasgos. Desde esta perspectiva, el problema central consiste en hallar las relaciones entre las puntuaciones observadas y la magnitud del o los rasgos que mide la prueba. El significado del término rasgo latente tiene una connotación estadística, en otras palabras, "... no implica que existan rasgos latentes o habilidades subyacentes en un sentido físico o fisiológico, ni tampoco que originen una conducta. Lo rasgos latentes son constructos derivados matemáticamente de relaciones empíricas observadas entre las respuestas a la prueba" (Anastasi y Urbina, 1998, p. 73). De esta forma, los términos *rasgo latente*, *variable latente*, *factor, atributo* y *habilidad* se refieren relaciones entre variables. Este postulado es un principio de organización conceptual que relaciona la ejecución del examinado con el constructo medido y es el punto de conexión entre el objeto de medida y el referente comportamental que permite cuantificarlo.

En segundo lugar se supone que la probabilidad de puntuar en un ítem y la magnitud de atributo que mide el mismo pueden relacionarse mediante un modelo matemático, la Función Característica del Ítem (FCI), cuya representación gráfica se conoce como Curva Característica del Ítem (CCI). La CCI es una función de incremento monotónico entre el nivel del atributo que mide el ítem y la probabilidad de puntuar en el mismo; es un modelo teórico que representa la relación entre un

[10] Algunas partes de esta sección se han tomado, con autorización de los autores, de Herrera, Sánchez & Jiménez (2001)

parámetro, nivel del rasgo notado como θ, y la probabilidad de puntuar en el ítem; en este sentido la CCI es diferente a la representación de la probabilidad empírica de acierto en un ítem en función de los puntajes observados en la prueba. La diferencia entre estas dos representaciones se abordó implícitamente cuando se compararon los métodos basados en el análisis de TC y los basados en la TRI, a través de la figura 7: la figura 7a representa la relación empírica entre la proporción de aciertos en un ítem y la puntuación observada en la prueba, mientras que la 7b representa un modelo logístico que relaciona la magnitud de atributo y la probabilidad de acierto en el ítem.

Sobre estos dos pilares básicos la distribución de probabilidad condicional de las puntuaciones en un ítem dicótomo dado un nivel de θ, se expresa en función de la respuesta al ítem como $f(U_i \mid \theta) = p_i(\theta)^{u_i}(1 - p_i(\theta))^{1-u_i}$, donde $p_i(\theta)$ es la probabilidad de puntuar en el ítem y u_i es la puntuación en el mismo, que solo toma valores 1 o 0. Puede verse que si la respuesta es acertada $(u_i = 1)$, la función es $f(U_i \mid \theta) = p_i(\theta)$; mientras que si la respuesta es errada, $(u_i = 0)$ entonces $f(U_i \mid \theta) = 1 - p_i(\theta)$. En palabras de García Cueto (1993) la CCI es la curva que relaciona las medias de estas distribuciones condicionales. En las figuras 1 a 3 se han mostrado las CCI de ítems en diferentes condiciones y en la figura 10 se muestran las CCI de tres ítems en los que las probabilidades de puntuar en los ítems i, j y k para un examinado con un nivel de atributo igual a 0, son 0.85, 0.04 y 0.52 respectivamente; es decir, $p_i(\theta = 0) = 0.85$, $p_j(\theta = 0) = 0.04$ y $p_k(\theta = 0) = 0.52$.

Supuestos de la TRI

Para que los modelos TRI sean aplicables, y por tanto los procedimientos que se derivan para la detección de DIF, deben cumplirse dos supuestos que se conocen como dimensionalidad del espacio latente e independencia local. Brevemente el primer supuesto hace referencia a la identificación de las dimensiones que intervienen en la respuesta a un ítem; mientras que la independencia local hace referencia al hecho de que dada una magnitud de atributo, las respuestas a un conjunto de ítems no están correlacionadas.

Si la respuesta a un ítem puede explicarse a partir de N dimensiones, entonces θ puede representarse mediante un vector de N componentes que se puede notar como $\theta = |\theta_1 ..., \theta_n ... \theta_N|$ y que se conoce como espacio latente. El número de dimensiones del espacio latente o número de rasgos latentes necesarios para explicar la respuesta al ítem, definen la dimensionalidad del mismo. En teoría se trata de un espacio $N - dimensional$ en el que cada examinado puede representarse como un punto según su nivel en cada una de las N dimensiones. La dimensionalidad se refiere a la definición completa de tal espacio. Sin embargo, los modelos de la TRI que han tenido mayor desarrollo en la actualidad y por tanto han sido utilizados con más frecuencia para el diseño de procedimientos para detección de DIF, son los que miden sólo un rango latente, conocidos

como modelos unidimensionales[11]. Pero técnicamente no es posible que un instrumento de medición psicológica evalúe una única dimensión exactamente identificable; así, la unidimensionalidad se refiere más bien a la identificación de un factor "dominante" y, en este sentido, no es una propiedad de todo-nada, sino que debe tomarse como una cuestión de grado. En Cuesta (1996) se encuentra una completa discusión del tema de la unidimensionalidad de las pruebas y algunas aproximaciones metodológicas para su verificación. Debe anotarse además, que algunos estudios empíricos como los de Harvey & Hammer (1999), Cuesta & Muñiz (1994, 1995) y Drasgow & Parsons (1993) han encontrado que los modelos TRI unidimensionales son relativamente robustos a violaciones moderadas de este supuesto.

De otra parte, hay independencia local si para un nivel de magnitud de atributo no existe relación entre las respuestas de los examinados a diferentes ítems (Hambleton, Swaminathan & Rogers, 1991). Estadísticamente este supuesto se expresa como el producto de probabilidades: dado un θ, la probabilidad de puntuar en un conjunto de ítems es igual al producto de las probabilidades de puntuar en cada uno de ellos. Siguiendo la notación utilizada hasta ahora, existe independencia local para un conjunto de k ítems, si $P(u_1,...u_i,..u_k \mid \theta) = P_1(\theta)*..*.P_i(\theta)*...*P_k(\theta) = \prod_{i=1}^{k} P_i(\theta)$. La independencia local está estrechamente relacionada con la unidimensionalidad; si ésta última se cumple también se cumple la independencia local, pero no la inversa; puede haber independencia local en datos de más de una dimensión. Si una única dimensión es suficiente para explicar las respuestas de los examinados, es razonable pensar que para un θ dado, las respuestas a un par de ítems cualquiera serán independientes ya que, teóricamente, el único elemento que determina la respuesta, es la magnitud de atributo. En Crocker & Algina (1986), Lord, (1980) y Lord & Novick (1966) se encuentra un tratamiento más detenido de la relación entre estos dos supuestos.

Modelos y parámetros

Verificados los supuestos, la tarea consiste en ajustar el modelo que represente de la manera más precisa la relación entre la magnitud de atributo (θ, conocido como parámetro de los individuos) y la probabilidad de puntuar en el ítem particular. Históricamente se han trabajado dos tipos de modelos, el primero de ellos propuesto por Lord (1952), se basa en la distribución normal acumulada y el segundo tipo, desarrollado gracias a los trabajos de Rasch, (1960) y Birnbaum, (1958, 1968), se basan en la ya conocida función logística. Actualmente los modelos de ojiva normal tienen poco uso debido a que los logísticos resultan más sencillos y económicos.

[11] Algunos trabajos sobre el desarrollo de modelos TRI multidimensionales se encuentran en Reckase (1997) y en McDonald (1997); McDonald (2000). En Shealy & Stout (1993b) y Bolt (1996) se encuentran aplicaciones de este tipo de modelos en la detección de DIF.

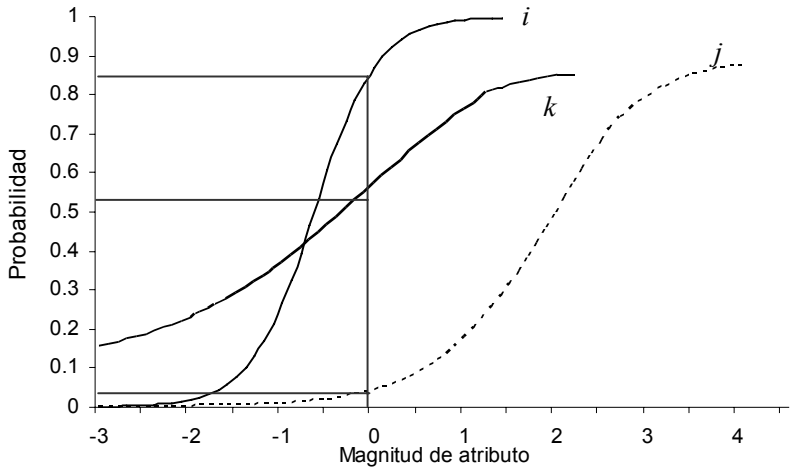


Figura 10 CCI de tres ítems con diferentes valores de parámetros para un mismo grupo de examinados (Tomada de Herrera, Sánchez & Jiménez (2001))

El otro criterio que determina el tipo de modelo es el número de parámetros de los ítems que se consideren relevantes para explicar la relación entre θ y $P_i(\theta)$. El modelo logístico tiene la forma $P_i(\theta) = \frac{e^{dx}}{1+e^{dx}}$, donde x es la expresión que relaciona los parámetros de los ítems que se consideren relevantes en el caso particular, y d es una constante que toma uno de dos valores, 1 o 1.7, y en el segundo caso permite una aproximación bastante precisa del modelo logístico a la función normal acumulada. Se han identificado tres parámetros de los ítems denominados dificultad, discriminación y pseudoazar, que se notan, para un ítem i, como b_i, a_i y c_i, respectivamente. Los modelos de tres parámetros (3P) tienen la forma $P_i(\theta) = c_i + (1-c_i)\frac{e^{dx}}{1+e^{dx}}$ con $x = a_i(\theta - b_i)$ y se ajustan cuando se supone que todos, b_i, a_i y c_i, son relevantes para explicar el comportamiento de las respuestas de los examinados al ítem. En el modelo de dos parámetros (2P) se supone que la probabilidad de acertar en el ítem por azar, si existe, es despreciable, y entonces se modifica la expresión haciendo $c_i = 0$. Finalmente, en el modelo de un parámetro (1P) o modelo de Rasch se supone además que la discriminación es constante para todos los ítems, se hace $a_i = 1$ y entonces la expresión queda sencillamente como $P_i(\theta) = \frac{e^{d(\theta - b_i)}}{1+e^{d(\theta - b_i)}}$. En la figura 10 se muestran las CCI de tres ítems con modelo logístico de tres parámetros.

El parámetro de dificultad se define en la misma escala de la magnitud de atributo, es el valor de θ en el punto de máxima pendiente de la CCI, lo que puede expresarse como el nivel de atributo necesario para puntuar en el ítem. Este parámetro define la posición de la CCI sobre la escala de θ; en la figura 10 la curva correspondiente al ítem *j* se ubica más hacia la derecha que los

otros dos y es el más difícil de los tres. Los valores del parámetro de dificultad para los tres ítems de dicha figura son $b_i = -.65,\ b_j = 2,\ b_k = -.1$. El parámetro de discriminación se define como la capacidad del ítem para distinguir entre examinados con altas y bajas magnitudes de atributo; puede asimilarse al nivel de inclinación de la curva y corresponde a la pendiente de la recta tangente a la CCI en su punto de mayor inclinación. La curva de mayor pendiente en la figura 10 es la del ítem *i*, a medida que la curva se hace más horizontal, la discriminación del ítem se acerca a *0*; así, el ítem de menos discriminación es el *k*. La discriminación para los tres ítems es $a_i = 1.7,\ a_j = .9,\ a_k = .55$. Finalmente, el tercer parámetro se define como la probabilidad de puntuar en un ítem por azar, corresponde a la probabilidad de acierto cuando θ es mínimo; sin embargo, dado que θ puede tomar valores entre $-\infty$ y ∞, *c* corresponde a la asíntota de la CCI cuando θ tiende a $-\infty$ pero este valor asintótico no es observable en las CCI puesto que generalmente se fija un valor mínimo de *-3* o *-4* para la escala de θ; así, el parámetro de pseudoazar puede verse como la mínima probabilidad de acertar. En la figura 10, el valor mínimo de θ es *-3* y en ese punto el ítem *k* tiene mayor probabilidad de acierto que los otros dos. Los valores del parámetro de pseudoazar para los tres ítems son: $c_i = c_j = 0$ y $c_k = .1$.

Escala de θ y puntuación verdadera en la prueba

Se ha afirmado previamente que θ puede tomar valores entre $-\infty$ y ∞ y que en la práctica se fija un intervalo de variación para tales valores. Lo que esto significa es que, en teoría, los modelos TRI pueden estar localizados en cualquier intervalo de valores sobre la recta real sin que esto afecte la relación entre la magnitud de atributo y la probabilidad de acertar al ítem. En la figura 11 se representa la CCI del ítem *i* de la figura 10 con $-3 \le \theta \le 3$ cuando se ha transformado para utilizar una escala entre 0 y 6. Obviamente cambiar la escala que se utilizó originalmente para las estimaciones, implica transformar también los valores de las estimaciones de los parámetros de los ítems. En el ejemplo de la figura 11 se ha realizado una transformación lineal de la forma $\theta' = m\theta + k$ con $m = 1$ y $k = 3$, de manera que un valor original de $\theta = -3$ quedó transformado en $\theta' = 1(-3) + 3 = 0$, mientras que para $\theta = 3$, $\theta' = 6$. Con este tipo de transformación, los parámetros de dificultad y discriminación quedan como $b_i' = mb_i + k$ y $a_i' = a/m$, respectivamente, mientras que el valor de c_i no cambia. Para el ítem del ejemplo, $b_i' = 1(-.65) + 3 = 2.35$, $a_i = 1.7/1 = 1.7$ y $c_i = 0$. Pero esta transformación lineal no es más que un ejemplo de los muchos tipos de transformaciones de la escala, que podrían utilizarse. Tanto Muñiz, (1997) como Hambleton, Swaminathan & Rogers (1991) explican en detalle los tipos de transformaciones más utilizadas.

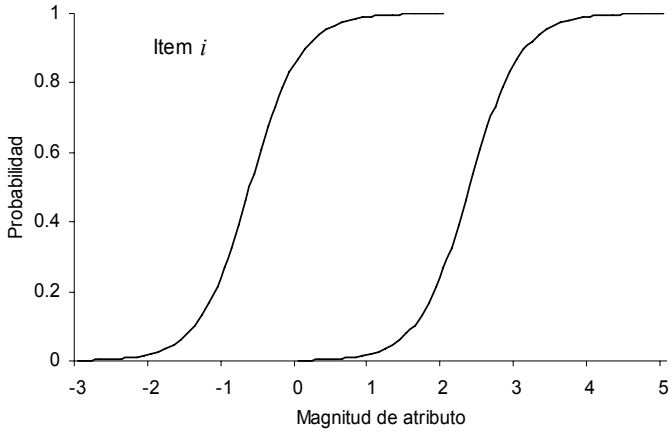


Figura 11: CCI del mismo ítem en dos intervalos diferentes para los valores de la magnitud de atributo, θ.

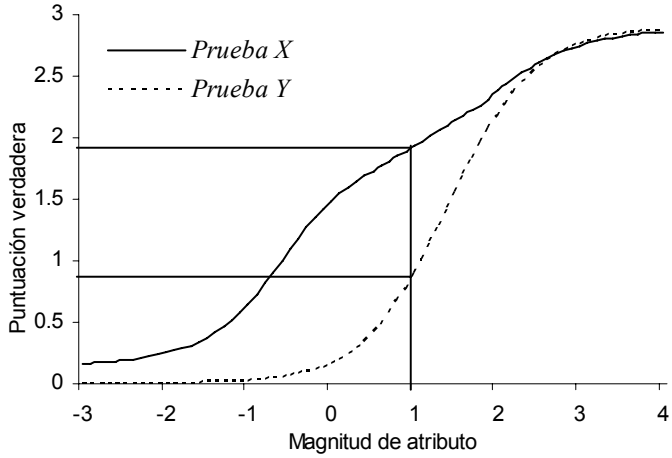


Figura 12. Curva característica de una prueba (X) conformada por los tres ítems de la figura 10 y una (Y) conformada por tres ítems de mayor dificultad

Un efecto práctico de las transformaciones de la escala de θ, dada su indeterminación, es la estimación de la puntuación verdadera de un sujeto en una prueba. Aunque el centro de atención de la TRI no es la estimación de dicha puntuación, (notada como τ_j para el individuo j) sino de su magnitud de atributo (θ_j), debe señalarse que la naturaleza de las dos escalas es diferente y tienen interpretaciones diferentes. θ es un parámetro de magnitud de atributo y es invariante puesto que su valor no depende de la prueba que responda el individuo; por su parte, τ es la puntuación que se predice o se espera en una prueba para un individuo con un determinado valor de θ; en consecuencia depende de los ítem de la prueba, y está en la misma escala de las puntuaciones observadas: $0 \leq \tau \leq k$ para una prueba de k ítems con modelos 1P o 2P, y $\sum_{i=1}^{k} c_i \leq \tau \leq k$ cuando se trata de modelos de tres parámetros. La puntuación esperada para un

individuo j no es otra cosa que la suma de las probabilidades de acierto en los ítems de la prueba dada su magnitud de atributo; esto es, $\tau_j = \sum_{i=1}^{k} P_i(\theta_j)$. En la figura 12 se representa la suma de las CCI de los ítems de la figura 10, que conforman una prueba X, y la suma de las CCI de otros tres ítems que conforman la prueba Y. Esta representación se conoce como curva característica del test (CCT) e indica las puntuaciones verdaderas en una prueba para individuos con diferentes magnitudes de atributo. Nótese que si se sabe que un individuo j tiene una magnitud de atributo de, por ejemplo, $\theta = 1$, se puede esperar (predecir) que acierte 2 ítems de la prueba X y sólo uno de la prueba Y, sin que haya respondido estas pruebas específicas.

Equiparación de puntuaciones

Otro aspecto relacionado con la indeterminación de la escala de θ es la equiparación de los resultados de pruebas diferentes aplicadas al mismo individuo o grupo de individuos, o de los resultados de la misma prueba aplicada a diferentes examinados. De acuerdo con Kolen (2004) la primera referencia sobre el tema, con el término de "comparabilidad", se puede encontrar en Flanagan (1951), sin embargo, la mayoría de revisiones citan a Angoff (1984) y a Lord (1980) como los pioneros en el tratamiento del tema. Estos últimos entienden la equiparación como un proceso que implica expresar el sistema de unidades de una prueba en el de otra, de manera que sus resultados sean comparables o equivalentes; el problema consiste entonces en hallar una escala o métrica común para dos o más medidas de un mismo atributo de manera que se puedan comparar resultados de individuos o de instrumentos diferentes. La construcción de dichos sistemas comunes tiene una enorme utilidad en la práctica cuando se trata de tomar decisiones sobre asignación de cupos educativos o empleos con base en pruebas diferentes de un mismo atributo o dominio, o en procesos que impliquen comparar el desempeño de dos candidatos que han presentado pruebas diferentes. En este sentido resulta interesante hacer notar que de una ausencia casi total del tema en los libros clásicos de psicometría, se ha pasado a un gran volumen de producción al respecto desde diferentes perspectivas; en los últimos años se ha observado un importante incremento en el número de publicaciones por parte de revistas científicas especializadas y la *Applied Psychological Measurement*, una de las publicaciones más tradicionales y consultadas en la actualidad, dedicó el N° 4 de 2004 exclusivamente al tema.

Pero afirmar que la equiparación de puntuaciones está relacionada con la indeterminación de la escala de θ, no implica que este problema sea propio de la TRI; por el contrario, si se tiene en cuenta que las estimaciones de los parámetros en los modelos TRI son invariantes, el tema pierde importancia puesto que conocidos los parámetros de los ítems, la estimación de la magnitud de atributo (no de la puntuación verdadera) es independiente del grupo de ítems incluidos en la prueba; y a su vez, conocida la magnitud de atributo de los individuos, la estimación de los parámetros de los ítems es independiente del grupo de examinados. Sin embargo, en la práctica suele ocurrir

que no se dispone del conocimiento de los parámetros de unos u otros. Si se han ajustado los modelos de los ítems y se dispone de las CCT en la misma escala como las que se presentan en la figura 12, entonces la equivalencia es inmediata: en el ejemplo de la figura, puede afirmarse que responder correctamente dos pregunta de la prueba X es equivalente a responder una de la prueba Y. Desde esta perspectiva, Hambleton, Swaminathan & Rogers (1991) sugieren que en el marco de la TRI no se hable de equiparación (*equating*) sino de 'escalamiento' (*scaling*) para referirse a la necesidad de elegir una escala común para las estimaciones de los parámetros tanto de los ítems como los individuos, cuando éstas se basan en pruebas diferentes o grupos diferentes (p. 125). Navas Ara (1996) y Kolen & Brennan (1995) constituyen dos de los trabajos más completos sobre el tema en los que se puede encontrar una detallada revisión de los diseños y los métodos desarrollados para la equiparación de puntuaciones desde la TCT y desde la TRI.

Háblese de equiparación, como es la costumbre, o de escalamiento, siguiendo la sugerencia de Hambleton, Swaminathan & Rogers (1991), en la práctica jamás se tendrán los mismos valores de los parámetros cuando éstos se estiman a partir de las respuestas de dos grupos diferentes, y en la detección de DIF resulta de gran importancia garantizar la comparabilidad entre tales estimaciones, lo que implica tenerlos expresados en la misma métrica. En el marco de la TRI este proceso de equiparación puede entenderse como una transformación lineal de los parámetros, del tipo que se presentó en el apartado anterior y que se ilustró en la figura 11: para las constantes m y k con $m > 0$, los parámetros de los individuos y de los ítems pueden expresarse como $\theta' = m\theta + k$, $a_i' = a_i / m$, $b_i' = mb_i + k$ y $c_i' = c_i$, sin que se altere la probabilidad de acierto en el ítem, es decir, se cumple que $P_i(\theta, a_i, b_i, c_i) = P_i(\theta', a_i', b_i', c_i')$. Así, el proceso consiste en encontrar los valores de las constantes m y k que permitan expresar unas estimaciones en la métrica de otras; la estrategia más utilizada es la denominada "prueba de anclaje" que consiste en incluir en las dos aplicaciones un grupo de ítems comunes que constituyen la base para la equiparación puesto que se dispone de dos estimaciones de sus parámetros. Este procedimiento también es aplicable cuando se incluyen 'individuos de anclaje' y se dispone de dos estimaciones de su magnitud de atributo; la siguiente exposición se enmarca en el primer caso, no sin mencionar que las transformaciones presentadas para el parámetro b, son aplicables a θ.

Disponiendo de dos estimaciones de los parámetros de un mismo grupo de ítems, los métodos más utilizados para el cálculo de las constantes son: el de regresión, los de media-desviación y los basados en la CCI. Sustentado en la relación lineal existente entre dos estimaciones diferentes de los parámetros del mismo ítem, el primer método consiste en hallar los valores de los parámetros del modelo de regresión $b_{a2} = mb_{a1} + k + e$, donde b_{a1} y b_{a2} son las dos estimaciones de b para el ítem de anclaje a, y e es el error aleatorio del modelo. Partiendo del mismo supuesto, el razonamiento de los métodos de media-desviación típica es el siguiente: Si $b_{a2} = mb_{a1} + k$, entonces

se cumple que $\bar{b}_{a2} = m\bar{b}_{a1} + k$ y que $s_{a2} = ms_{a1}$, donde $\bar{b}_{a1}$ y $\bar{b}_{a2}$ son las medias de las estimaciones de dificultad para los ítems de anclaje en la primera y segunda aplicación, respectivamente; y s_{a1} y s_{a2} son las desviaciones típicas de las mismas estimaciones. Despejando para m y para k se obtienen los valores de dichas constantes en términos de la media y desviación típica de las estimaciones de dificultad en las dos aplicaciones. Dos propuestas de modificación que han intentado mejorar este método general son las de Linn, Levine, Hastings & Wardrop (1981) y la de Stocking & Lord (1983). La primera consiste en calcular las medias y desviaciones típicas de la dificultad a partir de las estimaciones ponderadas de acuerdo con su error estándar; y la segunda consiste en un procedimiento iterativo que parte de estas estimaciones ponderadas pero va ajustando el peso de manera que se evite la influencia de posibles datos extremos (*outlier*).

Tanto el método de regresión como los de media-desviación típica sólo toman en consideración el parámetro de dificultad, los métodos basados en la CCI toman toda la información contenida en la ella. El procedimiento que se conoce hoy como el método de la función de respuesta al ítem, propuesto originalmente por Haebara, (1980 en Navas Ara, 1996) y Stocking & Lord (1983), deriva los valores de las constantes de equiparación minimizando las diferencias entre las puntuaciones verdaderas de los individuos, en las dos estimaciones. Si τ_{jX} y τ_{jY} son las puntuaciones verdaderas del individuo j en los ítems o subpruebas de anclaje X y Y, respectivamente; las constantes de equiparación, m y k, serán los valores que minimicen la función $F = \frac{1}{n}\sum_{j=1}^{n}\left(\tau_{jX} - \tau_{jY}\right)^2$. Siguiendo el mismo razonamiento, Divgi (1985) propuso el método conocido como χ^2 mínimo ya que minimiza una función de las diferencias entre las estimaciones de los parámetros en las dos aplicaciones, función que coincide con el estadístico χ^2 de Lord para la detección de DIF y que se presentará con algo más de detalle en la siguiente sección. Más recientemente Ogasawara (2001) propuso una estimación por mínimos cuadrados que no ha sido muy evaluada hasta el momento.

Hoy se dispone de una buena cantidad de trabajos que evalúan la eficiencia relativa de cada uno de estos métodos o las condiciones bajo las cuales resultan más recomendables; entre ellos se pueden citar Candell & Drasgow (1988), Lautenschlager & Park (1988), Baker & Al-Karni (1991), Kim & Cohen (1992), Millsap & Everson (1993), Baker (1996), Kaskowitz & De Ayala (2001), Hidalgo Montesinos & López Pina (2002), entre muchos otros. Sin embargo, los resultados no son concluyendo a favor de uno u otro procedimiento. Algunas investigaciones (Stocking & Lord, 1983; Kim & Cohen, 1992, 1994) sugieren que el método de función de respuesta al ítem puede resultar más preciso comparativamente, y Kaskowitz & De Ayala (2001) encontraron que este procedimiento tal como lo implementa el EQUATE de Baker (1995), es relativamente robusto a la magnitud de error de estimación de los parámetros. Por el contrario, Candell & Drasgow (1988) encontraron que el procedimiento de media-desviación típica ponderadas, que es más fácil

y económico, mostraba mejores resultados; para Park & Lautenschlager (1990) el mejor método de equiparación, considerando precisión y costos es el χ^2 mínimo, y Baker & Al-Karni (1991) no encontraron diferencias entre métodos de media-desviación típica y los basados en la CCI. Además, aunque algunas revisiones (Gómez Benito & Hidalgo Montesinos (1997b) parecen indicar que los métodos basados en la CCI se utilizan más frecuentemente que otros, una somera revisión de publicaciones actuales sobre trabajos aplicados en la detección de DIF, tampoco parece favorecer a unos u otros puesto que no se observa una clara predominancia en su uso.

Detección de DIF con base en la TRI

Los métodos que se basan en la TRI para la detección de DIF se sustentan en la comparación de las CCI de un ítem cuando se ajustan modelos separadamente para los dos grupos de interés. Brevemente, estos procedimientos ajustan modelos TRI para los dos grupos independientemente, evalúan su ajuste a los datos[12], expresan los parámetros en la misma métrica y comparan los dos modelos. La idea de fondo es que el DIF se manifiesta mediante diferencias en las CCI del ítem cuando se calcula para los dos grupos separadamente (ver figuras 2 y 3); así, si un ítem no tiene DIF las CCI de los grupos de referencia y focal deben coincidir salvo por pequeñas variaciones atribuibles al azar como se mostró en la figura 1. El punto clave es entonces la identificación del procedimiento que conduzca a una comparación más fina y precisa de las CCI para los dos grupos. Las diferentes propuestas metodológicas se han agrupado en tres: comparación de los modelos ajustados, comparación de los parámetros de los mismos y medidas del área de discrepancia entre las CCI.

Pero cualquiera que sea la estrategia utilizada, el proceso de detección de ítems con DIF desde la perspectiva de la TRI exige abordar previamente dos temas: la elección del algoritmo de estimación de los parámetros y el control del posible efecto de los ítem con DIF en el cálculo de las constantes de equiparación y en la estimación de θ. En primer lugar, la precisión en la estimación de los parámetros tanto de los individuos como de los ítems, y por tanto en la detección del DIF, depende en parte del procedimiento de estimación utilizado (McLaughlin & Drasgow, 1987; Lim & Drasgow, 1990; Cohen, Kim & Subkoviak, 1991; Kim & Cohen, 1994); los algoritmos de estimación más frecuentes han sido el de Máxima Verosimilitud Conjunta (MVC, Lord, 1980), Máxima Verosimilitud Marginal (MVM) de Bock & Aitkin (1981) y la Modal Bayesiana (MB) de Mislevy (1986) y Mislevy & Bock (1990). Aunque los dos primeros se basan en el mismo principio: maximizar la función de verosimilitud, una diferencia importante entre ellos es que el primero supone que el parámetro θ es conocido, mientras que la MVM supone que los $\hat{\theta}$ estimados constitu-

[12] En Andrich (1998) se encuentra una presentación y discusión del tema del ajuste de los modelos.

yen una muestra de la distribución de la magnitud de atributo[13]. El primer supuesto no se cumple en la práctica mientras que el segundo es menos difícil de cumplir y en todo caso, es más débil que el primero; la literatura disponible hasta el momento indica que las estimaciones por MVC son menos precisas que las arrojadas por los otros dos procedimientos, y además, se ha reportado una superioridad de las estimaciones bayesiana con grupos pequeños (Lim & Drasgow, 1990; Kim & Cohen, 1994).

El segundo aspecto hace referencia al posible 'efecto contaminante' de los ítems con DIF en la estimación de la magnitud de atributo como criterio de igualación de los grupos, y en el cálculo de las constantes de equiparación. Para controlar este efecto se han propuesto procedimientos de 'purificación' que buscan, en varias etapas o iterativamente, identificar los ítems con DIF y excluirlos para efectos de la estimación de θ o del cálculo de m y k. La propuesta pionera en este sentido fue la publicada en el texto de Lord, sugerida por Marco (1977, en Lord, 1980), que implica comparar las CCI de todos los ítems para identificar los que presentan DIF, estimar θ a partir de los ítems que no presentan DIF uniendo los dos grupos, estimar los parámetros de los ítems utilizando como valores fijos estas estimaciones de θ, y comparar nuevamente las CCI con esta estimaciones 'purificadas' de los parámetros de los ítems. Posteriormente Segal (1983, en Candell & Drasgow, 1988) propuso un procedimiento iterativo que va ajustando los valores de m y k obviando la re-estimación de θ; este consiste en realizar una equiparación con base en estimaciones iniciales de los parámetros para los dos grupos separadamente, identificar los ítems con DIF, hacer una nueva equiparación excluyéndolos, y recalcular los índices de DIF; estos dos últimos pasos se repiten hasta que en dos iteraciones consecutivas se identifiquen los mismos ítems con DIF. Drasgow, (1987) utilizó este procedimiento con el método de la función de respuesta al ítem para el cálculo de las constantes de equiparación y un año más tarde Candell & Drasgow (1988) demostraron su efectividad utilizando dos estrategias para calcular m y k: media –desviación típica robusta y método de función de respuesta al ítem.

Por su parte, Park & Lautenschlager (1990) compararon la efectividad de la propuesta de Segal (1983) con una modificación de la de Lord y propusieron una combinación de las dos. Esta se inicia con el ajuste de las constantes de equiparación siguiendo el procedimiento iterativo de Segal (1983) y el método de cálculo del χ^2 mínimo, una vez se logra el criterio de convergencia se re-estima θ para los dos grupos separadamente a partir de los ítems insesgados, entonces se re-estiman los parámetros de los ítems utilizando como valores fijos estas estimaciones de θ y finalmente, se identifican los ítems con DIF. De acuerdo con Lautenschlager, Flaherty & Park (1994) este procedimiento arroja resultados más satisfactorios que los dos anteriores; aunque Miller &

[13] En Lim & Drasgow (1990) se encuentra una explicación y comparación didáctica de estos algoritmos de estimación y posteriormente en este mismo capítulo se volverá sobre el tema

Oshima (1992) y Hidalgo Montesinos & López Pina (2002) lo consideran demasiado costoso en términos de cómputo. En consecuencia Miller & Oshima (1992) propusieron una modificación que consiste en un procedimiento de dos etapas que obvia el proceso iterativo inicial para calcular las constantes de equiparación pero calcula dichas constantes y estima los parámetros de los ítems dos veces. Hidalgo Montesinos & López Pina (2002) sintetizaron esta propuesta en un procedimiento de dos etapas con una sola estimación de los parámetros de los ítems. En la primera etapa se estiman estos parámetros para los dos grupos separadamente, se equiparan las métricas y se identifican y eliminan los ítems con DIF; en la segunda etapa se realiza la equiparación con los ítems libres de DIF y se obtiene el estadístico para detectar DIF para todos los ítems. Mediante un estudio con datos simulados, los autores mostraron que este procedimiento arroja resultados satisfactorios utilizando el χ^2 de Lord (1980) y las medidas de áreas de Raju (1988); Raju (1990) para la detección del DIF.

Sea cual sea el procedimiento de purificación elegido, una estrategia frecuentemente utilizada en los estudios de DIF es calcular las constantes de equiparación tomando como subprueba de anclaje todos los ítems de la prueba con excepción del que está siendo estudiado, sin embargo recientemente Wang & Yeh (2003) y Wang (2004) mostraron que este procedimiento arroja resultados satisfactorios solamente cuando ningún ítem de la prueba presenta DIF o si el ítem analizado es el único que lo presenta o, si hay varios ítems con DIF pero unos favorecen a un grupo y otros al otro grupo de manera que se presente el fenómeno de cancelación previamente descrito. Dado que estas condiciones son difíciles de satisfacer en la práctica o, si se cumplen, hacen innecesario el estudio, resulta más recomendable usar unos ítems determinados como subprueba de anclaje para el análisis de todos los demás. Esto implica decidir el mecanismo apropiado para la selección de dichos ítems y el número de ítems necesarios para obtener cálculo preciso de las constantes de equiparación. La mejor estrategia para la selección de los ítems de anclaje sería la identificación de algunos previamente calibrados y con evidencia de que no tienen DIF, sin embargo, esto sólo es posible si se dispone de bancos suficientemente amplios y analizados, condición que no se cumple en muchas situaciones reales (Fidalgo, 1996a, Navas-Ara, 1996). Algunos autores (Thissen, Steinberg & Wainer, 1988; Wang & Yeh, 2003) han sugerido utilizar un procedimiento menos costoso como el MH, para identificar algunos ítems insesgados que puedan utilizarse como subprueba de anclaje para la equiparación, y Wang, (2004) propuso un procedimiento iterativo para la selección de los ítems de anclaje, sin embargo, este último no ha sido aún evaluado. De otra parte, dos trabajos recientes se han ocupado del examinar el efecto del número de ítems de anclaje sobre la precisión de la equiparación: Según Kaskowitz & De Ayala (2001) 5 ítems es muy poco y recomiendan utilizar 15 o 25 para obtener resultados satisfactorios, sin embargo, Wang & Yeh (2003) sugieren que 4 ítems si realmente están libres de DIF, pueden ser suficientes para obtener una precisión adecuada.

Comparación de modelos

La primera referencia que generalmente se cita sobre la propuesta de comparación de modelos TRI para la detección de DIF es la aplicación de Thissen, Steinberg & Gerrard (1986). Utilizando

una prueba de "culpa sexual"[14] aplicada a grupos de hombres y mujeres, los autores ilustraron un procedimiento para probar la hipótesis de ausencia de DIF para uno o varios ítems simultáneamente, evaluando el ajuste de modelos TRI para los dos grupos de interés, mediante el estadístico G^2 de Bishop, Fienberg & Holland (1975). Posteriormente Thissen, Steinberg & Wainer (1988) y Thissen, Steinberg & Wainer (1993) formalizaron e ilustraron más ampliamente el procedimiento. Brevemente, este consiste en la comparación del ajuste de dos modelos, uno de los cuales, generalmente denominado modelo compacto (C), supone que los parámetros son iguales para los dos grupos, y otro, el modelo aumentado (A), supone que tales parámetros son diferentes. Los parámetros se estiman mediante MVM de Bock & Aitkin (1981) y la significación de las diferencias observadas entre los modelos se evalúa mediante el estadístico de razón de verosimilitud. Thissen, Steinberg & Wainer (1988, p. 153) describen el procedimiento en tres pasos:

1. Ajustar un modelo TRI (modelo A) para los dos grupos simultáneamente con uno o unos pocos ítems 'de anclaje', a los cuales se le impone la restricción de que los parámetros sean iguales para los dos grupos; esta restricción no se impone al ítem o ítems estudiados. Una vez estimados los parámetros se calcula el estadístico $G_A^2 = -2\,log\,L(A)$, donde $L(A)$ es la función de verosimilitud para las estimaciones de los parámetros del modelo.

2. Ajustar nuevamente el modelo (modelo C) imponiendo la restricción de igualdad de parámetros para el o los ítems estudiados y calcular $G_C^2 = -2\,log\,L(C)$.

3. Calcular el estadístico $G^2 = G_C^2 - G_A^2$ para probar la hipótesis de igualdad de los modelos o ausencia de DIF; éste sigue una distribución con tantos grados de libertad como la diferencia en el número de parámetros de los dos modelos. El rechazo de la hipótesis de igualdad de los modelos, conduce a la conclusión de que el ítem o ítems analizados presentan DIF.

Otras propuestas de procedimientos que siguen el mismo razonamiento son la de Muthén & Lehman (1985), la de Kelderman (1989) y la de Bock, Muraki & Pfeiffenberger (1988). La primera de ellas estima los parámetros mediante mínimos cuadrados generalizados para modelos TRI basados en la distribución normal acumulada, y posteriormente evalúa la significación de las diferencias entre los modelos mediante la razón de verosimilitud ya mencionada. Kelderman (1989) por su parte propuso un procedimiento para comparar modelos loglineales de un parámetro estimando los parámetros por máxima verosimilitud y sometiendo a prueba la hipótesis de no DIF mediante el mismo estadístico G^2. Finalmente, Bock, Muraki & Pfeiffenberger (1988) propusieron estimar los

[14] "…Mosher Forced Choice Sex Guilt Inventory (MFCSGI; Mosher (1966); Mosher (1968)) a 72-ítem questionnaire intended to measure a construct called "sex guilt"". (Thissen, Steinberg & Gerrard, 1986, p. 119)

parámetros por máxima verosimilitud marginal y evaluar las diferencias observadas utilizando el error estándar de dichas estimaciones. Fidalgo, (1996a) incluye en su texto una ilustración detallada del procedimiento de Thissen, Steinberg & Gerrard (1986) y en Thissen, Steinberg & Wainer (1993) se encuentran ejemplos numéricos de cada uno de los cuatro procedimientos.

Aunque estos procedimientos gozan de un adecuado sustento matemático y permiten evaluar una prueba completa o un grupo de ítems de manera simultánea, no se han generalizado en la práctica. Según Gómez Benito & Hidalgo Montesinos (1997b) las dos grandes desventajas del procedimiento de Thissen, Steinberg & Gerrard (1986) son su incapacidad para detectar diferencias pequeñas y su dependencia del tamaño de los grupos, mientras que el de Kelderman (1989) presenta dificultades para análisis simultáneos de pruebas largas y requiere demasiadas iteraciones para alcanzar un criterio de convergencia. Presumiblemente una razón válida para el escaso uso actual de todos los procedimientos descritos es su alto costo computacional en comparación con otros procedimientos disponibles hoy, incluyendo los demás basados en la TRI. De acuerdo con Kim & Cohen (1995) utilizando el procedimiento de Thissen, Steinberg & Gerrard (1986) y Thissen, Steinberg & Wainer (1988, 1993) iterativamente para purificar la medida de igualación de los grupos, encontraron que se requieren $3k+2ni-i^2$ calibraciones para el análisis de DIF de los ítems, donde k es el número de iteraciones necesarias y i es el número de ítems; así, con 3 iteraciones y 15 ítems (prueba más corta de lo habitual), el número de calibraciones sería 90. Además, aún si no se utiliza este procedimiento iterativo y se recurre al mecanismo de purificación mediante el χ^2_{MH}, (Thissen, Steinberg & Wainer, 1988), o si no se utiliza purificación alguna, puesto que requiere múltiples calibraciones, el procedimiento resulta más dispendioso que otros basados en TRI y definitivamente mucho más que la RL o el MH.

Comparación de parámetros: χ^2 de lord

La propuesta pionera en las técnicas de detección de DIF con base en modelos TRI es la de Lord, (1980) presentada en su libro sobre aplicaciones de los modelos TRI, en la cual toma parte de un trabajo suyo, Lord (1977), realizado con base en las respuestas de 2250 blancos y 2250 negros en un grupo de ítems de aptitud verbal del Scholastic Aptitud Test (*SAT*), y que había sido publicado en el texto de Ype. H. Poortinga tres años antes. A diferencia de los procedimientos expuestos en el apartado anterior, la técnica de Lord (1980) compara los vectores de parámetros estimados cuando se ajustan modelos TRI para los dos grupos separadamente. Brevemente el procedimiento consiste en estimar los parámetros de los ítems ajustando modelos de 1, 2 o 3 parámetros para los grupos de interés, poner las estimaciones en la misma métrica y finalmente someter a prueba estadística las diferencias observadas entre los vectores de estimaciones de los parámetros.

Si X_f y X_r son los vectores de parámetros para los grupos focal y de referencia, respectivamente, la hipótesis de ausencia de DIF, que se somete a prueba, puede expresarse como

$H_0 : X_f = X_r$. Ahora, si $\hat{X}_f$ y $\hat{X}_r$ son los vectores de las estimaciones de los mismos parámetros, las diferencias entre dichas estimaciones para los dos grupos pueden expresarse en un vector $V = \hat{X}_r - \hat{X}_f$ que tiene dimensión $p\ X\ 1$, donde p es el número de parámetros en el modelo; si se han ajustado modelos de tres parámetros $V = \begin{bmatrix} \hat{a}_{ir} \\ \hat{b}_{ir} \\ \hat{c}_{ir} \end{bmatrix}_r - \begin{bmatrix} \hat{a}_{if} \\ \hat{b}_{if} \\ \hat{c}_{if} \end{bmatrix} = \begin{bmatrix} \hat{a}_{ir} - \hat{a}_{if} \\ \hat{b}_{ir} - \hat{b}_{if} \\ \hat{c}_{ir} - \hat{c}_{if} \end{bmatrix}$, si se usan modelos TRI de dos parámetros $V = \begin{bmatrix} \hat{a}_{ir} \\ \hat{b}_{ir} \end{bmatrix} - \begin{bmatrix} \hat{a}_{if} \\ \hat{b}_{if} \end{bmatrix} = \begin{bmatrix} \hat{a}_{ir} - \hat{a}_{if} \\ \hat{b}_{ir} - \hat{b}_{if} \end{bmatrix}$, y finalmente, si se ajusta modelos Rasch, $V = \hat{b}_{ir} - \hat{b}_{if}$. El estadístico para someter a prueba la significación de las diferencias así observadas, conocido como el χ^2 de Lord, es $\chi^2 = V' \sum^{-1} V$, donde $\sum$ es la matriz de varianzas y covarianzas de las diferencias entre los parámetros. Esta matriz se estima como la suma de las matrices de varianzas y covarianzas de las estimaciones de los parámetros para los dos grupos; esto es: $S = S_r + S_f$, donde S_r y S_f son las matrices de varianzas y covarianzas de las estimaciones de los parámetros para el grupo de referencia y focal, respectivamente. Como es bien sabido, estas matrices son cuadradas y en este caso tendrán dimensión $p\ X\ p$. En estos términos el estadístico de prueba sigue una distribución χ^2 con p grados de libertad, tantos como parámetros comparados, y queda expresado como $\chi^2 = V' S^{-1} V$

Lord (1980, p. 217) advirtió que cuando se ajustan modelos de tres parámetros y los ítems difieren en el parámetro de discriminación, los más discriminativos mostrarán mayores diferencias entre los grupos comparados que los menos discriminativos; considerando además que el DIF se expresa a través de las diferencias entre la dificultad y la discriminación, el uso del parámetro de pseudoazar no parece relevante y en cambio presenta dificultades en su estimación. De esta manera el autor sugiere un procedimiento de tres pasos cuando se utilicen modelos de tres parámetros, así:

1. Ajustar un único modelo de tres parámetros tomando conjuntamente los grupos que se van a comparar, para obtener una estimación conjunta de c_i, para los dos grupos ($\hat{c}_{if} = \hat{c}_{ir} = \hat{c}_i$)

2. Ajustar modelos de tres parámetros separadamente para los dos grupos, fijando como valor de c_i, la estimación, $\hat{c}_i$, obtenida en el paso anterior. Estos modelos suelen notarse como 3P-c y no son modelos de dos parámetros puesto que incluyen el parámetro de pseudoazar, sin embargo, éste se considera constante para todos los grupos de interés.

3. Someter a prueba, mediante el estadístico χ^2 descrito antes, las diferencias observadas en las estimaciones de los parámetros de dificultad y discriminación, obtenidas en el paso anterior.

En todos los casos Lord (1980) recomienda que las estimaciones se realicen estandarizando las del parámetro de dificultad, en una escala con media 0 y desviación estándar 1, con el fin de que las estimaciones de todos los parámetros en todos los grupos queden expresadas en la misma escala. Además, para el cálculo de la media y desviación estándar que se utilizan para dicha estandarización, puede ser conveniente omitir los ítems con dificultad extrema (muy fáciles o muy difíciles) ya que las estimaciones de este parámetro para este tipo de ítem suelen tener error grande; en todo caso, estos ítems se omiten únicamente para el cálculo de dichos descriptivos y reciben el mismo tratamiento que los otros ítems, para todos los demás efectos.

Además de la limitación del procedimiento cuando se ajustan modelos de tres parámetros, los supuestos que lo sustentan han impuesto otras limitaciones que pueden ser serias en sus aplicaciones prácticas. En primer lugar, la prueba χ^2 de Lord es asintótica y en consecuencia hay una exigencia de tamaño de muestra para alcanzar la distribución esperada, exigencia que puede ser difícil de satisfacer cuando se trabaja con grupos minoritarios; de otra parte, supone que el parámetro de magnitud de atributo de los individuos (θ) es conocido, supuesto imposible de cumplir en los estudios aplicados; y finalmente, es aplicable únicamente cuando se utilizan algoritmos de estimación por máxima verosimilitud. Estas limitaciones, el desconocimiento de una tamaño de muestra mínimo para lograr la convergencia a la distribución χ^2 y algunos reportes como el de Linn, Levine, Hastings & Wardrop (1981) según el cual podía presentar una alta tasa de falsos positivos (FP) en comparación con medidas de área, hicieron que el procedimiento adquiriera mala prensa. De hecho, Camilli & Shepard (1994), en uno de los textos más consultados sobre métodos para la detección de DIF, no recomiendan su uso; otros autores como Hidalgo Montesinos & López Pina (1997) y López Pina, Hidalgo & Sánchez Meca (1993) no lo prefieren comparado con procedimientos como la RL y otros como Fidalgo, (1996a) y Millsap & Everson (1993) sugieren utilizarlo en combinación con otros procedimientos, sobretodo si resulta significativo.

Sin embargo, buena parte de los trabajos posteriores a la publicación de la propuesta de Lord (1980) se han dedicado al examen del funcionamiento del estadístico en condiciones diferentes a las supuestas por él. Con respecto al algoritmo de estimación de los parámetros y el supuesto θ conocido, los estudios han sido bastante consistentes en mostrar que las estimaciones por MVC que hace tal supuesto, no resultan recomendables; en cambio las estimaciones por MVM de Bock & Aitkin (1981) y MB de Mislevy, (1986) y Mislevy & Bock (1990) pueden mejorar sustancialmente la precisión de las estimaciones y por tanto, el funcionamiento del χ^2 de Lord. McLaughlin & Drasgow (1987) mostraron que cuando los parámetros de los ítems y de los individuos son todos desconocidos y se estiman mediante MVC, el χ^2 de Lord presenta una importante inflación del error tipo I, por encima de los valores nominales; Kim & Cohen (1994) replicaron este estudio utilizando estimaciones MVM y MB con modelos de 2 y 3 parámetros y encontraron que el error tipo I se mantuvo satisfactoriamente controlado con modelos de 2 parámetros y con 3 parámetros si se

fija un valor para c (3P-c). Los mismos autores (Cohen & Kim, 1993) habían encontrado además, que los resultados de los análisis tanto en la recuperación de los parámetros[15] como las tasas de falsos positivos y falsos negativos fueron más satisfactorios cuando se utilizó algoritmo de estimación bayesiana. Por su parte, Lim & Drasgow (1990) compararon las estimaciones por MVM y MB con modelos de dos parámetros y con ítems uni y multidimensionales; ambos procedimientos de estimación arrojaron resultados satisfactorios con ítems unidimensionales y con tamaños de grupos de 750 examinados aunque observaron una sobreestimación del parámetro a para ítems altamente discriminativos y con valores de b extremos; además, encontraron cierta superioridad de las estimaciones bayesianas sobre las de MVM con grupos pequeños (250 examinados por grupo). En cuanto al efecto del tamaño de muestra sobre el funcionamiento del χ^2 de Lord para la detección de DIF, en el mismo estudio anterior de Lim & Drasgow (1990) advirtieron que tanto las estimaciones por MVM como la MB son sensitivas al tamaño de muestra y ese efecto es mayor para las estimaciones por máxima verosimilitud. Además, con base en un estudio de Monte Carlo simulando datos con modelos de 2 parámetros Cohen & Kim (1993) recomendaron su uso sobre dos procedimientos de medidas de área entre las CCI, sobretodo cuando se trabaja con muestras pequeñas (100 examinados por grupo), pruebas cortas (hasta 20 ítems), diferentes distribuciones de la magnitud de atributo o altos porcentajes de ítems con DIF (hasta 20%). En el capítulo 6 se tratarán con más detalle los hallazgos referentes al efecto de los tamaños de grupos sobre el funcionamiento del estadístico.

Otro aspecto que se ha mencionado como una debilidad del procedimiento es la estimación de la matriz de varianzas y covarianzas, necesaria para el cálculo del estadístico de prueba. Según Thissen & Wainer (1982) la falta de precisión en la estimación de la matriz de varianzas y covarianzas de las estimaciones de los parámetros de los ítems puede aumentar el error en la detección de ítems con DIF; sin embargo, Kim & Cohen (1995) comparando el χ^2 de Lord con el K^2 de Pearson[16] encontraron resultados muy similares entre los dos estadísticos y entre éste último y dos medidas de área, a partir de estos resultados ellos concluyeron que "la carencia de las covarianzas fuera de la diagonal de matriz de varianzas y covarianzas, no parece afectar de manera

---

[15] La recuperación de los parámetros se evaluó mediante los cuadrados medios del error de las estimaciones con respecto al valor del parámetro y mediante las correlaciones entre éstos y sus estimaciones en las diferentes réplicas. Este procedimiento se describe más en detalle en la presentación de los estudios empíricos, en el capítulo 4.

[16] La diferencia entre estos dos estadísticos es que el K^2 de Pearson supone independencia entre las diferencias de las estimaciones de los parámetros y entonces la matriz de varianzas y covarianzas incluye 0 en todas las celdas fuera de la diagonal, mientras que en el χ^2 estima la matriz a partir de las matrices de varianzas y covarianzas de las estimaciones de los parámetros para los dos grupos, como se describió antes.

importante el acuerdo entre esas medidas en la detección de DIF" (p. 309), además, hallaron tasas de error dentro de los límites nominales.

Finalmente, debe anotarse que comparado con otros procedimientos basados en la TRI, el χ^2 es un procedimiento relativamente fácil de utilizar, cuenta con una prueba de significación para la identificación de los ítems con DIF, y algunos estudios apoyan su uso en la práctica. Con datos simulados utilizando modelos TRI de dos parámetros McCauley & Mendoza (1985) encontraron que el χ^2 de Lord resultó más efectivo que otros índices basados en la TRI, incluyendo los de medidas de área, para identificar ítems con DIF cuando éste se ha simulado generando los datos con base en modelos multidimensionales. Los resultados del estudio de Kim & Cohen (1995) con respuestas a una prueba de matemáticas y ajustando modelos de dos parámetros también apoyan su uso sobretodo cuando se compara en términos de costo computacional con la razón de verosimilitud de Thissen, Steinberg & Wainer (1988, 1993). Evaluando la efectividad del proceso de purificación de dos etapas sobre la efectividad de métodos basados en la TRI, Núñez Núñez, Hidalgo Montesinos & López Pina (2000) encontraron resultados satisfactorios para el χ^2 de Lord, en comparación con medidas de área. Hidalgo Montesinos & López Pina (2002) aplicaron un procedimiento de purificación de dos etapas con modelos de respuesta graduada utilizando una medida de área y el χ^2 de Lord y no encontraron una superioridad notoria de alguno de los dos procedimientos, aunque hacen notar que el último es un estadístico más conservador que la medida de área (p. 43). Nótese, sin embargo, que en estos estudios se obtuvieron estimaciones de máxima verosimilitud marginal y Cohen & Kim (1995) utilizaron también estimaciones bayesianas; además, en ningún caso se ajustaron modelos TRI de tres parámetros.

Medidas de área entre las CCI

Aunque las medidas de área parten de la misma idea básica común a todos los procedimientos basados en la TRI, difieren en que no comparan los modelos o valores de los parámetros sino que evalúan de alguna manera el área comprendida entre las dos CCI del ítem ajustadas para los dos grupos independientemente y expresadas en la misma métrica, desde esta perspectiva un ítem estará libre de DIF si dicha área es nula. La primera propuesta publicada para medir el área que separa a dos CCI fue la de Rudner (1977) y Rudner, Getson & Knight (1980a, 1980b), que se conoce hoy como una aproximación discreta para un intervalo finito de θ, y aparece ilustrada en la figura 13. Después de ajustar los modelos y equiparar las dos CCI con valores de θ en un intervalo fijo ($-3 \leq \theta \leq 3$), esta escala se divide en pequeños intervalos ($\Delta\theta$) y se calcula el área del rectángulo que tiene por un lado, el valor absoluto de la diferencia de probabilidades entre el grupo de referencia y el focal dado un valor θ_j ($\left| P_R(\theta = \theta_j) - P_F(\theta = \theta_j) \right|$), y por el otro lado $\Delta\theta$; el

índice de discrepancia es entonces $A = \sum_{\forall \theta_j} \left| P_R(\theta = \theta_j) - P_F(\theta = \theta_j) \right| \Delta\theta$, la suma de las áreas para todo el rango de valores de θ. En la figura 13 se ilustra el cálculo de este índice para un ítem con parámetros $a_r = 1.7,\ b_r = -.65,\ a_f = 1,\ b_r = .2$ y $c_r = c_f = 0$ haciendo $\Delta\theta = .1$, valor útil para efectos ilustrativos pero demasiado grandes para un estudio aplicado puesto que la precisión del índice depende de la amplitud del incremento.

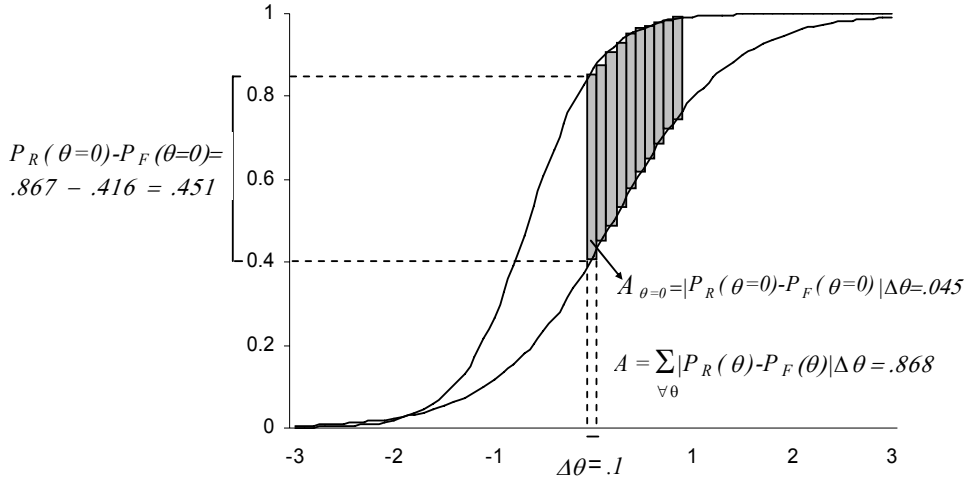


Figura 13: Ilustración de la medida de área sin signo propuesta por Rudner, (1977) y Rudner, Getson & Knight (1980a, 1980b) con $\Delta\theta = .1$

Algunas de las críticas a este índice sencillo, muy intuitivo y fácil de calcular, fueron la ausencia de un estadístico de prueba o un punto de corte que permitiera clasificar los ítems en sesgados e insesgados, su posible imprecisión y dar el mismo peso a las áreas en todos los puntos de la escala de θ. Varias propuestas posteriores han intentado superar estas limitaciones, algunas de ellas son los cuatro índices de Linn, Levine, Hastings & Wardrop (1981), los índices de diferencia de probabilidad de Linn & Harnisch (1981), la suma de cuadrados autoponderada de Shepard, Camilli & Williams (1984) o la medida de áreas con signo de Camilli & Shepard (1994). Aunque algunas de estas logran ponderar las diferencias teniendo en cuenta el número de examinados en cada intervalo y en consecuencia, mejorarían la precisión del cálculo, Shepard, Camilli & Williams (1985) no encontraron diferencias importantes entre las medidas ponderadas y las no ponderadas; además, todas las propuestas siguen adoleciendo de un valor crítico para identificar la discrepancia máxima atribuible al azar y son medidas aproximadas puesto que calculan áreas discretas para evaluar un área que es de naturaleza continua.

Desde una perspectiva matemática diferente Raju, (1988) propuso unas medidas exactas que toman en consideración la naturaleza continua de la escala de θ a partir de la definición misma de este tipo de áreas: integrando entre $-\infty$ e ∞ con respecto a θ. Si $F_F(\theta)$ y $F_R(\theta)$ son las funciones de respuesta al ítem para los grupos focal y de referencia respectivamente, el área exacta

signada (ESA, por las siglas en inglés) y el área no signada (EUA)[17] entre las respectivas CCI son $ESA = \int_{-\infty}^{+\infty} (F_R(\theta) - F_F(\theta))d(\theta)$ y $EUA = \int_{-\infty}^{+\infty} |F_R(\theta) - F_F(\theta)|d(\theta)$, respectivamente. Con estos puntos de partida Raju, (1988) derivó las fórmulas para calcular las medidas de área signadas y no signadas para modelos de 1, 2 y 3 parámetros, sin embargo cuando se ajustan modelos 3P, el cálculo del área es posible siempre y cuando el parámetro c sea igual para los dos grupos, si esta condición no se cumple, el área es infinita (p. 499). Con esta restricción, las medidas de área signada y no signada entre CCI para modelos de tres parámetros, son $ESA = (1-c)(b_r - b_f)$ y $EUA = (1-c)\left|\frac{2(a_r - a_f)}{da_r a_f} ln\left(1 + e^{\frac{da_r a_f (b_r - b_f)}{a_r - a_f}}\right) - (b_r - b_f)\right|$, respectivamente. Si además el ítem es igualmente discriminativo para los dos grupos ($a_r = a_f = a$), ésta última expresión se reduce a $EUA = (1-c)|b_r - b_f|$. Las expresiones para modelos de dos parámetros ignoran los términos que dependen de c, y para modelos de Rasch se reducen a la diferencia entre la dificultad: $ESA = b_r - b_f$ y $EUA = |b_r - b_f|$.

Aunque el mismo autor hizo notar que las diferencias entre las CCI son más importantes en los valores de θ con mayor número de examinados y en consecuencia puede ser apropiado calcularlas para intervalos cerrados, mostró que la diferencia entre las dos aproximaciones puede ser numéricamente importante y que algunos ítems detectados con las medidas de área exacta pueden parecer no sesgados con intervalo cerrado (ver ejemplo en Raju, 1980, p. 501). Dos años más tarde el mismo autor (Raju, 1990) describió las distribuciones muestrales de sus medidas de área y propuso los estadísticos de prueba para ESA y EUA, conocidos como $Z(ESA)$ y $Z(H)$, respectivamente, asintóticamente normales. La primera está definida para modelos de Rasch y modelos 2P y 3P siempre y cuando tengan la misma discriminación para los dos grupos[18], y tiene la forma general $Z(ESA) = \frac{\hat{b}_r - \hat{b}_f}{\sqrt{Var(ESA)}}$, donde la varianza se calcula dependiendo del tipo área (signada o no signada) y de modelo específico. La segunda prueba está definida para modelos de dos y

[17] A costa de la pureza en la escritura, en este documento se utilizarán las siglas en inglés ya que su uso está bastante generalizado y resultan más fáciles de leer que las respectivas siglas en castellano.

[18] Si los parámetros de discriminación son iguales para los dos grupos ($a_r = a_f$) las CCI no se cruzan, mientras que si $a_r \neq a_f$ estas se cruzan en algún punto de la escala de θ

tres parámetros con diferente discriminación y tiene la forma $Z(H) = \frac{H}{\sqrt{Var(H)}}$, donde H y su varianza se calculan dependiendo del tipo de modelo. Las expresiones para calcular tales valores pueden consultarse en Raju, (1990) o en Fidalgo, (1996a).

Partiendo del mismo punto pero integrando para un intervalo cerrado entre dos puntos θ_1 y θ_2, Kim & Cohen (1991) derivaron las expresiones para calcular las medidas de área entre las CCI, que se conocen como medidas de área cerradas signadas (CSA, por sus siglas en inglés) y no signada (CUA) para modelos de 1P, 2P y 3P. En este caso no existe la limitación de las áreas de Raju en los modelos de 3P; puesto que se limita el intervalo de θ, el área entre las CCI es finita aún cuando los modelos sean de tres parámetros con diferente valor de *c*. Para modelos de tres parámetros el área signada es $CSA = (c_r - c_f)(\theta_2 - \theta_1) + ln\left(\frac{\left(1 + e^{da_r(\theta_2 - b_r)}\right)^{\frac{1-c_r}{da_r}}\left(1 + e^{da_f(\theta_1 - b_f)}\right)^{\frac{1-c_f}{da_f}}}{\left(1 + e^{da_r(\theta_1 - b_r)}\right)^{\frac{1-c_r}{da_r}}\left(1 + e^{da_f(\theta_2 - b_f)}\right)^{\frac{1-c_f}{da_f}}}\right)$ y tiene la misma forma para modelos de 2P y 1P pero se va sintetizando a medida que se excluyen términos ya sea por el número de parámetros del modelo o porque alguno de éstos se suponen iguales para los dos grupos. Para modelos de Rasch, modelos de 2P con igual discriminación y modelos de 3P con igual discriminación y pseudoazar, el área no signada es sencillamente el valor absoluto del área signada respectiva, es decir $CUA = |CSA|$. Sin embargo, cuando estas condiciones no se cumplen y las dos CCI se cruzan en un punto θ_x, el cálculo del área no signada depende de la localización de dicho punto: Si este valor se encuentra fuera del intervalo ($\theta_x < \theta_1$ o $\theta_x > \theta_2$) el área no signada sigue siendo el valor absoluto de la signada, pero si las CCI se cruzan en un punto que cae dentro del intervalo de interés ($\theta_1 < \theta_x < \theta_2$), el cálculo del área debe considerar dicho valor. Las expresiones para obtener el valor θ_x y para el cálculo de todas estas medidas de área se encuentran en Kim & Cohen (1991).

En el mismo estudio ya citado, utilizando una prueba verbal de 50 ítems, en 10 de los cuales el DIF se indujo experimentalmente, y con grupos iguales de 1000 examinados cada uno, Kim & Cohen, (1991) encontraron resultados muy similares entre las medidas de área exactas y cerradas con modelos de 3P y 3P-c; además sus hallazgos mostraron que tanto las medidas de área cerradas tanto signadas como sin signo, funcionaron satisfactoriamente con ambos tipos de modelos, este resultado apoyaría su uso en comparación con las medidas exactas puesto que no hacen la restricción de igualdad del parámetro *c*, necesaria para el cálculo de éstas últimas. Posteriormente los mismos autores, Cohen, Kim & Baker (1993) derivaron las expresiones para calcular las medidas de área entre las CCI para modelos de respuesta graduada; ésta últimas sin embargo, han sido menos evaluadas.

A pesar de su sencillez, una revisión de los estudios publicados indican que actualmente las medidas de área parecen ser más utilizadas para simular el DIF que para detectarlo; además, algunos resultados de estudios comparativos no favorecen su uso. De acuerdo con los resultados de Cohen y Kim (1993), aunque las diferencias no son demasiado grandes, tanto el *Z(ESA)* como el *Z(H)* mostraron resultados más pobres que el χ^2 de Lord. Ajustando modelos de 1p Gómez Benito & Navas-Ara (2000) evaluaron el funcionamiento de las medidas de área exactas utilizando el estadístico de prueba (Raju, 1990) y un punto de corte calculado mediante el procedimiento de línea de base (Rogers & Hambleton, 1989; Hambleton & Rogers, 1989), tanto la tasa de detección en cuatro réplicas como el número de falsos positivos fueron muy similares para los dos procedimientos pero no alcanzaron los niveles de la RL o del MH. De otra parte, en un estudio ya citado sobre la efectividad del proceso de purificación de dos etapas, Núñez Núñez, Hidalgo Montesinos & López Pina (2000) encontraron resultados menos satisfactorios para las medidas de área en comparación con el χ^2 de Lord; sin embargo, cuando compararon las medidas de área para modelos de respuesta graduada (Cohen, Kim & Baker, 1993), Hidalgo Montesinos & López Pina (2002) encontraron resultados muy similares a los arrojados por el χ^2 de Lord.

El estudio de Kim & Cohen (1995) mostró que existe un buen acuerdo entre los tres procedimientos basados en la TRI (razón de verosimilitud, χ^2 de Lord y medidas de área *Z(ESA)* y *Z(H)* en la detección de DIF inducido por el uso de calculadora en una prueba de matemáticas. Este acuerdo y las similitudes entre los resultados arrojados por los tres procedimientos mejora cuando se utilizan procedimientos iterativos; a partir de estos resultados los autores recomiendan usar algún procedimientos de purificación de la escala y no utilizar una técnica sola en las aplicaciones prácticas, sino usar una combinación de ellas. De acuerdo con (Hambleton, Clauser Mazor y Jones, 1993) y Gómez Benito & Hidalgo Montesinos (1997b) los métodos basados en la TRI que estiman un nivel de habilidad o atributo latente gozan hoy de mucha aceptación, entre sus bondades se destacan generalmente su solidez matemática y su capacidad para detectar DIF uniforme y no uniforme. Sin embargo, la mayor limitación en las aplicaciones reales es su exigencia en el tamaño de muestra necesario para obtener estimaciones adecuadas de los parámetros de los modelos y el cálculo de las constantes de equiparación.

SEGUNDA PARTE:

ESTUDIOS EMPIRICOS

Capítulo 4:
EL EFECTO DEL TAMAÑO DE MUESTRA Y LA RAZÓN DE TAMAÑOS SOBRE EL MANTEL-HAENSZEL

Tanto los estudios teóricos como los experimentos con datos simulados han documentado suficientemente el efecto del tamaño de muestra sobre el poder y el error tipo I de muchos estadísticos, entre ellos el MH. De acuerdo con la revisión de Hills (1990), el MH es una opción adecuada cuando el tamaño de los grupos está entre 100 y 300; para Zieky, (1993) se requieren por lo menos 100 examinados en el grupo más pequeño y 500 en total; mientras que Mazor, Clauser & Hambleton (1992) y Fidalgo, Mellenbergh & Muñiz (1999) cuestionan su uso con muestras muy pequeñas. En un estudio con datos simulados, los primeros compararon 5 diferentes tamaños de muestra (2000, 1000, 500, 200 y 100 por grupo), y encontraron que un tamaño de 200 sólo es recomendable cuando interesa identificar únicamente los ítems con amplia magnitud de DIF; si se requiere mayor poder el tamaño de muestra debe incrementarse de manera importante sobre todo cuando hay impacto y, aún con muestras de 2000 examinados por grupo, los ítems más difíciles, con baja discriminación o con baja magnitud de DIF pueden no ser identificados por el MH. En este trabajo solo se evaluó DIF uniforme y los tamaños de los dos grupos fueron iguales en todas las condiciones experimentales. En una dirección similar pero evaluando solamente dos tamaños de muestra (100 y 200 por grupo), Fidalgo, Mellenbergh & Muñiz (1999) encontraron que el χ^2_{MH} es más sensitivo al tamaño de muestra que el $\hat{\alpha}_{MH}$ y recomendaron utilizar ambos estadísticos, dando prioridad al segundo, cuando se tengan muestran tan pequeñas como 200 examinados por grupo. Nuevamente en este estudio se trabajó con grupos de igual tamaño. En un estudio más reciente Fidalgo, Ferreres & Muñiz (2004) encontraron que con grupos de referencia entre 50 y 200 examinados y grupos focales entre 50 y 100, el poder del MH es muy bajo y solamente mejora si se utiliza un nivel de significación del 20% en vez del tradicional 5%.

Con tamaños de muestra grandes (500, 1000 y 3000 por grupo) Roussos & Stout (1996) compararon el comportamiento del error tipo I del MH y SIBTEST de Shealy & Stout (1993a), bajo diferentes valores de los parámetros de los ítems. Ellos encontraron que cuando no hay diferencia en la distribución de magnitud de atributo entre los grupos (impacto), ambos estadísticos adhieren bien al nivel de significancia nominal y no observaron grandes diferencias entre ellos; sin embargo, cuando hay impacto ambos procedimientos experimentan un aumento en su error tipo I con el tamaño de muestra; en particular, si los ítems tienen alta discriminación y baja dificultad. De otra parte, con muestras de 250 y 500 examinados por grupo, Rogers & Swaminathan (1993) evaluaron las propiedades distribucionales y el poder del MH y la RL para detectar uniforme y no uniforme. Aunque los factores que afectan el poder de los estadísticos no son los mismos cuando se trata de detectar DIF uniforme y no uniforme; como era de esperarse, el tamaño de muestra tuvo efecto importante sobre el poder del MH para detectar ambos tipos de DIF. Pero además, encon-

traron que el MH y RL mostraban poder similar para detectar DIF mixto, simulado variando tanto el parámetro de dificultad como el de discriminación. Aunque los dos trabajos reportaron resultados consistentes en cuando al efecto del tamaño de muestra, el primero solamente evaluó su efecto sobre el error tipo I y ambos utilizaron grupos de igual tamaño; en consecuencia sus hallazgos no son directamente generalizables a estudios étnicos en los cuales un grupo puede ser sensiblemente minoritario.

Miller & Oshima (1992) simularon las dos situaciones: grupos iguales (1000 examinados por grupo) y un grupo minoritario (1000 y 300 examinados) para explorar el funcionamiento de seis índices de DIF basados en la TRI y el MH, en función del tamaño de muestra, el número de ítems DIF y la magnitud del mismo. Los resultados mostraron que el poder de todos los estadísticos fue más alto con mayor tamaño de muestra y el número de falsos positivos de los seis índices TRI fue mayor cuando el tamaño de muestra fue menor, mientras que el MH mostró comportamiento más estable en los dos tamaños de muestra. Desde una aproximación similar, Narayanan & Swaminathan (1994) compararon el poder y error tipo I del MH y el SIBTEST para detectar DIF no uniforme bajo condiciones diferentes de tamaño de muestra, entre otros factores. Simulando nueve condiciones diferentes, resultantes de cruzar tres tamaños para el grupo de referencia (300, 500 y 1000) con tres para el grupo focal (100, 200 y 300), encontraron que el poder de ambos estadísticos se vio más afectado por el tamaño del grupo focal que por el de referencia. Dos años más tarde los mismos autores compararon el error tipo I y el poder del MH, la RL y el SIBTEST en la detección de DIF no uniforme manipulando 5 factores, entre ellos tamaño de muestra Narayanan & Swaminathan (1996). En este estudio analizaron cuatro tamaños cruzando dos para el grupo de referencia (500 y 1000) y dos para el grupo focal (200 y 500), además de observar el efecto del tamaño de muestra sobre el poder de los tres procedimientos, encontraron que la tasa de detección del DIF uniforme es similar para los tres procedimientos mientras que MH tiene poder muy inferior para detectar el no uniforme. A partir de los resultados de los dos últimos estudios, los autores concluyeron que es necesario adelantar trabajos con diferentes tamaños de muestra, tomando en consideración la razón de tamaños de los grupos de referencia y focal.

Esta última conclusión parece tener apoyo en uno de los resultados del trabajo de y Zwick, Thayer & Lewis (1999). Evaluando el procedimiento de estimación de Bayes empírico[19] para predecir la clasificación de los ítems, a partir de análisis futuros, en una de las categorías usadas por el ETS; ellos graficaron los valores del $MHD\text{-}DIF$ contra el estimador de Bayes empírico para 76 ítems de la subprueba verbal del *Graduate Record Examinations* (GRE). Los resultados mostraron que cuando los análisis se hicieron para hombre y mujeres, con tamaños de muestra de 4253 y 6491, respectivamente, los valores coincidían bastante ajustándose a una recta de 45°; sin em-

[19] En Zwick, Thayer & Lewis (1997, 1999) y Zwick & Thayer (2002) se pueden encontrar explicaciones detalladas de esta propuesta

bargo, cuando el análisis se realizó para 8736 blanco y 220 asiáticos, los resultados fueron menos estables y se observaron desviaciones importantes con respecto a ese patrón. Aunque su objetivo no era estudiar el efecto de las diferencias de tamaños de los grupos sobre el MH en las dos situaciones descritas, una posible hipótesis para explicar este resultado es el efecto de este factor.

A diferencia de lo ocurrido con el efecto del tamaño de muestra sobre el MH, el efecto de la razón de tamaños no ha sido estudiado de manera sistemática, lo cual limita la generalización de los hallazgos a situaciones en las cuales los grupos sean de tamaños similares entre sí o similares a las condiciones estudiadas. Este estudio tuvo como objetivo examinar el efecto del tamaño de muestra y de la razón de tamaños ($r = nr/nf$, donde nr es el tamaño del grupo de referencia y nf es el tamaño del grupo focal) sobre el error tipo I y el poder del M-H para detectar DIF uniforme, no uniforme y mixto en ítems dicótomos.

Método

Se adelantó un experimento de Monte Carlo en el que se manipularon los dos factores mencionados: tamaño de muestra del grupo de referencia (nr) y razón de tamaños de muestra de los grupos ($r = nr/nf$). Los tamaños del grupo de referencia fueron 500 y 1500; y el número de examinados en el grupo de referencia por cada uno del grupo focal fueron 1, 2, 2.5, 3, 4 y 5; así se obtuvieron 12 condiciones experimentales resultantes de cruzar estos dos factores, como se muestra en la tabla 4. Los tamaños del grupo de referencia se eligieron para representar dos situaciones diferentes entre sí y que pueden ser frecuentes en la práctica. Las razones de tamaños están entre 1, que puede encontrarse en estudios que comparan grupos según género, hasta situaciones en las que el grupo focal pertenece a una etnia o raza minoritaria (5 examinados del grupo mayoritario por cada uno del minoritario). También se consideraron el tipo de DIF (uniforme, no uniforme y mixto), la magnitud de DIF y los parámetros de los ítems, sin embargo, estos se fijaron dentro de cada condición experimental y los dos últimos se limitaron a valores frecuentes en los análisis aplicados. Además, se controlaron otros factores como la longitud de la prueba (una única prueba de 100 ítems) la distribución de magnitud de atributo en los dos grupos (no se simuló impacto) y el porcentaje de ítem DIF en la prueba (12%).

Generación de datos

Se simuló una única prueba unidimensional compuesta por 100 ítems ajustando modelos TRI de tres parámetros con igual probabilidad de acierto por azar ($c_i = 0,2 \quad \forall i$). Aunque la longitud de la prueba fue grande comparada con los instrumentos utilizados en la práctica, permitió controlar el porcentaje de ítem DIF dentro de la misma, que, según Clauser (1993); Narayanan & Swaminathan (1994), suele estar entre el 10% y el 15% en pruebas estandarizadas; además esta longitud de prueba permitió evaluar el comportamiento del error tipo I con diferentes tipos de ítems no

DIF. Los parámetros TRI de los ítems para el grupo de referencia se generaron siguiendo una distribución normal con media *0* y desviación estándar *1* para la dificultad ($b \approx n(0,1)$), una normal con media *0,5* y desviación estándar *0.2* para la discriminación ($a \approx n(0.5;0.2)$), y un valor constante, *0.2*, para el parámetro c ($c \approx u(0.2;0)$). En el anexo 1 se presentan los valores de los parámetros de los ítems no DIF de la prueba simulada.

Tabla 4

Descripción de las condiciones experimentales en los estudios de MH y RL

Condición experimental	Tamaño grupo de referencia	Tamaño grupo focal	Razón de tamaños
1	500	100	5
2	500	125	4
3	500	167	3
4	500	200	2,5
5	500	250	2
6	500	500	1
7	1500	300	5
8	1500	375	4
9	1500	500	3
10	1500	600	2,5
11	1500	750	2
12	1500	1500	1

Para la evaluación del poder del MH se estudiaron 12 ítems DIF: cuatro con DIF uniforme, cuatro con no uniforme y cuatro con mixto. El DIF uniforme se simuló aumentando el parámetro de dificultad para el grupo focal; en el no uniforme se aumentó el parámetro de discriminación y para el DIF mixto se aumentaron los valores de ambos parámetros. La tabla 5 muestra los parámetros de los ítems DIF y el área entre las CCI, de acuerdo con Raju, (1988); en el caso del DIF uniforme, el área se fijó en *0.4* y en el no uniforme ésta está entre *.3* y *.43*; valores que pueden encontrarse en análisis aplicados con pruebas estandarizadas y que corresponden a magnitudes que pueden ser detectadas por la mayoría procedimientos (Shepard, Camilli & Williams (1985)). En todos los casos los ítems DIF tuvieron dificultad moderada (entre –1 y 1) y discriminación superior a 0,7; esto permite controlar el efecto diferencial de los valores de los parámetros sobre el poder del MH y además, corresponden al tipo de ítems para los cuales el MH ha mostrado tasas de detección satisfactorias de acuerdo con lo reportado por Rogers & Swaminathan (1993).

Para obtener las matrices de datos se simularon en primer momento los vectores de θ, magnitud de atributo de los examinados, y posteriormente las respuestas de cada examinado en cada ítem. Los primeros se obtuvieron simulando distribuciones normales con media *0* y desviación *1* para ambos grupos ($\theta \approx n(0;1)$) y las respuestas de los examinados se simularon obteniendo las matrices de probabilidades de acertar en cada ítems para cada sujeto de acuerdo con un modelo logístico de tres parámetros ($P(u_i \mid \theta)$); finalmente, los valores *0* (falla en el ítem) o *1* (acierto) se

asignaron simulando para cada ítem y examinado, una distribución *bernoulli* con parámetro $P(u_i \mid \theta)$. Cada condición experimental se replicó 100 veces de manera que se analizaron 1200 pares de matrices de respuestas de diferentes grupos de individuos, en una única prueba con las características antes descritas.

Tabla 5

Parámetros de los ítems con DIF estudiados

Tipo de DIF	Número del Ítem	Grupo de referencia			Grupo focal			Área entre CCI
		a_R	b_R	c	a_f	b_f	c	
Uniforme	4	0.860	0.587	0.2	0.860	1.087	0.2	0.40
	24	0.858	-0.834	0.2	0.858	-0.334	0.2	0.40
	43	0.761	0.395	0.2	0.761	0.895	0.2	0.40
	73	0.755	-0.428	0.2	0.755	0.072	0.2	0.40
No uniforme	9	0.802	-0.413	0.2	1.302	-0.413	0.2	0.31
	49	0.841	-0.483	0.2	1.341	-0.483	0.2	0.30
	70	0.658	-0.564	0.2	1.158	-0.564	0.2	0.43
	82	0.815	0.012	0.2	1.315	0.012	0.2	0.30
Mixto	18	0.701	0.480	0.2	1.201	0.980	0.2	0.79
	42	0.704	-0.671	0.2	1.204	-0.171	0.2	0.78
	53	0.861	0.474	0.2	1.361	0.974	0.2	0.68
	83	0.966	-0.404	0.2	1.466	0.096	0.2	0.63

Análisis de los datos

Los análisis se realizaron en dos fases: la primera consistió en la evaluación de la calidad de los datos y en la segunda se identificó el efecto de los factores manipulados sobre el poder y el error tipo I del MH. Dentro de la primera fase se evaluó la recuperación de los parámetros y la precisión de las estimaciones a través de las réplicas. En primer lugar se estimaron por máxima verosimilitud, utilizando el BILOG 3 de Mislevy & Bock (1990), los parámetros de los ítems ($\hat{a}_{ir}$, $\hat{b}_{ir}$ y $\hat{c}_{ir}$ con número de ítems $i = 1,2,..,100$ y número de réplicas $r = 1,2,..,100$) y de los individuos ($\hat{\theta}_{jr}$ con número de individuos $j = 1,2,..,n$ y número de réplicas $r = 1,2,..,100$); estos se expresaron en una escala de media 0 y desviación estándar 1, utilizando el comando SCORE del BILOG 3. Uno de los archivos de de comandos del BILOG, creados en esta fase del análisis, se muestra en el anexo 2. La recuperación de los parámetros se evaluó mediante las estadísticas descriptivas de las estimaciones a través de las réplicas y su correlación de Pearson con los parámetros originales. Siguiendo a Cohen, Kane & Kim (2001), como índice de precisión de las estimaciones de los parámetros se utilizaron los cuadrados medios del error (CME) de las estimaciones con respecto al respectivo parámetro, a través de las réplicas, así:

$CME(\hat{a}_r) = \dfrac{\sum_{i=1}^{100}(\hat{a}_{ir} - a_i)^2}{100}$ y $CME(\hat{b}_r) = \dfrac{\sum_{i=1}^{100}(\hat{b}_{ir} - b_i)^2}{100}$ son los CME de las estimaciones de los parámetros de discriminación y dificultad, respectivamente, en la réplica *r,* y

$CME(\hat{\theta}_r) = \dfrac{\sum_{j=1}^{n}(\hat{\theta}_{jr} - \theta_j)^2}{n}$ son los CME de las estimaciones del parámetro de magnitud de atributo de los individuos para la misma réplica.

La segunda fase del análisis se inició con el cálculo del α_{MH}, el valor χ^2_{MH} y su significación, mediante el programa EZDIF de Waller (1998), que implementa un procedimiento de dos etapas para purificar el criterio de equiparación de los grupos; en el anexo 2 se muestra uno de los archivos de comandos creados para hacer estas estimaciones. El poder del estadístico se calculó mediante la tasa de detección (T_{MH}) de los ítems DIF a través de las réplicas utilizando valores de significación de 0.05 y 0.01; el error tipo I (E_{MH}) se calculó mediante la tasa de falsos positivos sobre los 88 ítems sin DIF en las 100 réplicas utilizando los mismos valores de significación. Posteriormente, mediante análisis de varianza se sometieron a prueba los efectos del tamaño de muestra del grupo de referencia (nr, con $nr_1 = 500$ y $nr_2 = 1500$), de la razón de tamaños (r, con $r_1 = 1$ $r_2 = 2$, $r_3 = 2{,}5$, $r_4 = 3$, $r_5 = 4$ y $r_6 = 5$) y de su interacción. Los modelos de análisis de varianza sometidos a prueba fueron $E_{MH} = \beta_1 nr + \beta_2 r + \beta_3 (nr)(r) + \varepsilon$ para la identificación de factores que afectan el error tipo I, y $T_{MH} = \beta_1 nr + \beta_2 r + \beta_3 (nr)(r) + \varepsilon$ para la estimación de los mismos efectos sobre el poder. El primero de ellos se analizó separadamente para cada tipo de ítem en términos de su dificultad y discriminación y el último para los tres tipos de DIF (uniforme, no uniforme y mixto). Los ítems no DIF se clasificaron en seis categorías resultantes de cruzar tres niveles de dificultad y dos de discriminación. Los tres niveles de dificultad fueron: a) baja si $b \leq -1.5$, b) media si $-1.5 \leq b < 1.5$ y c) alta si $b \geq 1.5$; y los dos niveles de discriminación fueron: a) baja si $a < 0.5$ y b) media si $0.5 \leq a \leq 1$. Dado el procedimiento utilizado para generar los parámetros de los ítems, descrito antes, ninguno de ellos tuvo una discriminación superior a 1. Finalmente, se hallaron las correlaciones bivariadas (producto momento de Pearson) y parciales entre los valores de los parámetros de los ítems no DIF y sus respectivas tasas de detección como DIF (tasa de FP) a través de las réplicas.

Resultados

Siguiendo las fases cumplidas en el análisis de los datos, los resultados se presentarán en cuatro apartes. En primer lugar se presentan los estadísticos descriptivos de las estimaciones de los parámetros TRI, en segundo lugar se muestran los resultados relacionados con la evaluación de la

calidad de los datos y de la precisión de las tales estimaciones, posteriormente se presentan los análisis relacionados con la identificación de los factores que afectaron el error tipo I, las tasas encontradas en las diferentes condiciones experimentales y las relaciones entre éstas y los parámetros de los ítems; finalmente, se muestran los resultados de la identificación de los factores que afectan el poder del MH para cada tipo de DIF.

Estadísticas descriptivas de las estimaciones

Las medias y desviaciones estándar de las estimaciones de los parámetros de los ítems en los diferentes grupos mostraron un comportamiento bastante satisfactorio teniendo en cuenta las distribuciones de a y b; y considerando que para los grupos focales el parámetro de dificultad se incrementó en 8 ítems y el de discriminación en otros 8. Como puede verse en la tabla 6, las medias y desviaciones estándar de $\hat{a}_i$ fueron próximas a $0,5$ y $0,2$, respectivamente, con ligeros incrementos en los grupos focales; las medias de $\hat{b}_i$ giraron alrededor de 0 y las desviaciones típicas fueron ligeramente inferiores a 1. La comparación de las estimaciones de los mismos parámetros entre los grupos de referencia y focal también arrojó resultados bastante satisfactorios. En la tabla 7 se muestran el promedio de las diferencias de la dificultad ($b_f - b_R$) y la discriminación ($a_f - a_R$) para los diferentes tipos de DIF en cada condición experimental. Las estimaciones reflejan de manera muy precisa las situaciones simuladas puesto que las diferencias en dificultad son próximas a $0,5$ en los ítems con DIF uniforme y mixto, similar comportamiento se observa en la discriminación para los ítems con DIF no uniforme y mixto; mientras que las diferencias en ambos parámetros son muy cercanas a 0 para los ítems sin DIF.

Tabla 6

Media y desviación típica de las estimaciones de los parámetros de los ítems en las condiciones experimentales de los estudios del MH y la RL.

Tamaño Referencia	Discriminación		Dificultad	
Tamaño focal	Media	D. Típica	Media	D. Típica
R 500	0.578	0.239	0.074	0.863
F 500	0.586	0.306	0.010	0.923
F 250	0.642	0.312	0.050	0.866
F 200	0.633	0.295	-0.013	0.854
F 167	0.622	0.276	0.110	0.858
F 125	0.660	0.289	0.130	0.823
F 100	0.704	0.298	0.065	0.787
R 1500	0.540	0.231	0.054	0.910
F 1500	0.611	0.330	-0.017	0.872
F 750	0.590	0.314	0.123	0.905
F 600	0.586	0.310	0.046	0.920
F 500	0.586	0.306	0.010	0.923
F 375	0.590	0.299	0.018	0.920
F 300	0.591	0.294	0.039	0.920

Tabla 7

Media de las diferencias de la estimaciones de dificultad y discriminación entre el grupo focal y de referencia, para los estudios del MH y la RL

Condición experimental	Discriminación				Dificultad			
	Uniforme	NoUniforme	Mixto	No DIF	Uniforme	NoUniforme	Mixto	No DIF
500 y 500	0.027	**0.522**	**0.525**	0.005	**0.401**	-0.132	**0.367**	-0.070
500 y 250	0.090	**0.573**	**0.589**	0.060	**0.424**	-0.088	**0.379**	-0.034
500 y 200	-0.005	**0.454**	**0.471**	0.007	**0.358**	-0.158	**0.332**	-0.134
500 y 167	-0.036	**0.414**	**0.404**	0.007	**0.499**	-0.019	**0.515**	-0.005
500 y 125	0.012	**0.438**	**0.499**	0.040	**0.543**	0.018	**0.480**	0.008
500 y 100	0.062	**0.539**	**0.543**	0.079	**0.419**	-0.057	**0.391**	-0.052
1500 y 1500	0.051	**0.572**	**0.602**	0.025	**0.353**	-0.117	**0.335**	-0.110
1500 y 750	0.019	**0.520**	**0.542**	0.008	**0.513**	0.008	**0.496**	0.030
1500 y 600	0.024	**0.514**	**0.521**	0.004	**0.437**	-0.081	**0.404**	-0.044
1500 y 500	0.020	**0.512**	**0.514**	0.005	**0.387**	-0.134	**0.365**	-0.080
1500 y 375	0.013	**0.483**	**0.511**	0.010	**0.417**	-0.117	**0.374**	-0.073
1500 y 300	0.010	**0.511**	**0.488**	0.013	**0.443**	-0.094	**0.406**	-0.047

Precisión de las estimaciones y recuperación de los parámetros

En la tabla 8 se presentan la media y desviación típica de los CME para $\hat{a}$, $\hat{b}$ y $\hat{\theta}$ y las correlaciones entre estas estimaciones y los parámetros respectivos. Puede observarse que los valores de CME (Subtabla 8a) de los parámetros de los ítems aumentaron ligeramente a medida que el tamaño de muestra disminuyó mientras que para el parámetro de magnitud de atributo éstos se mantuvieron alrededor de 0,07 en todas las condiciones experimentales. Además, los CME fueron mayores para la discriminación en las condiciones con grupo de referencia igual a 500 y comparables con los de la dificultad, cuando el grupo de referencia fue grande (1500); el mayor valor se observó para la discriminación en el grupo focal de 100 examinados.

De otra parte, exceptuando dos condiciones experimentales (grupo de referencia de 500 con grupos focales de 125 o 100), todas las correlaciones entre las estimaciones y los parámetros fueron superiores a 0,9; lo que garantiza una excelente recuperación de los parámetros. De manera similar a lo observado con los CME, las correlaciones (Subtabla 8b) de los parámetros de los ítems muestran una tendencia a disminuir ligeramente con el tamaño de muestra, mientras que se mantienen estables para la magnitud de atributo y son mejores para la dificultad que para la discriminación. En síntesis, los resultados de la primera fase del análisis apoyaron la calidad de los datos simulados y la adecuación de los procedimientos para generarlos; lo que permitió adelantar la segunda fase con un nivel de confianza satisfactorio sobre sus hallazgos.

Tabla 8

Media y desviación típica de los CME de $\hat{a}$, $\hat{b}$ y $\hat{\theta}$, y sus correlaciones con los respectivos parámetros para cada condición experimental.

8a. Cuadrados medios

Tamaño Referencia	Atributo		Discriminación		Dificultad	
Tamaño focal	Media	D. Típica	Media	D. Típica	Media	D. Típica
R 500	0.074	0.008	0.127	0.023	0.091	0.019
F 500	0.074	0.012	0.082	0.043	0.086	0.017
F 250	0.069	0.010	0.142	0.024	0.124	0.026
F 200	0.070	0.010	0.157	0.025	0.136	0.029
F 167	0.075	0.011	0.189	0.032	0.153	0.027
F 125	0.076	0.011	0.219	0.041	0.175	0.035
F 100	0.084	0.011	0.254	0.046	0.193	0.032
R 1500	0.071	0.011	0.072	0.013	0.068	0.016
F 1500	0.067	0.010	0.040	0.007	0.062	0.015
F 750	0.076	0.012	0.060	0.011	0.073	0.017
F 600	0.066	0.012	0.067	0.011	0.078	0.016
F 500	0.073	0.014	0.077	0.012	0.084	0.016
F 375	0.068	0.011	0.099	0.019	0.099	0.018
F 300	0.072	0.012	0.118	0.019	0.111	0.019

8b. Correlaciones

Tamaño Referencia	Atributo		Discriminación		Dificultad	
Tamaño focal	Media	D. Típica	Media	D. Típica	Media	D. Típica
R 500	0.965	0.002	0.936	0.012	0.954	0.009
F 500	0.965	0.003	0.961	0.007	0.957	0.008
F 250	0.967	0.003	0.928	0.012	0.937	0.013
F 200	0.969	0.004	0.921	0.013	0.931	0.014
F 167	0.967	0.003	0.905	0.016	0.923	0.014
F 125	0.970	0.005	0.889	0.021	0.912	0.018
F 100	0.969	0.005	0.862	0.090	0.903	0.016
R 1500	0.963	0.002	0.963	0.007	0.965	0.008
F 1500	0.970	0.001	0.980	0.003	0.969	0.008
F 750	0.966	0.002	0.970	0.006	0.963	0.008
F 600	0.966	0.002	0.966	0.005	0.960	0.008
F 500	0.965	0.003	0.961	0.006	0.958	0.008
F 375	0.966	0.003	0.950	0.009	0.950	0.009
F 300	0.966	0.003	0.940	0.009	0.944	0.009

Factores que afectan el error tipo I del MH

Los análisis de varianza para identificar los factores con efecto significativo sobre el error tipo I mostraron que, con muy pocas excepciones, los factores afectan el error tipo I sobre todo cuando los ítems tienen dificultad media. En la tabla 9 se resumen los resultados; tanto el tamaño del grupo de referencia como la razón de tamaños afectaron de manera significativa la tasa de falsos positivos (FP) para ítems de discriminación media y dificultad media o baja. Además, la razón de tamaños tuvo efecto significativo para ítems de alta dificultad y discriminación media.

Sin embargo, la significación estadística de los efectos sólo tiene interés aplicado si implica una inflación importante del error tipo I por encima de los valores nominales. En la tabla 10 se mues-

tran las medias de FP calculadas sobre las 100 réplicas de los 88 ítems no DIF, y las proporciones de ítems que fueron falsamente identificados como DIF entre el 6% y el 10% de las réplicas (tasa FP ligeramente alta, $0.5 < FP \leq .1$) y en más del 10% de las réplicas (tasa FP alta, $FP > .1$) para las diferentes condiciones experimentales. Con tamaño del grupo de referencia pequeño ($nr = 500$) la tasa total de FP se mantuvo igual o inferior al valor nominal; no obstante, con tamaños de muestra grandes ($nr = 1500$) se observaron valores ligeramente superiores al valor nominal. Además, con *nr* pequeño el porcentaje de ítem falsamente identificados como DIF en más del 10% de las réplicas, no superó el 2%, este porcentaje se incrementó con *nr* grande y con los tamaños de muestra máximos ($nr = nf = 500$), más del 80% de los ítems fueron falsamente identificados como DIF en más del 5% de las réplicas y 23% de los mismos resultaron FP más del 10% de las veces.

Tabla 9

Valor F y significación para los efectos sobre el error tipo I del MH por tipo de ítem

Tipo de ítem	Factor	Valor F	Significación
Dificultad baja y Discriminación baja	Tamaño referencia	2.3	0.135
	Razón R/f	0.8	0.543
	Interacción	1.7	0.174
Dificultad baja y Discriminación media	Tamaño referencia	**11.1**	**0.002**
	Razón R/f	**4.9**	**0.003**
	Interacción	0.8	0.520
Dificultad media y Discriminación baja	Tamaño referencia	**37.4**	**0.000**
	Razón R/f	**5.4**	**0.000**
	Interacción	**2.8**	**0.026**
Dificultad media y Discriminación media	Tamaño referencia	**117.7**	**0.000**
	Razón R/f	**15.4**	**0.000**
	Interacción	**5.7**	**0.000**
Dificultad alta y Discriminación baja	Tamaño referencia	1.9	0.175
	Razón R/f	0.7	0.609
	Interacción	**3.2**	**0.028**
Dificultad alta y Discriminación media	Tamaño referencia	1.5	0.228
	Razón R/f	**5.7**	**0.002**
	Interacción	2.2	0.098

La tabla 11 muestra las tasas de FP totales y las proporciones de ítems con tasa FP ligeramente alta y alta, para los diferentes tipos de ítems según los valores de dificultad y discriminación. En general, la tasa de FP se mantuvo muy cercana al valor nominal; sin embargo, lo superó ligeramente para ítems con discriminación media y dificultad baja o media. También fueron estas categorías de ítems las que en mayor porcentaje resultaron identificados como DIF; casi el 60% de los que tenían baja dificultad y discriminación media, y similar porcentaje de los de dificultad y discriminación medias, resultaron identificados como DIF en más del 5% de las réplicas.

Tabla 10

Tasa de FP total del MH y proporción de ítems FP para cada condición experimental

Tamaños de grupos (razón)	Tasa FP total		Proporciones de ítems FP	
	$\alpha = 0.05$	$\alpha = 0.01$	$0.5 < FP \leq .1$	$FP > .1$
500 y 500 (1)	0.05	0.01	0.28	0.01
500 y 250 (2)	0.06	0.01	0.50	0.01
500 y 200 (2.5)	0.05	0.01	0.36	0.02
500 y 167 (3)	0.05	0.01	0.36	0.00
500 y 125 (4)	0.04	0.01	0.20	0.00
500 y 100 (5)	0.03	0.01	0.16	0.00
1500 y 1500 (1)	0.08	0.02	0.59	0.23
1500 y 750 (2)	0.07	0.02	0.52	0.09
1500 y 600 (2.5)	0.06	0.01	0.58	0.08
1500 y 500 (3)	0.06	0.01	0.41	0.08
1500 y 375 (4)	0.06	0.02	0.45	0.02
1500 y 300 (5)	0.05	0.01	0.45	0.02

Tabla 11

Tasa de falsos positivos total del MH y proporción de ítems FP según tipos de ítems

Dificultad y discriminación	Tasa FP total		Proporciones de ítems FP	
	$\alpha = 0.05$	$\alpha = 0.01$	$0.5 < FP \leq .1$	$FP > .1$
b baja y a baja	0.05	0.01	0.35	0.03
b baja y a media	0.06	0.02	0.54	0.04
b media y a baja	0.05	0.01	0.36	0.02
b media y a media	0.06	0.02	0.48	0.09
b alta y a baja	0.04	0.01	0.28	0.00
b alta y a media	0.05	0.01	0.25	0.03

Las correlaciones entre la tasa de FP del ítem con $\alpha = 0.05$ y los valores de los parámetros de dificultad y discriminación fueron de -$.127$ y $.246$, respectivamente y similares resultados se observaron cuando se calcularon las correlaciones parciales controlando el efecto de nr, el de r o ambos simultáneamente (ver tabla 12). Las correlaciones parciales de la tasa de FP con la discriminación estuvieron entre .25 y .27, mientras que las del parámetro de dificultad estuvieron alrededor de -.13; sin embargo, cuando se observaron las correlaciones bivariadas dentro de cada condición experimental, aparecieron algunas diferencias de interés. En la tabla 12 se puede observar que para el parámetro de discriminación, éstas son mayores para mayores tamaños de muestra: son bajas y no significativas cuando $nr = 500$ y $nf < nr$, modestas cuando y $nf < nr$, y llega a ser alta ($.6$) con los máximos tamaños de los grupos. Para la dificultad, sin embargo, la mayoría de las correlaciones son inferiores a $.2$.

Tabla 12

Correlaciones entre la tasa de FP del MH y los parámetros de dificultad y discriminación de los ítems

12a. Correlaciones bivariadas

	Discriminación		Dificultad	
Tamaños de grupos (razón)	Correlación	Significación	Correlación	Significación
500 y 500 (1)	**0.39**	0.00	**-0.27**	0.01
500 y 250 (2)	0.07	0.55	-0.17	0.12
500 y 200 (2.5)	0.08	0.45	-0.19	0.07
500 y 167 (3)	-0.02	0.87	0.08	0.48
500 y 125 (49	0.18	0.09	0.06	0.56
500 y 100 (5)	0.13	0.24	0.00	0.99
1500 y 1500 (1)	**0.60**	0.00	-0.15	0.16
1500 y 750 (2)	**0.37**	0.00	-0.19	0.08
1500 y 600 (2,5)	0.23	0.03	**-0.32**	0.00
1500 y 500 (3)	**0.43**	0.00	-0.12	0.28
1500 y 375 (4)	**0.34**	0.00	-0.18	0.09
1500 y 300 (5)	0.24	0.02	-0.14	0.21

12b. Correlaciones parciales

	Discriminación		Dificultad	
	Correlación	Significación	Correlación	Significación
Controlando *nr*	0.26	0.00	-0.13	0.00
Controlando *r*	0.25	0.00	-0.13	0.00
Controlando *nr* y *r*	0.27	0.00	-0.14	0.00

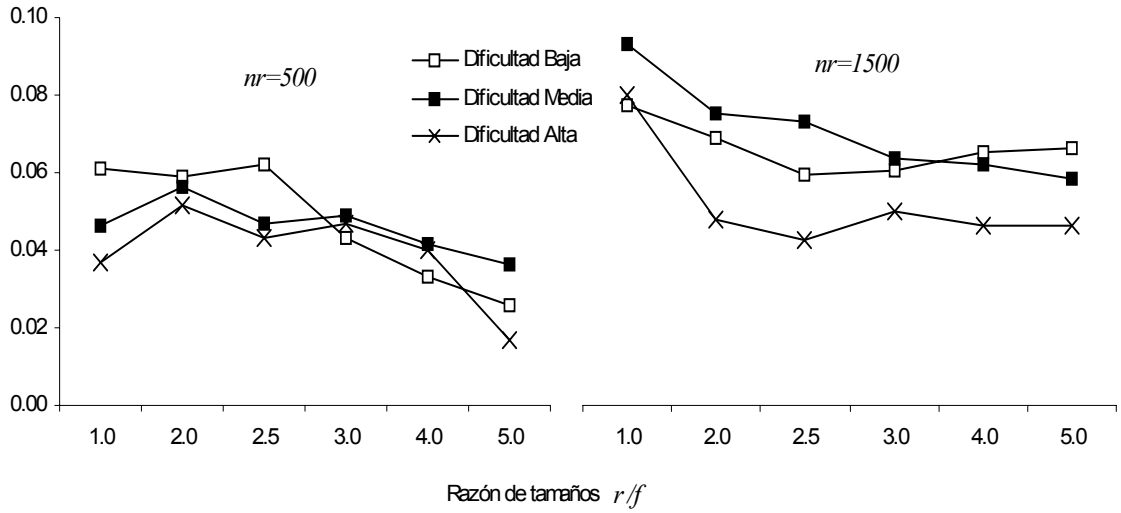


Figura 14. Tasa de FP del MH en función de la razón de tamaños según nivel de dificultad y tamaño del grupo de referencia

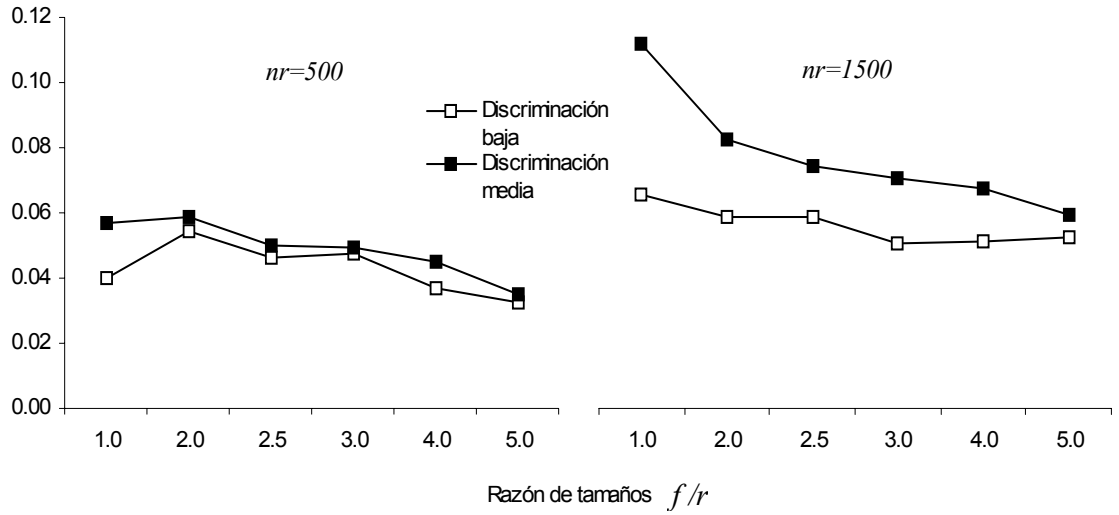


Figura 15. Tasa de falsos positivos del MH en función de la razón de tamaños según nivel de discriminación y tamaño del grupo de referencia

Finalmente, las figuras 14 y 15 muestran las tasas de FP con $\alpha = 0.05$ en función r y nr para los diferentes niveles de dificultad y discriminación, respectivamente. Cuando $nr = 500$ la tasa de FP se mantiene por debajo o alrededor del valor nominal para todas las razones de tamaños y lo niveles de dificultad y discriminación del ítem. Sin embargo, cuando $nr = 1500$ estos valores son más altos sobre todo cuando se trata de ítems con dificultad media o discriminación media; en general, los ítems con esta última característica son los que tienen mayor probabilidad de ser falsamente detectados como DIF y su tasa de FP puede elevarse por encima del 10% cuando los grupos son grandes.

Factores que afectan la potencia del MH

En la tabla 13 se presenta un resumen de los resultados de los análisis de varianza cuando la variable dependiente fue la tasa de detecciones correctas. Exceptuando la interacción entre tamaño del grupo de referencia y razón de tamaños para identificar DIF no uniforme, todos los factores tuvieron efecto significativo; estos resultados permiten afirmar el poder del MH para detectar DIF en ítems dicótomos se deja afectar de manera significativa por los tamaños de ambos grupos y por su interacción, con la excepción antes mencionada.

La tabla 14 y la figura 16 muestran las tasas de detección del MH en cada condición experimental para cada tipo de DIF. El MH mostró tasas bastante altas para detectar DIF uniforme y mixto con $nr = 1500$; con $nr = 500$ éstas se mantuvieron por encima de 0.8 (con $\alpha = 0.05$) para grupos focales de 200 o más en el caso de DIF uniforme, y de 125 o más en el caso del DIF mixto. Por el contrario el poder para detectar DIF no uniforme fue muy bajo con una tasa promedio del 74% para los tamaños de muestra máximos y todas las demás inferiores a este valor. Sin embargo, cuando se examinó la dispersión de las tasas de detección dentro de la misma condición

experimental para los grupos de cuatro ítems que tienen el mismo DIF, se encontró que ésta es importante cuando el ítem tiene DIF no uniforme; bajo todas las condiciones experimentales el ítem mejor identificado fue el número 70 y el peor fue el número 82 (los parámetros aparecen en la tabla 2) y la diferencia se incrementó con el tamaño de muestra. Por ejemplo, con grupos de 500 examinados cada uno, las tasas de detección fueron .45, .48, .66 y .12 para los ítems 9, 49, 70 y 82, respectivamente y cuando los grupos tienen 1500 examinados cada uno éstas fueron .87, .86, .96 y .25 para los mismos ítems; en el primer caso el rango de las tasas de detección es de .54 y en segundo es igual a .71. La dificultad del ítem 70, mejor identificado, es la mayor en valor absoluto mientras que la del 82, peor identificado, es muy cercana a 0. En el anexo 3 aparece la información completa de cada uno de los 12 ítems DIF, incluyendo la CCI, los valores de los parámetros y las tasas de detección.

Tabla 13

Valor F y significación para los efectos sobre las tasas de detección del MH por tipo de DIF

Tipo de DIF	**Factor**	**Valor F**	**Significación**
Uniforme	Tamaño referencia	327.2	0.000
	Razón R/f	54.1	0.000
	Interacción	42.7	0.000
No Uniforme	Tamaño referencia	26.7	0.000
	Razón R/f	3.5	0.016
	Interacción	0.1	0.994
Mixto	Tamaño referencia	26.8	0.000
	Razón R/f	7.9	0.000
	Interacción	5.9	0.001

Tabla 14

Tasa de detección del MH para cada condición experimental, por tipo de DIF

Tamaños de grupos (razón)	DIF Uniforme		DIF No uniforme		DIF Mixto	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
500 y 500 (1)	0.96	0.89	0.43	0.24	0.95	0.91
500 y 250 (2)	0.88	0.68	0.25	0.07	0.91	0.83
500 y 200 (2,5)	0.83	0.65	0.20	0.08	0.89	0.78
500 y 167 (3)	0.72	0.51	0.14	0.02	0.89	0.78
500 y 125 (4)	0.60	0.32	0.09	0.04	0.80	0.56
500 y 100 (5)	0.35	0.14	0.05	0.01	0.44	0.23
1500 y 1500 (1)	1.00	0.99	0.74	0.61	1.00	1.00
1500 y 750 (2)	0.99	0.99	0.54	0.32	1.00	0.99
1500 y 600 (2,5)	0.99	0.98	0.54	0.32	1.00	0.99
1500 y 500 (3)	0.99	0.98	0.53	0.36	0.99	0.97
1500 y 375 (4)	0.99	0.95	0.40	0.21	0.98	0.94
1500 y 300 (5)	0.96	0.88	0.38	0.17	0.96	0.90

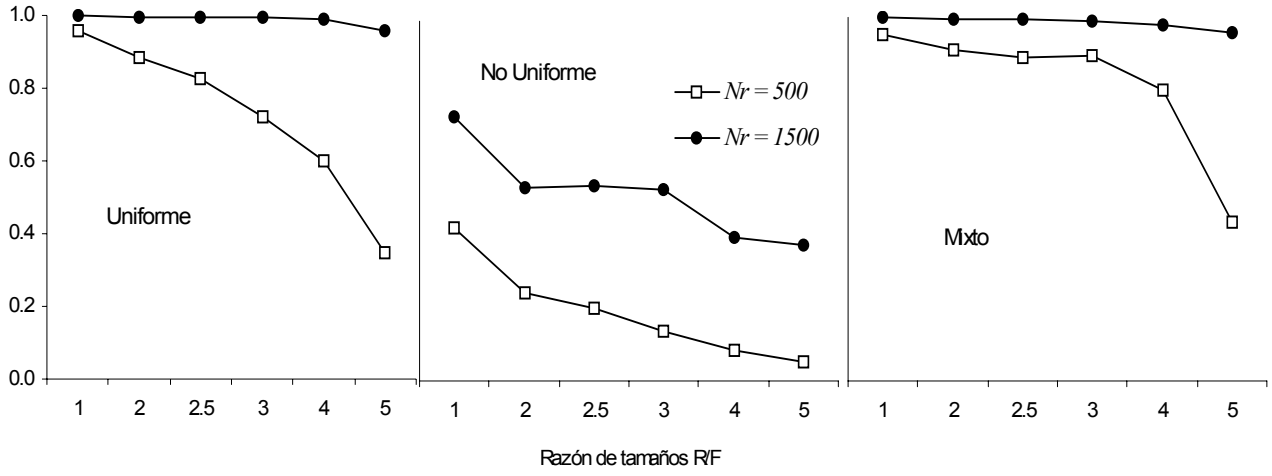


Figura 16. Tasa de detección del MH con $\alpha = 0.05$ en función de la razón de tamaños para cada tipo de DIF.

De otra parte, la prueba de Tukey para comparaciones pareadas post-hot arrojó diferencias significativas entre grupos iguales ($r = 1$) y grupos desiguales para la tasa de detección del MH cuando el DIF es uniforme, y diferencias entre $r = 5$ y las demás, cuando el DIF es mixto. En la figura 16 puede verse que cuando el tamaño de grupo de referencia es igual a 1500, las tasas de detección se mantienen altas para DIF uniforme y mixto, pero cuando el tamaño del grupo de referencia es igual a 500, las tasas descienden en función del tamaño del grupo focal en todos los tipos de DIF.

Discusión y conclusiones

En primer lugar los resultados relacionados con la precisión de las estimaciones y la recuperación de los parámetros a través de las réplicas fueron bastante satisfactorios. El comportamiento de los CME y las correlaciones entre las estimaciones y los parámetros estuvo dentro de lo esperado puesto que el tamaño de muestra afecta la precisión de las estimaciones de los parámetros de los ítems, mientras que el número de ítems afecta la precisión de las estimaciones del parámetro de magnitud de atributo (Cohen, Kane & Kim, 2001). En efecto, los CME de la dificultad y la discriminación crecieron y las correlaciones bajaron ligeramente a medida que el tamaño de los grupos disminuyó; mientras que los CME y las correlaciones de las estimaciones de θ se mantuvieron más o menos estables a través de las diferentes condiciones, todas simuladas con una sola prueba de 100 ítems. El procedimiento de generación de los datos, el número de réplicas utilizado y el número de ítems de la prueba permitieron una adecuados recuperación de los parámetros TRI tanto de los ítems como de los individuos.

Además, las diferencias simuladas entre los grupos en los parámetros de dificultad y discriminación para diferentes tipo de DIF se mantuvieron de manera bastante precisa a través de las réplicas de manera que, desde el marco de la TRI, el DIF que se simuló se mantiene en las dife-

rentes condiciones experimentales tanto en dirección como en magnitud. Dado que para los ítems con DIF uniforme y no uniforme se introdujeron diferencias en un solo parámetro; mientras que para el DIF mixto ésta se introdujo en la misma cantidad para los dos parámetros, la magnitud total de DIF es mayor para este último tipo de DIF (ver anexo 1) y cabe esperar mayores tasas de detección.

Sin embargo, los resultados satisfactorios en la precisión de las estimaciones de los parámetros no obvian la limitación de este tipo de estudios en cuanto a la generalidad de sus resultados. Como ocurre con cualquier experimento controlado, los resultados del presente estudio no son generalizables a situaciones que incluyan ítems DIF con valores de los parámetros más extremos de los utilizados aquí, o con diferentes magnitudes de DIF o con grupos que tengan diferentes distribuciones de la magnitud de atributo.

Tal vez el resultado más evidente y nada sorprendente, de los ANOVA para la identificación de factores que afecta el error tipo I, es el efecto significativo de los tamaños de los grupos sobre las tasas de falsos positivos del MH, lo que puede resultar más interesante es que dicho efecto no fue igual en las diferentes condiciones experimentales ni para los diferentes tipos de ítems categorizados según los valores de sus parámetros de dificultad y de discriminación. Cuando el ítem tenía dificultad media, o dificultad baja con discriminación media, tanto el tamaño del grupo de referencia como la razón de tamaños afectó el error tipo I del MH; aunque la tasa de falsos positivo se mantuvo por debajo del valor nominal para las condiciones con 500 examinados en el grupo de referencia, y ligeramente superiores al valor nominal cuando el grupo de referencia tenía 1500 examinados. De acuerdo con Bradley (1978), para $\alpha = 0.05$, tasas de falsos positivos entre 0.025 y 0.075 no implicarían una inflación importante del error tipo I; solamente la tasa total de falsos positivos obtenida con los máximos tamaños de grupo (grupos iguales de 1500 examinados) superó ligeramente este límite llegando al 8%; además, en esta misma condición experimental el 23% de los ítems individuales fueron incorrectamente detectados en más del 10% de las réplicas pero en ningún caso, en el 20% de las mismas. Estos resultados permitirían concluir que a pesar del efecto significativo de los tamaños de ambos grupos, no se observó una inflación importante del error como para que no resulte recomendable en la práctica.

Sin embargo, el mayor interés aplicado de este tipo de investigación es la identificación de las condiciones bajo las cuales la tasa de falsos positivos experimenta una inflación importante o se mantiene controlada. Cuando el grupo de referencia cambió de 500 a 1500, el error del MH creció de 3,8% a 6,3% (en promedio) y cuando la razón de tamaños cambió de 1 a 5, la tasa de falsos positivos bajó de 6,5% a 4%. Además, los ítems con discriminación media (mayores valores de discriminación en este estudio) tuvieron mayor probabilidad de resultar falsamente identificados como DIF; sin embargo, esta tasa se mantuvo alrededor del valor nominal cuando el tamaño del grupo mayoritario fue de 500 examinados y se elevó hasta casi el 10% cuando este grupo tuvo 1500 sujetos. Los resultados son menos obvios cuando se trata de ver la relación del parámetro

de dificultad y el error tipo I, y parecen mostrar una interacción entre el valor de dicho parámetro y los tamaños de los grupos: cuando el grupo mayoritario fue de 500 sujetos y la razón de tamaños estuvo entre 1 y 2,5, (grupos focales entre 200 y 500) los ítems con mayor probabilidad de resultar FP fueron los más fáciles pero esta relación cambió cuando la razón de tamaños aumentó; además, para $nr = 1500$ los ítems de dificultad media tuvieron mayor tasa de FP mientras la razón de tamaños estuvo entre 1 y 3. En el primer caso la tasa de FP se mantuvo controlada pero en el segundo experimentó una ligera inflación.

Estos resultados sugieren la existencia de un efecto de los valores de los parámetros de los ítems que podría evaluarse de manera más adecuada sin utilizar las categorías construidas para el presente estudio. De hecho, un resultado interesante fue la asociación entre los parámetros de los ítems y las tasas de falsos positivos. El error tipo I del MH mostró correlación positiva y significativa con la discriminación del ítem en grupos de 500 examinados cada uno y en todas las condiciones con grupo de referencia de 1500; tales correlaciones estuvieron entre 0,6 (para grupos iguales de 1500) y 0,24 (para $Nr = 1500$ y $r=5$) y, en general, disminuyeron a medida que la razón de tamaños aumentó. Aunque los valores de estas correlaciones fueron en general, modestos, parecen sugerir no solo una relación directa entre los parámetros de los ítems y la tasa de FP, sino una interacción entre éstos y los tamaños de los grupos. Esta hipótesis, sin embargo, amerita un análisis más sistemático observando valores de dificultad y discriminación más extremos que los utilizados en este estudio.

En síntesis, los resultados de esta parte del estudio permiten concluir que los psicólogos interesados en el uso del MH para el análisis de sus pruebas deben tener en cuenta que con un grupo de referencia de 500 examinados y un grupo focal que puede variar entre 100 y 500, tendrán controlado el error tipo I; si el grupo de referencia es de 1500 examinados, un grupo focal de hasta 300 no implicará inflación importante del error tipo I y con grupos de igual tamaño, éste puede exceder ligeramente el valor deseado. Sin embargo, no todos los ítems de la prueba tienen igual probabilidad de ser falsamente detectados como DIF; para los ítems más discriminativos esta probabilidad puede exceder ligeramente el valor nominal cuando el grupo mayoritario tiene 1500 examinados y el grupo minoritario tiene más de 500.

De otra parte, los resultados del segundo grupo de ANOVA mostraron que todos los factores considerados tuvieron efecto significativo sobre el poder del MH para detectar cualquier tipo de DIF; exceptuando la interacción $r \; X \; n_r$ para detectar DIF no uniforme. Nuevamente, el efecto del tamaño de muestra del grupo de referencia no resulta sorprendente y está bastante sustentado en la literatura especializada; además, las tasas totales de detección de DIF no uniforme resultaron tan bajas (entre 9% y 43% para $n_r=500$, y entre 38% y 74% para $n_r=1500$) que la identificación de los efectos significativos parece tener poco interés aplicado. Algunos resultados parcialmente comparables habían sido reportados por Swaminathan & Rogers (1990), Rogers & Swaminathan (1993), Uttaro & Millsap (1994) y Narayanan & Swaminathan (1996); en la tabla 15 se muestran las tasas de

detección encontradas en dichos estudios y en el actual, con $\alpha = 0.05$ para diferentes tamaños de grupos. Los resultados más disímiles son los del primer estudio con tasas de detección nulas; sin embargo, a diferencia de las demás, estas tasas fueron calculadas como el número de ítems detectados de 4 ítems DIF en una única réplica, y todos los ítems con DIF no uniforme tenían dificultad de 0. Los demás resultados son numéricamente comparables, pero ameritan algunos comentarios sobre las condiciones experimentales de cada uno de los estudios en particular sobre las magnitudes de DIF (áreas entre las CCI) y los parámetros de los ítems. En el estudio de Rogers & Swaminathan (1993) se simularon magnitudes de DIF entre .2 y .8, en Narayanan & Swaminathan (1996) estas estuvieron entre .4 y 1, y en el presente estudio estuvieron entre .3 y .43; así, los resultados encontrados permiten concluir que para magnitudes de DIF entre .2 y 1, medidas como el área entre las CCI, pueden esperarse tasas de detección del MH entre el 40% y el 46% para el DIF no uniforme con grupos iguales de 500 examinados, y pueden se mayores a medida que los tamaños de los grupos aumentan.

Tabla 15

Tasas de detección del MH para DIF no uniforme reportadas en diferentes estudios

Tamaños de grupos (razón)	Swaminathan & Rogers (1990)	Rogers & Swaminathan (1993)	Narayanan & Swaminathan (1996)	Estudio actual
250 y 250 (1)	.0	.35		
500 y 500 (1)	.0	.46	.41	.43
1500 y 1500 (1)				.74
500 y 250 (2)				.25
1000 y 500 (2)			.49	
1500 y 750 (2)				.54
500 y 200 (2,5)			.31	.20
1500 y 600 (2,5)				.54
500 y 125 (4)				.09
1000 y 200 (4)			.35	
1500 y 375 (4)				.40

En cuanto a los parámetros de los ítems, los mismos estudios han reportado un efecto importante de los valores de dificultad y discriminación sobre el poder del MH. Rogers & Swaminathan (1993) encontraron 56% de detección para ítems con baja dificultad y baja discriminación ($b_r = b_f = -1.5;\ .42 \le a_r \le .6;\ .74 \le a_f \le .89$), 46% para ítems de alta dificultad y alta discriminación ($b_r = b_f = 1.5$ y los mismos valores de discriminación) y solo 5% para ítems de dificultad media ($b_r = b_f = 0;\ .42 \le a_r \le 1.1;\ .74 \le a_f \le 1.7$). Utilizando los mismos valores de dificultad y valores de discriminación más diversos, Narayanan & Swaminathan (1996) encontraron tasas se detección de 66% y 53% para ítems con baja y alta dificultad, respectivamente; y entre 15% y 22% para ítems de difi-

cultad media. Con base en estos resultados los autores habían notado que con ítems muy fáciles o muy difíciles, el MH podía presentar resultados satisfactorios para detectar DIF no uniforme. De otra parte, Uttaro & Millsap (1994) con grupos iguales de 500 por grupo encontraron tasas de error tipo II altas (mayor de .2) para DIF no uniforme cuando las CCI se cruzaban hacia $\theta = 0$, con o sin impacto y para dos longitudes de prueba diferentes. En el presente estudio con 1500 examinados en el grupo de referencia, la tasa de detección estuvo entre 60% (con *r=5*) y 96% (con *r=1*) para un ítem con $b = -0.56,\ a_r = 0.66$ y $a_f = 1.16$, estuvo entre 50% y 86% para uno con $b = -0.48,\ a_r = 0.84\, b$ y $a_f = 1.34$; mientras que sólo alcanzó un máximo de 25% para un ítem con *b=0.012* (Ver anexo 3). Este resultado consistente con los anteriores, permite concluir que con tamaños de muestra grande el MH podría detectar de manera satisfactoria el DIF no uniforme de aquellos ítems que tengan dificultad tan baja como *-0,4* y discriminación moderada; además, esta tasa crece a medida que la razón de tamaños disminuye y llega a ser bastante alta con grupos de igual tamaño. Sin embargo, dado que en este estudio se controlaron los valores de los parámetros de los ítems DIF, no es posible observar la tasa de detección para ítems con diferentes valores de dificultad y discriminación, de manera que se puedan identificar unos rangos de valores de los parámetros y de tamaños de grupo para los cuales el MH resulte una técnica adecuada en la detección de DIF no uniforme.

Cuando se trató de detectar DIF uniforme o mixto con un grupo de referencia de 1500 examinados, las tasas de detección del MH se mantuvieron satisfactoriamente altas independientemente de la razón de tamaños. Sin embargo, cuando el tamaño del grupo de referencia fue de 500 examinados, la razón de tamaños tuvo efecto importante sobre el poder del estadístico. Una tasa de detección de al menos 90% (con $\alpha = 0.05$) sólo se obtuvo con grupos iguales para DIF uniforme o con una razón de tamaños de máximo 2, para DIF mixto. La tasa de detección de DIF uniforme descendió un 8% cuando la razón de tamaños cambió de 1 a 2 examinados del grupo de referencia por cada uno del grupo focal. De acuerdo con estos hallazgos el psicólogo aplicado puede tener la tranquilidad de que con 1500 examinados en el grupo de referencia y un grupo focal hasta cinco veces menor, el MH puede detectar satisfactoriamente los ítems con DIF uniforme o mixto si la magnitud del mismo es al menos de .4, definida como área entre las CCI (Raju, (1988) y no hay impacto o diferencias reales en la distribución de atributo entre los grupos. Pero, si el grupo de referencia es de solo 500 examinados debe tener un grupo focal de igual tamaño para tener resultados similares, y de mínimo 200 para tener tasas de detección superiores al 80%. De otra parte, sin considerar el tamaño de los grupos y a pesar de la poca variabilidad de los valores de los parámetros de los ítems DIF, pudo observarse que la tasa de detección del DIF uniforme aumentó con la discriminación del ítem (Ver anexo 3). Mazor, Clauser & Hambleton (1992) y Fidalgo, Mellenbergh & Muñiz (1999), simulando solamente DIF uniforme habían reportado hallazgos en el mismo sentido.

Capítulo 5:
EL EFECTO DEL TAMAÑO DE MUESTRA Y LA RAZÓN DE TAMAÑOS SOBRE LA REGRESION LOGISTICA

Aunque ya es bien sabido desde la estadística que el tamaño de muestra suele tener algún efecto sobre los procedimientos como la regresión logística, sigue siendo de interés aplicado evaluar la magnitud de tal efecto y estimar los tamaños de muestra óptimos, si existen, que brinden un poder adecuado y mantengan controlado el error tipo I, cuando se utilizan para objetivos particulares. Así, desde que Swaminathan & Rogers (1990) propusieron formalmente el uso de la regresión logística y el estadístico de prueba para la detección de ítems DIF, evaluaron el efecto del tamaño de muestra sobre su poder y error tipo I para detectar DIF uniforme y no uniforme, comparándolo con el MH. Utilizando grupos iguales de 250 y 500 examinados cada uno, ellos encontraron tasas muy bajas de FP (alrededor del 1%), calculadas a partir del número de ítems falsamente identificados en un única réplica, con respecto al total de ítems no DIF; además, con grupos de 500 obtuvieron precisiones del 100% y del 75% en la detección de DIF uniforme y no uniforme, respectivamente; mientras que con grupos de 250, la precisión fue de 75% para el DIF uniforme y 50% para el no uniforme. En este trabajo, sin embargo, solamente se realizaron réplicas de una condición experimental para comparar el poder relativo de la RL con el MH pero las tasas de FP y de detecciones correctas de cada procedimiento se calcularon como porcentaje de ítem detectados sobre los totales respectivos: 4 ítems DIF y 36 no DIF en una condición experimental, 6 DIF y 48 no-DIF en otra, y 8 DIF con 64 no-DIF en la última.

En los siguientes estudios, como en la mayoría de investigaciones sobre el tema, tanto el error tipo I poder como el poder se calculan como la proporción de detecciones falsas o correctas, en un número de réplicas mayor de 1. En un estudio reportado en el capítulo anterior, los mismos autores (Rogers & Swaminathan, 1993) utilizando los mismos tamaños de grupos (250 y 500 por grupo), encontraron un incremento en las tasas de detección de DIF tanto uniforme como no uniforme, cuando el tamaño de los grupos aumentó de 250 a 500 examinados en cada uno; en este estudio, sin embargo, no se evaluó el efecto sobre el error tipo I. Nuevamente con grupos iguales de 250, 500 y 1000 examinados, Jodoin & Huff (2001) encontraron un importante incremento del error tipo I de la RL con el aumento del tamaño de muestra, usando como estadístico de prueba el χ^2_{RL} de dos grados de libertad propuesto por Swaminathan & Rogers (1990), presentado en el capítulo 2. Con los grupos más pequeños, el mínimo porcentaje de ítems DIF en la prueba (10%) y sin impacto, ellos encontraron un error tipo I de 5.6 para $\alpha = 0.05$; sin embargo, esta tasa creció de manera importante a medida que se incrementaron los tamaños de los grupos y el porcentaje ítems DIF, sobre todo con diferencias en la distribución de magnitud de atributo entre los grupos. La tasa de FP alcanzó a ser de 17.7, 28.5 y 49.2 para grupos de 250, 500 y 1000, respectivamente, cuando el porcentaje de ítems DIF fue de 20% y las distribuciones de magnitud de atributo de

los grupos diferían tanto en media como en varianza. Además, con el incremento del tamaño de los grupos también aumentaron las tasas de detección de DIF uniforme y no uniforme; por ejemplo, con grupos iguales de 250 examinados cada uno y sin impacto, ellos encontraron tasas de 76.3% y 30% para DIF uniforme y no uniforme respectivamente, mientras que para grupos de 500 examinados éstas llegaron a 97% y 36% y con grupos de 1000 ascendieron a 100% y 80%.

Aunque los trabajos presentados hasta ahora reportaron resultados consistentes en cuando al efecto del tamaño de muestra sobre el poder y el error tipo I de la RL cuando se usa como estadístico de prueba el χ^2_{RL}, han utilizado grupos de igual tamaño y no se han ocupado del posible efecto de las diferencias de tamaños de los grupos. Narayanan & Swaminathan (1996) analizaron cuatro combinaciones de tamaños cruzando dos para el grupo de referencia (500 y 1000) y dos para el grupo focal (200 y 500), además de observar el efecto del tamaño de muestra sobre el poder de los tres procedimientos que estudiaron (MH, RL y SIBTEST), encontraron que los resultados parecían sugerir un efecto de la razón de tamaños y concluyeron que es necesario investigar esta hipótesis.

En un estudio similar Jodoin & Gierl (2001) analizaron el comportamiento del error tipo I y el poder de la RL con seis combinaciones de tamaños de grupo resultantes de cruzar $n's$ de 250, 500 y 1000, con $n_f \leq n_r$. Utilizando como estadístico de prueba el χ^2_{RL} ellos encontraron una tasa de falsos positivos de 5.3% para grupos iguales de 250 examinados sin impacto y con el 10% de ítems DIF en la prueba. Este valor se incrementó con el tamaño de muestra llegando a 13.1% para grupos de 1000 examinados con media de magnitud de atributo diferente, y a 15.8% con estos mismos grupos cuando el porcentaje de ítem DIF fue de 20%. Además, los autores encontraron incrementos de las tasas de detección de los ítems DIF con el aumento de los tamaños de los grupos, pero dichos incrementos no fueron regulares. En la figura 17 se han graficado las tasas de detección reportadas por Jodoin & Gierl (2001, pag. 341 y 343) para los diferentes tamaños de los grupos con idéntica distribución de la magnitud de atributo, no se reportó el poder para grupos de 1000 examinados con 20% de ítems DIF en la prueba puesto que en esta condición se presentó una inflación importante del error tipo I. Los resultados de la gráfica permiten observar que el poder aumenta con los tamaños del grupo focal cuando el grupo de referencia es igual a 500 pero se presenta un descenso cuando el tamaño del grupo de referencia crece y la diferencia entre los dos grupos es mayor (tamaños de 1000 y 250); esta tendencia parece mantenerse cuando las distribuciones de magnitud de atributo son diferentes, sin embargo, en estas condiciones la inflación del error tipo I fue importante y por tanto los autores no reportaron las tasas de detecciones correctas. Este comportamiento y los hallazgos reportados antes por Narayanan & Swaminathan (1996) parece apoyar la hipótesis de que la razón de tamaños de los grupos puede tener algún efecto sobre el comportamiento del χ^2_{RL} en la detección de DIF.

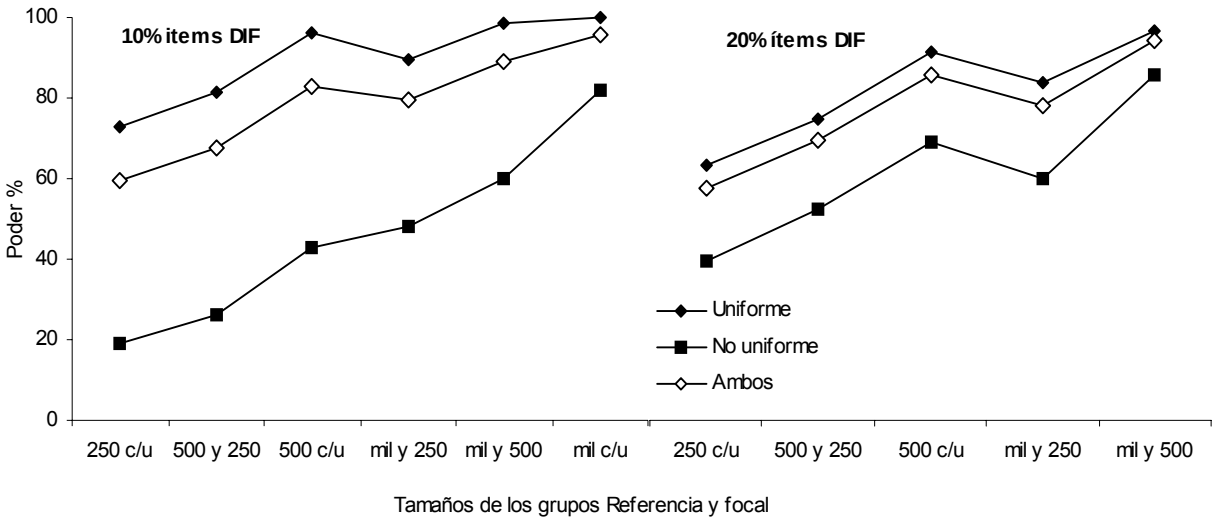


Figura 17. Tasa de detección del χ^2_{RL} reportadas por Jodoin & Gierl (2001) para diferentes tamaños de los grupos

Debe anotarse sin embargo, que el principal interés de Jodoin & Huff (2001) y Jodoin & Gierl (2001) era el comportamiento del poder y en error tipo I del χ^2_{RL} en comparación con otros dos procedimientos que consideran el tamaño del efecto, y bajo diferentes condiciones especialmente de diferencias en las distribuciones de magnitud de atributo entre los grupos. Ellos utilizaron tres estadísticos de prueba para la RL: el χ^2_{RL} tradicional y dos que combinan el χ^2_{RL} con el $R^2\Delta$ de Zumbo, (1999); el primero de ellos ($\chi^2_2 R^2\Delta$) con dos grados de libertad, prueba la hipótesis de ausencia de DIF tanto uniforme como no uniforme simultáneamente, mientras que el otro ($\chi^2_1 R^2\Delta$), con 1 grado de libertad, prueba las hipótesis separadamente para cada tipo de DIF. Aunque los hallazgos de los dos estudios han sido consistentes en mostrar que en diferentes condiciones experimentales, los estadísticos que consideran el tamaño del efecto, especialmente el $\chi^2_1 R^2\Delta$, resultan más efectivos que el χ^2_{RL} en el control del error tipo I manteniendo tasas de detección comparables, éste último sigue siendo más utilizado en los estudios aplicados. Siguiendo el procedimiento del estudio presentado en el capítulo anterior, esta investigación buscó evaluar el efecto del tamaño de muestra y de la razón de tamaños de los grupos, sobre el poder y el error tipo I de la RL utilizando como estadístico de prueba el χ^2_{RL} .

Método

En este estudio se siguió el mismo procedimiento descrito en el capítulo anterior con los mismos datos simulados en las mismas condiciones experimentales. En síntesis, se realizó un experimento de Monte Carlo con doce condiciones experimentales (ver tabla 4) resultantes de cruzar dos tamaños de muestra del grupo de referencia ($nr = 500,\ 1500$) y seis niveles de razón de

tamaños de los grupos ($r = {nr}/{nf} = 1,\ 2,\ 2.5,\ 3,\ 4,\ 5$). Se simuló una única prueba unidimensional compuesta por 100 ítems siguiendo modelos TRI de tres parámetros con $b \approx n(0,1)$, $a \approx n(0.5;0.2)$ y $c = 0.2$, con los parámetros que aparecen en el anexo 1. Dentro de esta se incluyeron 12 ítems DIF, 4 para cada clase de DIF, con los parámetros que aparecen en la tabla 5. Estos últimos ítems tuvieron dificultad entre –1 y 1 y discriminación entre 0,7 y 1. Los parámetros de los individuos se obtuvieron simulando distribuciones normales con media 0 y desviación 1 para ambos grupos ($\theta \approx n(0;1)$), posteriormente se calcularon las probabilidades de acierto de cada examinado en cada ítem siguiendo el modelo logístico de tres parámetros y; finalmente, se obtuvieron las matrices de respuestas de los individuos en los ítems simulando distribuciones *bernoulli* con parámetro $P(u_i \mid \theta)$. Cada condición experimental se replicó 100 veces de manera que se analizaron 1200 pares de matrices de respuestas de diferentes grupos de individuos en la prueba.

La primera fase del análisis de datos, consistente en la estimación de los parámetros tanto de los individuos como de los ítems y en la evaluación de la calidad de los datos y de la estimación de los parámetros TRI, se había adelantado ya en el primer estudio. Las estimaciones se realizaron por máxima verosimilitud utilizando el BILOG 3 (Mislevy & Bock, 1990) y, después de expresarlos en una escala común, se obtuvieron sus descriptivos, las correlaciones de Pearson entre las estimaciones y los parámetros, y los cuadrados medios del error de éstas con respecto a los parámetros siguiendo los procedimientos de Cohen, Kane & Kim (2001).

La segunda fase del análisis, específica para este estudio, se inició con las estimaciones de los parámetros de la regresión logística $\hat{\tau}_0, \hat{\tau}_1, \hat{\tau}_2$ y $\hat{\tau}_3$ y el cálculo del valor p para $\hat{\tau}_2$ y $\hat{\tau}_3$ (valores que se notarán como p_2 y p_3) siguiendo un procedimiento de dos etapas para eliminar del criterio de equiparación de los grupos, los ítems identificados como DIF; este procedimiento se adelantó mediante el programa EZDIF Waller, (1998). Con base en estos resultados se identificaron los ítems DIF con dos valores de significación ($\alpha = 0.05$ y $\alpha = 0.01$), en ambos casos se declaró el ítem DIF cuando $p_2 < \alpha$ o $p_3 < \alpha$ en la respectiva réplica. Para cada condición experimental se calculó el poder y el error tipo I de la RL mediante la tasa de detecciones correctas (T_{RL}) de los ítems DIF a través de las réplicas, y la tasa de FP (E_{RL}) sobre los 88 ítems no DIF en las 100 réplicas, respectivamente. Posteriormente, siguiendo el mismo procedimiento descrito en el capítulo anterior, se llevaron a cabo los análisis de varianza para la identificación de los factores que afectaron el poder y el error tipo I de la RL, sometiendo a prueba los modelos $T_{RL} = \beta_1 nr + \beta_2 rt + \beta_3 (nr)(rt) + \varepsilon$ y $E_{RL} = \beta_1 nr + \beta_2 rt + \beta_3 (nr)(rt) + \varepsilon$, respectivamente. El primer análisis se realizó para cada tipo de DIF y el segundo para las diferentes categorías de ítem según su dificultad y discriminación; estas categorías fueron las mismas del primer estudio, resultantes de cruzar tres niveles de dificultad (baja si $b \leq -1.5$, media si $-1.5 \leq b < 1.5$ y alta si

$b \geq 1.5$) y dos de discriminación (baja si $a < 0.5$ y media si $0.5 \leq a \leq 1$). Finalmente se hallaron las correlaciones bivariadas (producto momento de Pearson) y parciales entre los valores de los parámetros de los ítems no DIF y la tasa de FP, calculada como el porcentaje de falsas detecciones través de las réplicas.

Resultados

Dado que los datos utilizados en este estudio fueron los mismos que habían sido evaluados en el primero, la calidad de los datos y de las estimaciones de los parámetros ya estaba garantizada. Los resultados de las estadísticas descriptivas, las correlaciones y los CME se mostraron en las tablas 6 a 8. Aquí se presentarán los resultados específicos de este estudio: los factores que afectan el error tipo I y el poder de la RL.

Tabla 16

Valor F y significación de los efectos sobre el error tipo I de la RL por tipo de ítem

Tipo de ítem	Factor	F	Significación
Dificultad baja y Discriminación baja	Tamaño referencia	0.2	0.686
	Razón R/f	2.7	**0.043**
	Interacción	1.2	0.312
Dificultad baja y Discriminación media	Tamaño referencia	0.3	0.560
	Razón R/f	1.2	0.348
	Interacción	1.0	0.433
Dificultad media y Discriminación baja	Tamaño referencia	0.2	0.619
	Razón R/f	2.6	**0.036**
	Interacción	0.7	0.593
Dificultad media y Discriminación media	Tamaño referencia	2.0	0.161
	Razón R/f	0.6	0.637
	Interacción	0.7	0.590
Dificultad alta y Discriminación baja	Tamaño referencia	1.3	0.269
	Razón R/f	0.6	0.670
	Interacción	2.7	**0.051**
Dificultad alta y Discriminación media	Tamaño referencia	0.8	0.370
	Razón R/f	1.2	0.352
	Interacción	3.4	**0.023**

Factores que afectan el error tipo I de la RL

En la tabla 16 se muestran el valor F y su significación de los análisis de varianza para identificar los factores con efecto significativo sobre el error tipo I de la RL, por tipo de ítem. De acuerdo con estos resultados el tamaño del grupo de referencia no afectó de manera significativa el error tipo I de la RL, la razón de tamaños tuvo efecto significativo para ítems con baja discriminación y hubo interacción significativa entre los dos factores, para ítems con alta dificultad. Con grupos de referencia de 500 examinados y $\alpha = 0.05$, las tasas de FP estuvieron entre 0 y 15% con media

de 6.13%; cuando el grupo de referencia aumentó a 1500 sujetos, estas tasas estuvieron en el mismo rango con promedio de 6.26%. El comportamiento fue bastante similar para $\alpha = 0.01$: con $nr = 500$ las tasas de FP estuvieron entre 0 y 5% con media de 1.26%, y con $nr = 1500$ el rango de variación fue de 0 a 7% con media de 1.46%.

En las tablas 17 se muestran las medias de error (porcentaje promedio de FP de los ítems no DIF) para las diferentes condiciones experimentales. En todas ellas se observaron tasas promedio de FP ligeramente superiores al valor nominal pero sin exceder los intervalos de Bradley (1978). Además, dentro de cada tamaño del grupo de referencia, las tasas de FP se mantuvieron más o menos estables a lo largo de las diferentes razones de tamaños. En la misma tabla aparece el porcentaje de ítems con tasas de FP entre el 6% y el 10% y superiores al 10%, con $\alpha = 0.05$, para cada condición experimental. En cada condición resultaron falsamente detectados en más de 5 de las 100 réplicas, entre el 49% y el 74% de los ítems; aunque el mayor porcentaje se presentó con los máximos tamaños de grupos (1500 examinados por grupo), y el menor se presentó con tamaños pequeños (500 y 125 examinados), no se observó un incremento sistemático de dicho porcentaje con el aumento del tamaño del grupo de referencia o con la disminución de la razón de tamaños.

Los resultados del mismo análisis para cada una de las seis categorías de tipos de ítems según los valores de dificultad y discriminación aparecen en la tabla 18. Nuevamente se observan tasas promedio de FP ligeramente superiores al valor nominal con máximos de 6.8% para $\alpha = 0.05$ y 2.3% para $\alpha = 0.01$. Además, los porcentajes de ítems con tasas de FP ligeramente altas (entre 6% y 10%) y altas (superiores a 10%) tendieron a disminuir a medida que la dificultad del ítem aumentó.

Tabla 17

Tasa promedio de FP (%) de la RL y porcentaje de ítem falsamente detectados para cada condición experimental

Razón de tamaños	Tasas promedio (%) de FP				% de ítems FP con $\alpha = 0.05$			
	$nr = 500$		$nr = 1500$		$nr = 500$		$nr = 1500$	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$6 \le FP < 10$	$FP \ge 10$	$6 \le FP < 10$	$FP \ge 10$
1	6.4	1.4	6.8	2.3	58	3	68	6
2	6.5	1.4	6.0	1.4	67	2	46	3
2.5	6.2	1.2	6.2	1.2	47	9	48	5
3	6.2	1.3	6.3	1.2	55	5	68	5
4	5.7	1.2	5.9	1.2	47	2	53	2
5	5.8	1.1	6.5	1.4	51	1	62	3

Tabla 18

Tasa de FP (%) de la RL y porcentaje de ítems falsamente detectados para los tipos de ítem según dificultad y discriminación

Dificultad y discriminación	Tasas promedio (%) de FP		% de ítems FP con $\alpha = 0.05$	
	$\alpha = 0.05$	$\alpha = 0.01$	$6 \leq FP < 10$	$FP \geq 10$
b baja y a baja	6.5	1.3	60	03
b baja y a media	6.4	1.4	65	02
b media y a baja	6.3	1.4	58	04
b media y a media	6.1	1.4	52	04
b alta y a baja	5.4	1.0	36	06
b alta y a media	6.6	1.4	58	03

Dado que la razón de tamaños tuvo efecto significativo sobre el error tipo I para dos categorías de ítems y además hubo un efecto de interacción entre ésta y el tamaño del grupo de referencia para otras dos, en la figura 18 se presentan las tasas de FP con $\alpha = 0.05$ en función de r y nr para cada una de dichas categorías y en el anexo 4 se presentan las mismas tasas para los dos valores de significación utilizados. En general con $\alpha = 0.05$ los valores variaron entre 2% y 9%; cuando los ítems tuvieron dificultad media, estos estuvieron entre 5.5% y 9.0% y no se observaron variaciones importantes para los tamaños de grupo de referencia o las razones de tamaño. Los ítems con baja dificultad tuvieron tasas de FP entre 4.4% y 8.8% con valores máximos cuando los grupos fueron de igual tamaño ($r=1$), una notable disminución cuando la razón de tamaños aumentó de 1 a 2 y ligeros incrementos cuando ésta pasó de 4 a 5. La mayor variabilidad en las tasas de falsos positivos se observaron para ítems con alta dificultad, los valores variaron entre 2% y 9% y se observó una clara interacción entre los dos efectos. Además, entre estos ítems se observaron mayores diferencias según la categoría de la discriminación: la tasa de FP para ítems con baja discriminación estuvo entre 2.0% y 7.3% (media de 5.4%) con los mínimos valores para razones entre 1 y 2.5; mientras que para ítems con discriminación media estas tasas estuvieron entre 5.0% y 9.0% (media de 6.6%) con máximos valores cuando $nr= 500$ y $r= 2$ y cuando $nr=1500$ y $r=1$ o $r=5$. En el anexo 5 se muestran las tasas de detección y valores de los parámetros para los ítems de esta última categoría.

Las correlaciones bivariadas y parciales entre los valores de los parámetros y la tasa de FP con $\alpha = 0.05$ fueron despreciables en todas las condiciones con diferente tamaño de grupo ($r > 1$) tanto para el parámetros de dificultad como para el de discriminación; sin embargo la correlación de Pearson entre la dificultad del ítem y la tasa de FP resultó modesta y significativa en las dos condiciones con grupos focal y de referencia de igual tamaño: con grupos de 500 examinados cada uno, ésta fue -.22 ($p<.05$) y con grupos iguales de 1500 examinados, fue de -.254 ($p<.05$). De igual manera, las correlaciones parciales controlando el tamaño del grupo de referencia, la razón de tamaños o ambos simultáneamente, fueron todas muy cercanas a 0 y no significativas;

sin embargo, la correlación parcial entre la dificultad y tasa de FP controlando el tamaño de muestra del grupo referencia, fue de -0.24 ($p<.005$) cuando la razón fue igual a 1, y nula en las demás condiciones. En el anexo 6 aparecen las correlaciones tanto bivariadas como parciales entre los parámetros y las tasas de FP de la RL.

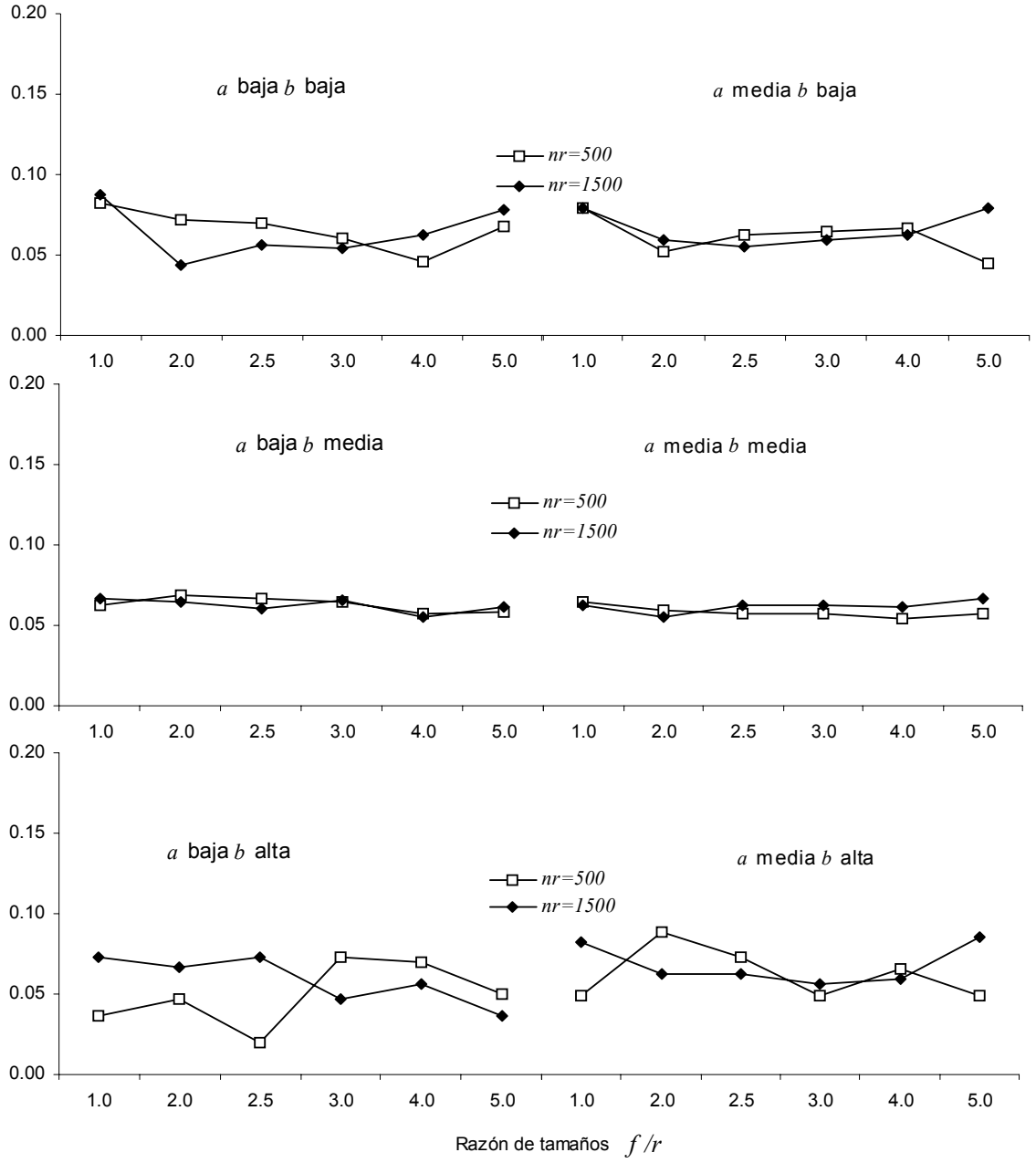


Figura 18. Tasa de FP del χ^2_{RL} según tamaño del grupo de referencia y razón de tamaños para las diferentes categorías de ítems

Factores que afectan las tasas de detecciones correctas de la RL

En la tabla 19 se presentan el valor F y su significación para los análisis de varianza cuando la variable dependiente fue la tasa de detecciones correctas de la RL para los tres tipos de DIF. A

diferencia de lo observado en los resultados relacionados con el error tipo I, todos los factores tuvieron efecto significativo sobre el poder de la RL calculado como el porcentaje de detecciones correctas a través de las réplicas.

Tabla 19

F y significación para los efectos sobre las tasas de detección de la RL por tipo de DIF

Tipo de DIF	Factor	Valor F	Significación
Uniforme	Tamaño referencia	197.6	0.000
	Razón R/f	39.5	0.000
	Interacción	9.2	0.000
No Uniforme	Tamaño referencia	397.4	0.000
	Razón R/f	31.9	0.000
	Interacción	13.6	0.000
Mixto	Tamaño referencia	199.7	0.000
	Razón R/f	24.4	0.000
	Interacción	18.1	0.000

Tabla 20

Tasa de detección de la RL para cada condición experimental, por tipo de DIF

Tamaño de grupos (razón)	**Uniforme**		**No uniforme**		**Mixto**	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
500 y 500 (1)	0.17	0.04	0.91	0.74	0.97	0.89
500 y 250 (2)	0.13	0.05	0.72	0.41	0.91	0.72
500 y 200 (2,5)	0.12	0.02	0.58	0.24	0.87	0.59
500 y 167 (3)	0.06	0.02	0.55	0.21	0.76	0.44
500 y 125 (4)	0.10	0.02	0.42	0.15	0.73	0.37
500 y 100 (5)	0.07	0.01	0.44	0.13	0.62	0.23
1500 y 1500 (1)	0.40	0.16	1.00	0.99	1.00	1.00
1500 y 750 (2)	0.24	0.08	0.99	0.95	1.00	0.99
1500 y 600 (2,5)	0.22	0.09	0.98	0.91	1.00	0.99
1500 y 500 (3)	0.21	0.06	0.98	0.88	1.00	0.99
1500 y 375 (4)	0.19	0.04	0.91	0.74	0.99	0.96
1500 y 300 (5)	0.14	0.04	0.89	0.69	0.97	0.90

La tabla 20 y la figura 19 muestran las tasas de detección de la RL en cada condición para cada tipo de DIF. Cuando se trató de detectar DIF uniforme con grupos iguales del máximo tamaño y $\alpha = 0.05$ sólo se alcanzó una tasa de detección del 40%, ésta disminuyó un 16% cuando la razón de tamaños aumentó de 1 a 2, y en todas las demás condiciones se encontraron tasas inferiores a esta última. Con estos resultados la identificación de los factores que afectan el poder para la detección de DIF uniforme carece de interés aplicado. El panorama fue muy diferente con los otros dos tipos de DIF. Con grupo de referencia de 1500 examinados las tasas de detección de DIF no uniforme y mixto fueron satisfactoriamente altas ($T_{RL} \geq .9$) independientemente de la razón de tamaños, solamente se observó un pequeño descenso con grupos focales menores de 500 para DIF

no uniforme. Sin embargo, con 500 examinados en el grupo de referencia estas tasas disminuyeron con el tamaño del grupo focal, para DIF no uniforme solamente se alcanzó una tasa por encima del 90%, con grupos iguales, ésta descendió casi un 20% (con $\alpha = 0.05$ **)** cuando la razón de tamaños aumentó de 1 a 2 y llegó a 42% y 44% con razones de tamaños de 4 y 5, respectivamente. El descenso en la tasa de detección del DIF mixto fue más paulatino pero sistemático con el aumento de la razón de tamaños; dicha tasa disminuyó casi un 30% cuando el grupo focal pasó de 250 a 100 examinados llegando a un 62% en este último caso.

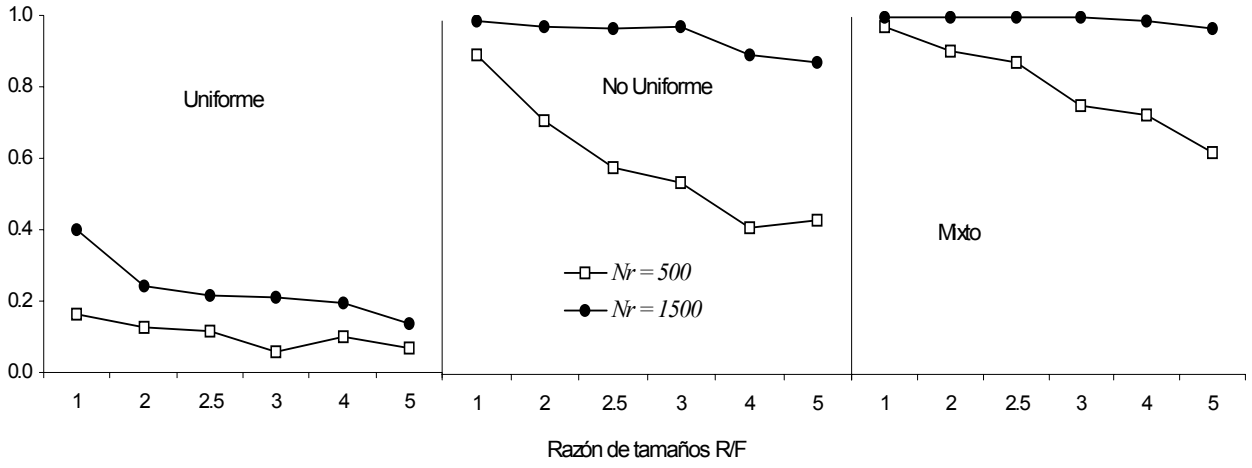


Figura 19. Tasa de detección de la RL con $\alpha = 0.05$ en función de la razón de tamaños para cada tipo de DIF

El examen de las tasas de detección de los diferentes ítems DIF no mostró una relación aparente entre ésta y los valores de los parámetros de dificultad y discriminación. En el anexo 7 aparecen las tasas de detección para cada ítem DIF en cada condición experimental (anexo 7.1) y un resumen de las mismas junto con la información sobre los parámetros del ítem y la magnitud de DIF (anexo 7.2). Puede observarse que el comportamiento de la tasa de detección es muy similar para los diferentes ítems: altas (mayores de .85) para todos los valores de la razón de tamaños cuando el DIF es no uniforme o mixto y $nr = 1500$, con descensos importantes para estos mismos tipos de DIF (de 92% a 35% para DIF no uniforme y de 100% a 55% para DIF mixto) a medida que crece la razón de tamaños cuando $nr = 500$; y bajas (entre 2% y 42%) y con descensos importantes pero menos regulares a través de la razón de tamaños, para DIF uniforme.

Discusión y conclusiones

Un primer resultado un tanto sorprendente fue el efecto nulo del tamaño del grupo de referencia sobre la tasa de FP, como estimador del error tipo I de la RL; ésta aumentó apenas un .13% en promedio, cuando el tamaño del grupo de referencia pasó de 500 a 1500 examinados. Este resultado es aparentemente contradictorio con los hallazgos de Jodoin & Huff (2001) y Jodoin & Gierl (2001), quienes encontraron una importante inflación del error tipo I por encima de límites admisibles, con el aumento del tamaño de los grupos, sobretodo cuando había diferencias en la distribu-

ción de la magnitud de atributo. Sin embargo, esa aparente contradicción puede explicarse ya que los estudios reportados no utilizaron un procedimiento de purificación del criterio de igualación de los grupos. Los hallazgos de Narayanan & Swaminathan (1996) son más coherentes con los resultados encontrados aquí; aunque ellos no reportaron prueba de significación del efecto del tamaño de grupo sobre el error tipo I, las tasas de FP fueron similares o apenas ligeramente superiores alcanzando un máximo valor de 8.9%; además en el estudio de Narayanan & Swaminathan (1996) la tasa de FP aumentó apenas un .75% en promedio, cuando el tamaño del grupo de referencia pasó de 500 a 1000 examinados. De otra parte, con grupos iguales de 1000 examinados Navas-Ara & Gómez Benito (2002a) encontraron que la tasa de FP en la detección de DIF uniforme disminuyó de .22 a .02 cuando se utilizó un procedimiento de dos etapas para purificar la medida de magnitud de atributo. Aunque en este último estudio las autoras calcularon la tasa de FP como la proporción de ítems incorrectamente identificados como DIF sobre el total de ítem no DIF promediadas para tres réplicas, sus resultados apoyan el uso del procedimiento de dos etapas para mantener controlado el error tipo I aún con tamaños de muestra grandes, haciendo que este factor resulte sin efecto significativo. De esta manera estos hallazgos de la presente investigación permiten concluir que el usuario de la RL para la identificación de DIF mantendrá bastante controlado el error tipo I incluso con grupos de 1500 examinados y grupos focales iguales o hasta 5 veces menores, cuando utilice un procedimiento de dos etapas para purificar el criterio de equiparación y en condiciones similares a los simuladas aquí.

De otra parte, aunque la razón de tamaños tuvo efecto significativo sobre la tasa de FP para ítems con baja discriminación y $b<1.5$, y hubo efecto de interacción entre los dos factores analizados (r y nr) para ítems con alta dificultad, no se observaron inflaciones tan importantes del error tipo I como para cuestionar el uso de la RL con algún ítems. De acuerdo con los intervalos propuestos por Bradley, (1978) para $\alpha = 0.05$ las tasas de falsos positivos admisibles deberían estar entre 2.5% y 7.5%, y para $\alpha = 0.01$ se espera $0.5\% \leq FP \leq 1.5\%$. Solamente en dos casos se superaron ligeramente estos límites: a) en ítems con dificultad baja y grupos de igual tamaño ($r=1$) las tasas de FP estuvieron alrededor de 8% con máximo de 8.8% y b) en ítems con alta dificultad y discriminación media cuando $nr= 500$ y $r= 2$ y cuando $nr=1500$ y $r=1$ o $r=5$; en este caso las tasas de FP estuvieron entre 8.3% y 9.0%. Utilizando valores de discriminación más extremos que los simulados en el presente estudio pero con puntos de corte similares, Narayanan & Swaminathan (1996) habían encontrado un resultado similar puesto que en su estudio las mayores tasas de FP se presentaron para ítems con alta discriminación: con $\alpha = 0.05$ ellos encontraron una tasa promedio de 9.9% para ítems con baja dificultad y alta discriminación, 8.8% para ítems con dificultad media y alta discriminación, y entre 6.4 y 7.5 para ítems con baja discriminación. Puesto que los valores de discriminación categorizados por Narayanan & Swaminathan (1996) como altos

corresponden a los que en el presente estudio hemos denominado medios pero ellos utilizaron un rango de valores más altos[20], estos resultados parecen apoyar la conclusión de que los ítems más discriminativos tienen mayor probabilidad de resultar falsamente identificados como DIF cuando se utiliza la RL, aunque tal probabilidad es apenas ligeramente superior al valor nominal.

Sin embargo, algunos resultados interesantes del presente estudio parecen sugerir que el parámetro de dificultad también puede estar relacionado con la tasa de falsos positivos pero en interacción con otros factores. En primer lugar, para las dos categorías de ítems con dificultad alta hubo efecto de interacción entre la razón de tamaños y el número de examinados en el grupo de referencia pero la tasa de FP superó los límites de Bradley (1978) solamente para ítems de discriminación media y para algunos valores de r (ver anexos 4 y 5). Este resultado sugiere algún efecto del parámetro de dificultad en interacción con el de discriminación y la razón de tamaños de los grupos, sin embargo, dado el procedimiento utilizado para la generación de los parámetros de los ítems en el presente estudio, la categoría de ítems con alta dificultad y discriminación media solamente quedó conformada por 3 ítems con similares valores de parámetros así que esta hipótesis y la identificación de las condiciones en las cuales el error tipo I puede resultar alto, si es que existen, requieren más investigación. También resulta interesante en este grupo de ítems que todos presentaron un incremento en las tasas de FP con grupo de referencia de 1500 examinados cuando la razón de tamaños aumentó de 4 a 5 (ver anexo 5.3), aunque no se trata de un resultado concluyente puede sugerir un aumento del error tipo I con grupos de referencia y razones de tamaños grandes. Para verificar esta hipótesis es necesario planear estudios con valores mayores de r y valores más heterogéneos de los parámetros de los ítems.

En segundo lugar, las correlaciones entre las tasas FP y los valores de los parámetros sólo resultaron significativas para el parámetro de dificultad con grupos de igual tamaño y tomaron valores negativos; la única correlación parcial significativa también se encontró con el parámetro de dificultad para razón de tamaños igual a 1 controlando el efecto del tamaño del grupo de referencia y además, las tasas de FP superaron los límites de Bradley (1978) para ítems con baja dificultad y grupos de igual tamaño. Aunque las tasas promedio solo superaron ligeramente estos límites llegando a un máximo de 8.8%, algunos ítems individuales con dificultad negativa, como el 5, 11, 25, 52, 69, 72, 84, 89, 98 y 99 (ver valores de los parámetros en el anexo 1) tuvieron tasas entre 10% y 15% con grupos de igual tamaño. Este resultado muestra la existencia de un efecto de la razón de tamaños sobre el error tipo I del χ^2_{RL} para ítems con baja dificultad y discriminación entre 0 y 1. Para efectos aplicados estos hallazgos permiten concluir que los ítems muy fáciles pueden tener una probabilidad mayor (hasta del 15% con $\alpha = 0.05$**)** de resultar falsamente identificados como

[20] El intervalo de valores de discriminación de los ítems sin DIF fue $.29 \leq a \leq 1.4$ en el estudio de Narayanan & Swaminathan (1996), y $0 \leq a \leq .92$ en el presente trabajo.

DIF utilizando la RL si se tienen grupos iguales tan pequeños como de 500 examinados cada uno o tan grandes como 1500 por grupo; esta probabilidad sin embargo, se mantendrá muy cercana a los valores nominales para estos mismos ítems si de tienen grupos de diferente tamaño y condiciones similares a las del presente estudio.

Los resultados del segundo grupo de análisis tendientes a la identificación de los factores que afectan el poder de la RL, estimado a partir de la tasa de detecciones correctas de cada ítems DIF sobre 100 réplicas, mostraron que tanto el tamaño del grupo de referencia como la razón de tamaños y su interacción, tuvieron efecto significativo para todos los tipos de DIF. Este resultado, sin embargo, tiene interés aplicado si la significación estadística del efecto implica un cambio importante en la tasa de detecciones como para recomendar o cuestionar el uso del estadístico en los trabajos aplicados. El examen de las tasas de detección de DIF uniforme mostró resultados muy poco alentadores: tasas de detección entre 2% y 42% con valores máximos para grupos iguales de 1500 examinados cada uno, y valores inferiores al 20% para todas las condiciones e ítems cuando el grupo de referencia tenía 500 examinados. Rogers & Swaminathan (1993) habían advertido la posible debilidad del χ^2_{RL} con 2 grados de libertad para detectar DIF uniforme y Navas-Ara & Gómez Benito (1994) habían mostrado que efectivamente las tasas de detección de DIF uniforme de la RL sin purificar la variable de equiparación de los grupos eran inferiores a las del MH. Sin embargo, varios estudios han reportado tasas superiores a las halladas aquí para la detección de DIF uniforme: Rogers & Swaminathan (1993) encontraron tasas promedio de 66% y 80% para grupos iguales de 250 y 500, respectivamente; los valores medios hallados por Jodoin & Huff (2001) fueron superiores a 71% con grupos iguales de 250, 500 y 1000 examinados; Jodoin & Gierl (2001) encontraron tasas medias entre 60% y 92% para grupos entre 250 y 1000 examinados; y en el estudio de Gómez Benito & Navas-Ara (1996) la RL logística identificó más del 90% de ítems DIF en cada una de cuatro réplicas con grupos iguales de 1000 examinados. En todos estos estudios sin embargo, se simularon magnitudes de DIF promedio superiores al valor fijado en el presente estudio para el DIF uniforme: .4. En conclusión, estos resultados del presente estudio no apoyan el uso de la RL con el χ^2_{RL} para la detección de DIF uniforme cuando la magnitud de DIF sea tan pequeña como un área entre CCI de .4 y en condiciones similares a las simuladas aquí.

A diferencia de lo observado con DIF uniforme, los resultados fueron bastante satisfactorios cuando se trató de detectar los otros dos tipos de DIF. Con grupo de referencia de 1500 examinados las tasas de detección de DIF no uniforme y mixto se mantuvieron por encima de .89 para todas las razones de tamaños. Sin embargo, con grupo de referencia de 500 examinados las tasas de detección disminuyeron con el tamaño del grupo focal para ambos tipos de DIF; este descenso fue considerable cuando la razón de tamaños cambió de 1 a 2 y se trató de detectar DIF no uniforme, mientras que para DIF mixto los decrementos fueron más regulares a medida que aumentó la razón de tamaños. De hecho, la prueba Tukey para comparaciones pareadas post-hot arrojó

diferencias significativas entre grupos iguales ($r=1$) y grupos diferentes para detectar DIF no uniforme con $nr=500$, mientras que para DIF mixto las diferencias estuvieron entre las tres condiciones con $1 \leq r \leq 2.5$ y las demás. De acuerdo con estos resultados se puede concluir que con grupos de referencia grandes la RL tendrá un poder muy satisfactorio en la detección de DIF no uniforme y mixto aún cuando el grupo focal sea 5 veces menor; sin embargo, cuando el grupo de referencia sea tan pequeño como de 500 examinados, será necesario un grupo focal de igual tamaño y uno de al menos 200 examinados para detectar DIF no uniforme y mixto respectivamente, con un poder superior al 85% (con $\alpha=.05$). Debe anotarse, sin embargo, que las diferencias entre las tasas de detección de DIF no uniforme y mixto pueden deberse a la magnitud de DIF empleada en cada tipo: en el primer caso las áreas entre CCI estuvieron entre .3 y .43 mientras que en el segundo estuvieron entre .63 y .79. Si se tiene en cuenta que en pruebas estandarizadas las magnitudes de DIF suelen ser moderada (entre .4 y .6 aproximadamente) y que los ítems con mayor área entre las CCI mostraron mejores tasas de detección para ambos tipos de DIF (ver anexo 7), puede esperarse que en condiciones similares, el comportamiento de la RL sea comparable para ambos tipos de DIF.

Finalmente, a diferencia de lo reportado en varios estudios (Rogers & Swaminathan, 1993; Narayanan & Swaminathan, 1996) no se encontraron asociaciones importantes entre los valores de los parámetros de los ítems DIF y las tasas de detección en ningún tipo de DIF. Debe anotarse, sin embargo, que los valores de los parámetros de los ítems DIF fueron muy similares entre sí y más homogéneos que los utilizados en otros estudios. Para el grupo de referencia todos los ítems con DIF tuvieron $.66 \leq a_r \leq .97$ y $-.83 \leq b_r \leq .59$ con incrementos para el grupo focal en la dificultad o en la discriminación, dependiendo del tipo de DIF. En estas condiciones los cambios en las tasas de detección no son atribuibles a los valores de los parámetros sino al tipo de DIF, el tamaño del grupo focal y la razón de tamaños, factores de interés en el presente estudio.

Capítulo 6:

EVALUACION DEL EFECTO DE TRES FACTORES SOBRE EL χ^2 DE LORD EN LA DETECCION DE DIF

A diferencia de ocurrido con la producción investigativa sobre los procedimientos que no se basan en la TRI, como el MH, los estudios sobre el χ^2 de Lord no han hecho énfasis en la evaluación del efecto del tamaño de muestra; sino que se han ocupado de aspectos como el algoritmo de estimación de los parámetros, el tipo de modelo TRI, el procedimiento de purificación de la escala o el procedimiento de equiparación. Sin embargo, para efectos del presente estudio resulta de interés la revisión de los reportes de las tasas de FP y de detecciones correctas en los diferentes tamaños de grupos utilizados.

Algunos de los estudios que reportan resultados referentes al comportamiento del error tipo I son los de Candell & Drasgow (1988), Lim & Drasgow (1990), Cohen & Kim (1993), Kim y Cohen, 1994 y Hidalgo Montesinos & López Pina (2002). En primer lugar,Candell & Drasgow (1988) compararon los resultados del Ji cuadrado de Lord con y sin purificación de la escala para la equiparación de los parámetros, con grupos iguales de 300 y 500 examinados; estimando los parámetros por MVM, utilizando un procedimiento iterativo para purificación de la escala y con $\alpha=.005$ encontraron tasas de FP entre 0 y 4%, muy similares para los dos tamaños de muestra. Por su parte, Lim & Drasgow (1990) evaluaron la efectividad del Ji cuadrado de Lord utilizando dos algoritmos diferentes de estimación (MVM y MB) en modelos de un parámetro uni y multidimensionales con grupos iguales de 750 y 250 examinados; los resultados mostraron que en general, el error tipo I, calculado como la proporción de detecciones incorrectas sobre 50 réplicas, se mantuvo bastante controlado alrededor de los valores nominales para ambos tamaños de muestra; por ejemplo, con datos unidimensionales y $\alpha=.05$ los porcentajes de FP estuvieron entre 2.9% y 4.6% para $n=250$ y entre 4.8% y 7.4% para $n=750$. De manera similar, comparando la efectividad del estadístico con dos medidas de área ($Z(ESA)$ y $Z(H)$) Cohen & Kim (1993) utilizaron tamaños de muestra de 100 y 500 examinados por grupo y encontraron que, con muy pocas y pequeñas excepciones, el error tipo I, calculado como el promedio de FP en cinco réplicas, se mantuvo dentro de los niveles nominales. Los mismos autores (Kim y Cohen, 1994) estudiaron el error tipo I utilizando MVM y MB para la estimación de los parámetros y ajustando diferentes modelos TRI (3p, 3p-c y 2p) con grupos de 1000 y 250 examinados. Los resultados mostraron que, a diferencia de lo reportado por McLaughlin & Drasgow (1987) estimando los parámetros por MVC, para modelos 3p-c y 2p el error tipo I, evaluado como el porcentaje de FP a través de 100 réplicas, se mantuvo por debajo de los niveles nominales para ambos tamaños de muestra (entre 3% y 4.2% para $n=1000$ y entre 2% y 3.2% para $n=250$, con $\alpha=.05$); sin embargo para modelos de tres parámetros se presentaron importantes inflaciones, por ejemplo con $\alpha=.05$ éste llegó al 45% para $n=1000$ y a 37% para $n=250$

cuando los parámetros de estimaron por MVM. Finalmente, en un estudio más reciente Hidalgo Montesinos & López Pina (2002) han evaluado el estadístico de Lord para detectar DIF en ítems de respuesta graduada; utilizando diferentes porcentajes de ítems con DIF y diferentes magnitudes de DIF y con grupos de 250, 500 y 1000 examinados, encontraron porcentajes de falsos positivos entre 0.68% y 6.34% para $n=250$, entre 0.03% y 15.25% para $n=500$ y entre 0.92% y 37.88% con $n=1000$ cuando no se utilizó un procedimiento de purificación; con un procedimiento de dos etapas, estas tasas fueron menores de 1.5 para $n=250$ y $n=500$ y menores a 2.1% con los mayores tamaños de muestra.

En lo que tiene que ver con la potencia del estadístico, en el primer estudio citado (Candell & Drasgow, 1988) con $\alpha=.005$ se encontraron tasas de detecciones correctas entre 50% y 80% para $n=300$ y entre 70% y 90% para $n=500$. Con el mismo nivel de significación, Lim & Drasgow (1990) encontraron que el poder del estadístico, calculado a partir de las detecciones de los ítems DIF, se mantuvo por encima del 72% con tamaños de grupo de 750 examinados y estuvo entre 12% y 68% con $n=250$. Por otra parte, a juzgar por los resultados reportados, las tasas de detección encontradas en el estudio de Cohen & Kim (1993) no fueron muy altas con grupos pequeños; para las diferentes condiciones con grupos de 100 examinados los porcentajes de falsos negativos fueron superiores al 75% mientras que con grupos de 500 examinados éstos se mantuvieron por debajo del 40%[21]. Finalmente, Hidalgo Montesinos & López Pina (2002) encontraron porcentajes de detecciones correctas entre 2.5% y 95.3% para $n=250$, entre 11% y 100% para $n=500$ y superiores al 31.5% para grupos de 1000 examinados, cuando no se utilizó purificación de la escala de equiparación; estos porcentajes fueron muy similares utilizando un procedimiento de dos etapas para purificar la escala: entre 2.5% y 97.9% para los grupos más pequeños, superiores a 11.5% para $n=500$ y mayores del 32% para los grupos de mayor tamaño.

Aunque los resultados de estos estudios no son comparables entre sí puesto que se han obtenido en condiciones diferentes y con objetivos también diferentes, parecen sustentar un efecto del tamaño de los grupos sobre las tasas de FP y de detecciones correctas del Ji cuadrado de Lord. Además, a la luz de los mismos resultados parece razonable esperar tasas de error tipo I alrededor de los valores nominales cuando los parámetros se estiman con MVM o con MB y se tienen grupos de hasta 1000 examinados; sin embargo, las tasas de detecciones correctas parecen variar de manera importante dentro de un mismo tamaño de muestra. Debe anotarse además, que todos los estudios simularon grupos de igual tamaño y ninguno de ellos evaluó el efecto del tamaño de grupo ni, obviamente, la razón de tamaños; además, aunque en algunos de los estudios se

[21] En el trabajo citado los autores reportan número medio de falsos negativos y falsos positivos para las diferentes condiciones estudiadas, los porcentajes que se presentan aquí se han calculado considerando la longitud de la prueba y el porcentaje de ítems con DIF en cada condición experimental.

simularon diferentes tipos de DIF, en ninguno de ellos se reportaron resultados diferenciales para los diferentes tipos. Este estudio tiene como objetivo evaluar el efecto de tres factores: tamaño de muestra, razón de tamaños y longitud de las prueba, sobre el estadístico de Lord en la detección de DIF uniforme, no uniforme y mixto.

Método

Siguiendo un procedimiento similar al de los estudios anteriores se realizó un experimento de Monte Carlo con datos simulados manipulando tres factores. Además de los factores analizados en los dos estudios anteriores: tamaño del grupo de referencia (*nr = 500, 1000*) y la razón de tamaños (*r= 1, 2, 2.5, 3, 4 y 5* examinados en el grupo de referencia por cada uno del focal), se simularon dos longitudes de prueba (*k= 50 , 100* ítems). Al cruzar completamente estos tres factores resultaron 24 condiciones experimentales como se muestra en la tabla 21. Las razones para la selección de los niveles de los dos primeros factores se presentaron en el capítulo 4; además teniendo en cuenta que la longitud de la prueba afecta la precisión de la estimación de θ (Cohen & Kim, 2001) en este estudio se simularon dos tamaños diferentes para evaluar su posible efecto sobre la estimación de los parámetros de magnitud de atributo y por consiguiente, sobre el funcionamiento del estadístico. Dentro de cada condición se consideraron el tipo de DIF (uniforme, no uniforme y mixto), los valores de los parámetros de los ítems y el porcentaje de ítems con DIF; éste último sin embargo, se anidó dentro de la longitud de la prueba. Finalmente, para examinar la relación entre los valores de los parámetros de los ítems y las variables dependientes, se crearon tres categorías de dificultad y tres de discriminación. Los niveles de dificultad fueron baja cuando $b_i < -1.5$, media cuando $-1.5 \leq b_i \leq 1.5$ y alta cuanto $b_i > 1.5$; para la discriminación se crearon los mismos niveles para $a_i < .2$, $.2 \leq a_i \leq .8$ y $a_i > .8$, respectivamente.

Generación de los datos

Para garantizar la comparabilidad de los resultados, en este estudio se utilizaron los mismos datos de los dos estudios anteriores pero a diferencia de los mismos, aquí se ajustaron modelos de dos parámetros. Conocidas las limitaciones que presenta este estadístico con modelos de tres parámetros (Lord, 1980, Kim y Cohen, 1994), esta estrategia metodológica permitió comparar la eficiencia del Ji cuadrado con los procedimientos estudiados en los dos capítulos anteriores y, a la vez, observar el funcionamiento de éste, en condiciones poco óptimas de ajuste de los modelos. Los parámetros de los 100 ítems para el grupo de referencia se generaron siguiendo distribuciones normales para la dificultad y la discriminación ($b \approx n(0,1)$ y $a \approx n(0.5;0.2)$) y una constante para el parámetro de pseudoazar ($c = 0.2$). Para simular el DIF, el valor del parámetro de dificultad se aumentó en cuatro ítems para el grupo focal (DIF uniforme), el parámetro de discriminación se aumentó en otros cuatro (DIF no uniforme) y en otros tantos se aumentaron los valores de ambos parámetros (DIF mixto). La prueba corta se conformó con los 50 primeros ítems, los valores

de los parámetros se encuentran en el anexo 1. En síntesis, se simularon dos pruebas unidimensionales de longitud 100 y 50 ítems con 12% y 14% de ítems con DIF, respectivamente.

Tabla 21

Descripción de las condiciones experimentales en el estudio del Ji cuadrado de Lord

Condición	Tamaño referencia	Tamaño focal	Razón de tamaños	Número de ítems	Condición	Tamaño referencia	Tamaño focal	Razón de tamaños	Número de ítems
1	500	100	5	50	13	1500	300	5	50
2	500	100	5	100	14	1500	300	5	100
3	500	125	4	50	15	1500	375	4	50
4	500	125	4	100	16	1500	375	4	100
5	500	167	3	50	17	1500	500	3	50
6	500	167	3	100	18	1500	500	3	100
7	500	200	2.5	50	19	1500	600	2.5	50
8	500	200	2.5	100	20	1500	600	2.5	100
9	500	250	2	50	21	1500	750	2	50
10	500	250	2	100	22	1500	750	2	100
11	500	500	1	50	23	1500	1500	1	50
12	500	500	1	100	24	1500	1500	1	100

Para generar las bases de datos de los individuos se utilizaron también los vectores de parámetros que se habían obtenido en los estudios anteriores, es decir igual distribución de magnitud de atributo para los grupos focal y de referencia distribuidos normalmente con media 0 y desviación típica de 1 ($\theta \approx n(0,1)$). Para cada condición experimental se generaron las dos matrices de probabilidades de acierto de cada individuo en cada ítem de la prueba siguiendo un modelo logístico de tres parámetros, $(P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}$ con $i = 1,2..k$ y $j = 1,2..n_f$ o $j = 1,2..n_r$); a partir de estas matrices se generaron las de respuestas simulando, para cada celda, una distribución Bernouilli con probabilidad igual a $P_i(\theta_j)$. Estas bases de datos se replicaron 50 veces para un total de 2400 bases de datos, 1200 pares focal-referencia.

Análisis de datos

En los análisis de los datos se siguió la misma estrategia de los anteriores estudios: una primera fase de evaluación de la precisión de las estimaciones y recuperación de los parámetros, y una segunda fase dedicada a la identificación de los factores con efecto sobre el poder y el error tipo I del estadístico. En primer lugar se estimaron los parámetros de los ítems y de los individuos y se expresaron en la misma métrica, utilizando el BILOG 3 de Mislevy & Bock (1990). En el anexo 8 se muestra un archivo de comandos utilizados en esta fase del estudio, puede verse que se trata fundamentalmente del mismo tipo de análisis de los estudios anteriores, con tres diferencias: se ajustaron modelos de 2 parámetros, se grabó la matriz de varianzas y covarianzas, y se varió el número de ítems de la prueba de acuerdo con la condición experimental. Posteriormente se obtu-

vieron las estadísticas descriptivas de las estimaciones a través de las réplicas, su correlación de Pearson con los parámetros originales y los cuadrados medios del error (CME) de las estimaciones con respecto al respectivo parámetro, siguiendo el procedimiento de Cohen, Kane & Kim (2001), como se describió en el capítulo 4.

En la segunda fase del análisis se calculó el χ^2 de Lord y su significación para cada ítem en cada réplica, para lo cual se construyeron macros en *Visual Basic* utilizando las matrices de varianzas y covarianzas estimadas por el BILOG en la fase anterior. La única función de las macros fue leer las estimaciones obtenidas previamente y hacer los productos matriciales necesarios para obtener el valor del estadístico, no se utilizó procedimiento de purificación alguno. Para cada ítem se calculó su tasa de detección a través de las 50 réplicas con valores de significación de .05, .01 y .1. La potencia de la prueba se calculó como la tasa promedio de detección (T_{χ^2}) de los ítems con DIF y su error tipo tipo I (E_{χ^2}) se calculó mediante la tasa promedio de detección de los ítems sin DIF. Mediante análisis de varianza se sometieron a prueba los efectos de los tres factores de interés y sus interacciones de segundo orden; los modelos de análisis de varianza sometidos a prueba fueron entonces $E_{\chi^2} = \beta_1 nr + \beta_2 r + \beta_3 k + \beta_4 (nr)(r) + \beta_5 (nr)(k) + \beta_6 (r)(k) + \varepsilon$ para identificar los factores que afectan el error tipo I, y el mismo con variable dependiente T_{χ^2} para la estimación de los mismos efectos sobre la potencia. Los análisis de varianza para evaluar los efectos sobre el error tipo I se realizaron para cada tipo de ítem según su nivel de dificultad y discriminación, mientras que los análisis de los factores que afectan la potencia del estadístico, se realizó para cada tipo de DIF. Finalmente, se hallaron las correlaciones entre los valores de los parámetros de los ítems sin DIF y sus respectivas tasas de detección como DIF (tasa de FP) a través de las réplicas.

Resultados

En la tabla 22 se presentan las estadísticas descriptivas de las estimaciones de los parámetros de dificultad y discriminación sobre las réplicas para los diferentes tamaños de los grupos de referencia y focal y en la tabla 23 aparecen las diferencias medias entre las estimaciones de los mismos parámetros para el grupo focal y de referencia ($\hat{b}_f - \hat{b}_R$ y $\hat{a}_f - \hat{a}_R$) calculadas sobre las réplicas dentro de cada condición experimental. Las medias de las estimaciones de discriminación estuvieron entre 0.5 y 0.65 con desviaciones estándar entre 0.22 y 0.32 mientras que las estimaciones del parámetro de dificultad tuvieron medias entre -0.12 y 0.02 con desviaciones estándar entre 0.81 y 0.94. Además cuando se calcularon las diferencias entre las estimaciones de los parámetros para los grupos focal y de referencia, se encontraron valores bastante cercanos a 0 para ambos parámetros para ítems con DIF, para el parámetro de discriminación cuando los ítems tenían DIF uniforme, y para la dificultad cuando los ítems tenían DIF no uniforme; en los demás casos las diferencias medias estuvieron cerca al 0.5, con valores entre 0.364 y 0.575 para las diferencias

en dificultad y entre 0.389 y 0.596 para las de discriminación. Teniendo en cuenta las distribuciones seguidas para simular los parámetros de los ítems ($b \approx n(0,1)$ y $a \approx n(0.5;0.2)$), y los incrementos en los valores de los mismos para simular los diferentes tipos de DIF, estos resultados indican un excelente comportamiento de los datos simulados.

Tabla 22

Descriptivos de las estimaciones de los parámetros de los ítems según tamaño de los grupos en el estudio del Ji cuadrado de Lord

		Discriminación		Dificultad	
Tamaño de grupo		**Promedio**	**Desviación**	**Promedio**	**Desviación**
Referencia	500	0.53	0.23	-0.07	0.85
	1500	0.50	0.22	-0.07	0.91
Focal	100	0.65	0.28	-0.08	0.81
	125	0.61	0.27	0.02	0.85
	167	0.59	0.27	0.02	0.88
	200	0.60	0.28	-0.11	0.88
	250	0.59	0.29	-0.12	0.88
	300	0.56	0.28	-0.08	0.94
	375	0.55	0.28	-0.12	0.94
	500	0.55	0.29	-0.12	0.94
	600	0.55	0.30	-0.08	0.94
	750	0.56	0.30	0.00	0.93
	1500	0.59	0.32	-0.10	0.88
	Total	**0.58**	**0.29**	**-0.07**	**0.90**

La tabla 24 presenta los CME de las estimaciones y las correlaciones de Pearson entre éstas y los parámetros respectivos, para las estimaciones de los parámetros de los ítems. Tanto los CME como las correlaciones mostraron mayor precisión de las estimaciones del parámetro de discriminación que del de dificultad; sin embargo, todos los resultados fueron satisfactorios: en el primer caso los CME estuvieron por debajo de 0.05 y las correlaciones fueron superiores a 0.86, mientras que para las estimaciones de dificultad los CME estuvieron entre 0.39 y 0.496 y las correlaciones entre 0.79 y 0.84. De otra parte, en ambos casos las correlaciones aumentaron y los CME disminuyeron con el aumento del tamaño del grupo.

Los resultados son aún más satisfactorios para las estimaciones del parámetro de los individuos, que se presentan en la tabla 25. Los CME estuvieron entre 0.068 y 0,131 y las correlaciones entre 0.934 y 0.97. Además, los CME fueron consistentemente menores y las correlaciones consistentemente mayores en las condiciones experimentales con longitud de prueba mayor (100 ítems). El promedio de correlación entre theta y sus estimaciones fue de 0.938 en las condiciones con $k= 50$ y de 0.967 con $k=100$; similarmente, los promedios de los CME fueron 0.13 y 0.07 para los dos tamaños de prueba, respectivamente.

Tabla 23

Diferencias promedio entre los grupos en las estimaciones de los parámetros de los ítems dentro del estudio del Ji cuadrado de Lord

Condición experimental	Sin DIF		DIF uniforme		DIF no uniforme		DIF Mixto	
	a_f - a_r	b_f - b_r	a_f - a_r	b_f - b_r	a_f - a_r	b_f - b_r	a_f - a_r	b_f - b_r
1	0.077	-0.058	0.038	**0.425**	**0.493**	-0.016	**0.520**	**0.418**
2	0.079	-0.042	0.059	**0.430**	**0.539**	-0.026	**0.521**	**0.419**
3	0.052	0.039	0.021	**0.566**	**0.389**	0.055	**0.453**	**0.535**
4	0.050	0.058	0.007	**0.575**	**0.436**	0.061	**0.461**	**0.541**
5	0.027	0.060	-0.028	**0.515**	**0.422**	-0.007	**0.410**	**0.548**
6	0.026	0.055	-0.025	**0.519**	**0.426**	0.013	**0.426**	**0.568**
7	0.028	-0.070	0.027	**0.403**	**0.453**	-0.107	**0.476**	**0.371**
8	0.028	-0.081	0.023	**0.409**	**0.454**	-0.114	**0.463**	**0.386**
9	0.019	-0.087	-0.002	**0.383**	**0.528**	-0.107	**0.489**	**0.411**
10	0.017	-0.094	0.002	**0.393**	**0.501**	-0.111	**0.474**	**0.409**
11	-0.019	-0.084	-0.044	**0.416**	**0.489**	-0.117	**0.455**	**0.370**
12	-0.020	-0.099	-0.049	**0.416**	**0.468**	-0.127	**0.421**	**0.396**
13	0.019	-0.045	-0.004	**0.410**	**0.523**	-0.086	**0.488**	**0.427**
14	0.015	-0.052	-0.005	**0.427**	**0.503**	-0.090	**0.449**	**0.444**
15	0.011	-0.085	0.009	**0.407**	**0.463**	-0.114	**0.473**	**0.392**
16	0.014	-0.090	-0.001	**0.423**	**0.470**	-0.116	**0.468**	**0.408**
17	0.011	0.032	-0.005	**0.407**	**0.482**	0.063	**0.446**	**0.419**
18	0.015	0.041	0.006	**0.507**	**0.513**	-0.050	**0.439**	**0.399**
19	0.012	-0.041	0.009	**0.432**	**0.527**	-0.073	**0.506**	**0.424**
20	0.011	-0.050	0.017	**0.449**	**0.514**	-0.068	**0.491**	**0.440**
21	0.015	0.034	0.006	**0.507**	**0.523**	0.020	**0.522**	**0.521**
22	0.015	0.032	0.018	**0.524**	**0.513**	0.028	**0.511**	**0.537**
23	0.040	-0.065	0.049	**0.364**	**0.573**	-0.080	**0.593**	**0.379**
24	0.040	-0.073	0.053	**0.385**	**0.573**	-0.081	**0.596**	**0.384**
Total	0.023	-0.035	0.008	**0.408**	**0.449**	-0.053	**0.443**	**0.407**

Tabla 24

CME y correlaciones de las estimaciones de los parámetros de los ítems según tamaño de grupo, en el estudio del Ji cuadrado de Lord

Tamaño del grupo	Discriminación				Dificultad			
	CME		Correlación		CME		Correlación	
	Media	D.S.	Media	D.S.	Media	D.S.	Media	D.S.
100	0.049	0.004	0.86	0.001	0.496	0.112	0.79	0.032
125	0.037	0.004	0.86	0.002	0.460	0.104	0.80	0.034
167	0.028	0.003	0.88	0.000	0.447	0.109	0.80	0.037
200	0.027	0.003	0.90	0.001	0.452	0.112	0.81	0.033
250	0.026	0.004	0.91	0.002	0.440	0.114	0.82	0.034
300	0.018	0.003	0.91	0.002	0.421	0.107	0.83	0.034
375	0.016	0.002	0.92	0.001	0.424	0.111	0.83	0.034
500	0.015	0.003	0.93	0.001	0.414	0.113	0.83	0.035
600	0.014	0.002	0.94	0.001	0.399	0.112	0.84	0.036
750	0.014	0.002	0.94	0.000	0.390	0.108	0.83	0.037
1500	0.019	0.002	0.95	0.001	0.398	0.116	0.84	0.037

Tabla 25

CME y correlaciones de las estimaciones del parámetro de los individuos dentro de cada condición experimental en el estudio del Ji cuadrado de Lord

Condición experimental	CME		Correlación	
	Media	Desviación	Media	Desviación
1	0.131	0.017	0.943	0.008
2	0.073	0.007	0.970	0.003
3	0.130	0.016	0.942	0.007
4	0.072	0.007	0.969	0.003
5	0.130	0.014	0.939	0.007
6	0.071	0.005	0.967	0.003
7	0.127	0.013	0.940	0.006
8	0.068	0.006	0.969	0.003
9	0.127	0.014	0.940	0.007
10	0.068	0.005	0.968	0.002
11	0.129	0.012	0.937	0.006
12	0.069	0.005	0.967	0.002
13	0.130	0.011	0.934	0.006
14	0.069	0.004	0.966	0.002
15	0.129	0.010	0.935	0.005
16	0.070	0.003	0.966	0.002
17	0.130	0.011	0.935	0.006
18	0.071	0.004	0.966	0.002
19	0.126	0.011	0.936	0.005
20	0.068	0.003	0.966	0.002
21	0.129	0.010	0.936	0.005
22	0.070	0.003	0.966	0.001
23	0.125	0.009	0.939	0.005
24	0.068	0.002	0.968	0.001

Factores que afectan el error tipo I del Ji cuadrado de Lord

La tabla 26 muestra los resultados del análisis de varianza para identificar los factores que afectan la tasa de falsos positivos del Ji cuadrado de Lord, para las diferentes categorías de ítems según su dificultad y discriminación. En todos los casos resultaron significativos los efectos principales del tamaño del grupo de referencia y la razón de tamaños y su interacción, mientras que la longitud de la prueba y su interacción con los otros factores no mostró efecto significativo. Los estadísticos descriptivos de las tasas de detecciones incorrectas con $\alpha=0.05$ para las dos longitudes de prueba fueron muy similares; con la prueba de 50 ítems éstas estuvieron entre 0 y 32% con media de 2.6% y desviación estándar de 3.6% y para la prueba de 100 ítems estuvieron entre 0 y 44% con media de 2.9% y desviación estándar de 3.9%. Por el contrario, en las condiciones experimentales con 500 examinados en el grupo de referencia el promedio de detecciones incorrectas estuvo entre 0 y 14% con media de 1.8% (D.E= 2.2%) y esta media se incrementó a 3.8 (D.E.=4.7%) con un máximo de 44%, cuando el tamaño del grupo de referencia aumentó a 1500 examinados. Por otra parte, las mayores tasas de detecciones incorrectas se observaron con gru-

pos de igual tamaño (*r=1*) y las mínimas con *r=2*. Con grupos iguales el porcentaje promedio de detecciones incorrectas fue de 5.5% (D:E=6.5%) con un máximo de 44% mientras que para *r=2* el promedio fue de 1.4 (D.E= 1.8%) con máximo de 8% y en los demás valores de *r* se observaron tasas entre 0 y 20% con medias entre 2.3% y 2.6%.

Tabla 26

F y significación del efecto de los diferentes factores sobre el error tipo I del Ji cuadrado de Lord

Factor	Dificultad baja		Dificultad media		Dificultad alta	
	F	Significación	F	Significación	F	Significación
Tamaño referencia	30.52	**0.00**	104.45	**0.00**	10.89	**0.001**
Razón de tamaños	6.77	**0.00**	33.15	**0.00**	4.80	**0.001**
Longitud de prueba	2.06	0.153	1.36	0.244	2.81	0.097
*r*k*	1.12	0.354	0.31	0.910	0.70	0.624
*r*nr*	10.92	**0.00**	34.46	**0.00**	4.03	**0.00**
*k*nr*	2.06	0.15	0.59	0.44	2.17	0.14

Factor	Discriminación baja		Discriminación media		Discriminación alta	
	F	Significación	F	Significación	F	Significación
Tamaño referencia	15.91	**0.000**	171.56	**0.000**	34.79	**0.000**
Razón de tamaños	20.06	**0.000**	50.34	**0.000**	8.45	**0.000**
Longitud de prueba	2.80	0.095	0.05	0.819	0.02	0.896
*r*k*	0.21	0.960	0.34	0.891	0.37	0.870
*r*nr*	4.54	**0.00**	45.21	**0.00**	10.02	**0.00**
*k*nr*	0.53	0.47	0.00	0.96	0.04	0.85

En el anexo 9 se muestran los promedios de FP del Ji cuadrado en cada condición experimental para los tres valores de α (0.01, 0.05 y 0.1) y en la figura 20 se presentan las tasa medias con $\alpha = 0.05$ según el tamaño del grupo de referencia y la razón de tamaños. Cuando el grupo mayoritario tenía 500 examinados las tasas medias de FP se mantuvieron dentro de los intervalos de Bradley, (1978) (entre 2.5% y 7.5%) independientemente de la razón de tamaños; el incremento del grupo de referencia a 1500 examinados representó una inflación del error tipo I solamente cuando los dos grupos eran iguales, y además, independientemente del tamaño del grupo mayoritario, las menores medias de detecciones falsas se presentaron cuando la razón de tamaños era igual a 2. De otra parte, cuando se calcularon los porcentajes de ítems sin DIF con tasas de detección por encima del intervalo de Bradley, (1978) para $\alpha = 0.05$ se observó un comportamiento similar. En las dos condiciones con grupos iguales de 1500 examinados más del 55% de los ítems sin DIF fueron falsamente detectados en más del 7.5% de las réplicas, mientras que en las condiciones con la razón de tamaños igual a 2, no se observaron tasas fuera del intervalo para la prueba de 50 ítems y menos de un 3% para la prueba de 100 ítems (ver tabla 27).

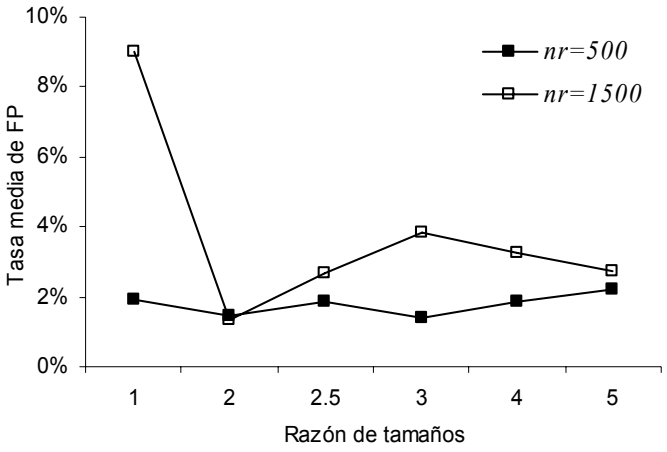


Figura 20. Medias de FP del Ji cuadrado de Lord con $\alpha = 0.05$ según tamaño del grupo de referencia y razón de tamaños.

Tabla 27

Porcentaje de ítems sin DIF con tasas de detección por encima de 7.5% con $\alpha = 0.05$ en cada condición experimental

Razón de tamaños	Longitud de Prueba	*nr=500*	*nr=1500*
1	50	2.3	55.8
	100	4.5	55.7
2	50	0.0	0.0
	100	2.3	1.1
2.5	50	7.0	7.0
	100	3.4	10.2
3	50	0.0	16.3
	100	1.1	18.2
4	50	2.3	11.6
	100	2.3	11.4
5	50	2.3	2.3
	100	6.8	4.5

De otra parte se encontró una correlación de 0.26 ($p<0.01$) entre la discriminación del ítem y la tasa de detección con $\alpha = 0.05$ y de -0.17 ($p<0.01$) entre ésta y la dificultad. Sin embargo, como puede verse en el anexo 10, para todas las categorías de dificultad el error tipo I se mantuvo dentro de los límites esperados con grupos de referencia de 500 examinados y sólo se presentó una pequeña elevación con grupos iguales de 1500 examinados. Sin embargo, cuando se compararon los promedios de detecciones incorrectas por categoría de discriminación los resultados fueron diferentes: Para los ítems con baja discriminación todas los promedios estuvieron dentro del intervalo esperado, para los ítems con discriminación media el único valor fuera del intervalo se presentó con grupos iguales de 1500 personas mientras que para los ítems con alta discriminación se presentaron elevaciones con *r=1*, *r=3* y *r=4*, llegando al casi el 25% con grupos iguales. En todo caso, las tasas más bajas se observaron con *r=2*.

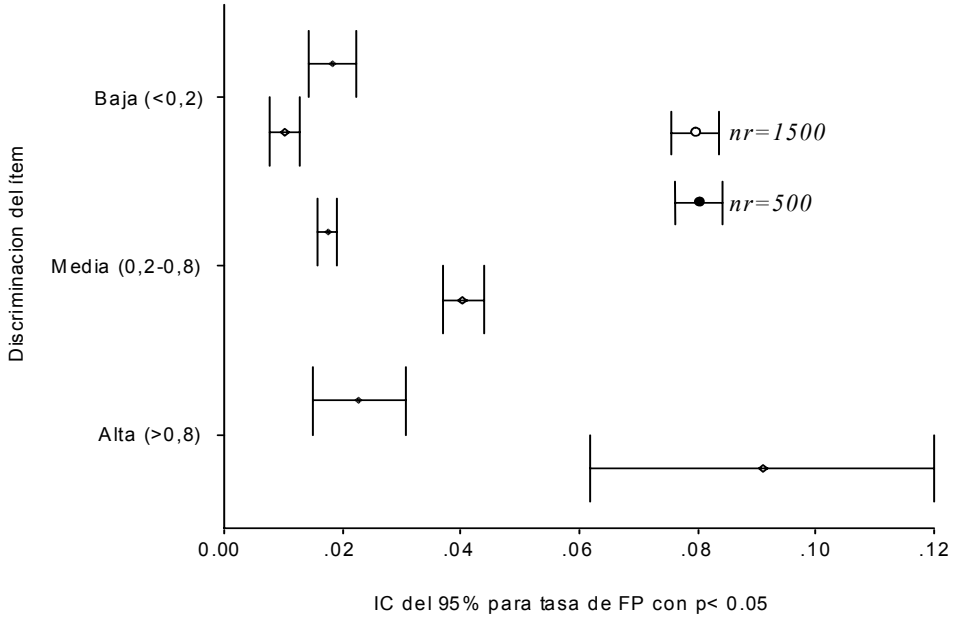


Figura 21. Intervalos del 95% de confianza para las medias de FP del Ji cuadrado de Lord en las diferentes categorías de ítems según su discriminación.

Finalmente, la figura 21 presenta los intervalos del 95% de confianza para el promedio de detecciones incorrectas en las diferentes categorías de discriminación de los ítems. Aunque dentro de cada categoría los promedios son significativamente diferentes para los dos tamaños del grupo de referencia, las tasas de detecciones incorrectas presentaron una elevación importante solamente cuando con ítems de alta discriminación y grupo de referencia de 1500 examinados.

Factores que afectan la potencia del Ji cuadrado de Lord

Los resultados de los análisis de varianza para identificar los factores que afectan el poder del Ji cuadrado de Lord (tabla 28) mostraron que tanto el tamaño del grupo de referencia como la razón de tamaños y su interacción tuvieron efecto significativo sobre la tasa de detección en todos los tipos de DIF, mientras que la longitud de la prueba afectó significativamente la tasa de detección de DIF no uniforme y mixto. En la tabla 29 puede verse que para todos los tipos de DIF el promedio de la tasa de detección fue mayor para mayor tamaño del grupo de referencia, disminuyó sistemáticamente con el aumento de la razón de tamaños para DIF uniforme y no uniforme, y presentó ligeras variaciones con la razón de tamaños para DIF mixto. Sin embargo, la tasa promedio de detección de DIF uniforme fue igual para las dos longitudes de prueba, aumentó un 3% con el número de ítems para DIF mixto, y un 6% para DIF no uniforme, y tuvo efecto significativo para estos últimos tipos de DIF.

En el anexo 9 se muestran los promedios de las tasas de detección de los tres tipos de DIF en cada una de las condiciones experimentales con los tres niveles de significación y además, en la figura 22 se presentan los promedios de las tasas de detección con $\alpha = 0.05$ de los tres tipos de

DIF en función de las variables manipuladas en el estudio. Estos resultados mostraron diferencias en el poder del estadístico para detectar los tres tipos de DIF. Las tasas de detección de DIF uniforme con grupo de referencia de 500 examinados estuvieron entre 31%, con $r=5$, y 79% con grupos iguales; cuando el tamaño del grupo de referencia creció a 1500, las tasas promedio se mantuvieron por encima de 90% para $r<3$ y estuvieron entre 65% y 74% en las demás condiciones. De otra parte, las tasas de detección más bajas se encontraron para DIF no uniforme con 500 examinados en el grupo de referencia, éste promedio fue inferior al 10% cuando $r>2.5$ y aunque tuvo un crecimiento importante con los demás valores de r, sólo llegó a 83% con grupos iguales. Finalmente, las tasas medias de detección de DIF mixto fueron bastante altas (entre 97% y 100%) en todas las condiciones con grupo de referencia de 1500 examinados y cuando el grupo de referencia tenía 500 personas siempre y cuando la razón de tamaños fuera menor de 5, los mínimos valores (61% y 69%) se encontraron en las dos condiciones con el máximo valor de r.

Tabla 28

F y significación del efecto de los diferentes factores sobre la potencia del Ji cuadrado de Lord

Factor	DIF uniforme		DIF no uniforme		DIF mixto	
	F	Significación	F	Significación	F	Significación
Tamaño referencia (*nr*)	125.27	**0.000**	1041.55	**0.000**	75.63	**0.000**
Razón de tamaños (*r*)	19.58	**0.000**	55.40	**0.000**	13.49	**0.000**
Longitud de prueba (*k*)	0.02	0.902	9.99	**0.003**	7.06	**0.010**
*r*k*	0.05	0.998	0.10	0.992	0.38	0.857
*r*nr*	11.26	**0.00**	34.95	**0.00**	11.72	**0.00**
*k*nr*	0.02	0.89	0.10	0.75	4.22	0.04

Tabla 29

Estadísticos descriptivos de la tasa de detección del Ji cuadrado de Lord según los factores manipulados.

	Uniforme			No uniforme			Mixto		
	Mínimo	Media	Máximo	Mínimo	Media	Máximo	Mínimo	Media	Máximo
nr=500	12	58	90	0	29	92	58	89	100
nr=1500	52	83	100	62	91	100	94	100	100
r=1	72	85	96	68	89	100	90	98	100
r=2	48	79	100	32	69	98	82	97	100
r= 2.5	38	71	100	22	63	100	62	93	100
r=3	52	71	88	0	52	100	84	97	100
r=4	52	70	84	4	45	94	82	97	100
r=5	12	49	86	0	42	98	58	82	100
k=50	16	71	100	0	56	100	58	92	100
k=100	12	71	100	0	62	100	64	95	100

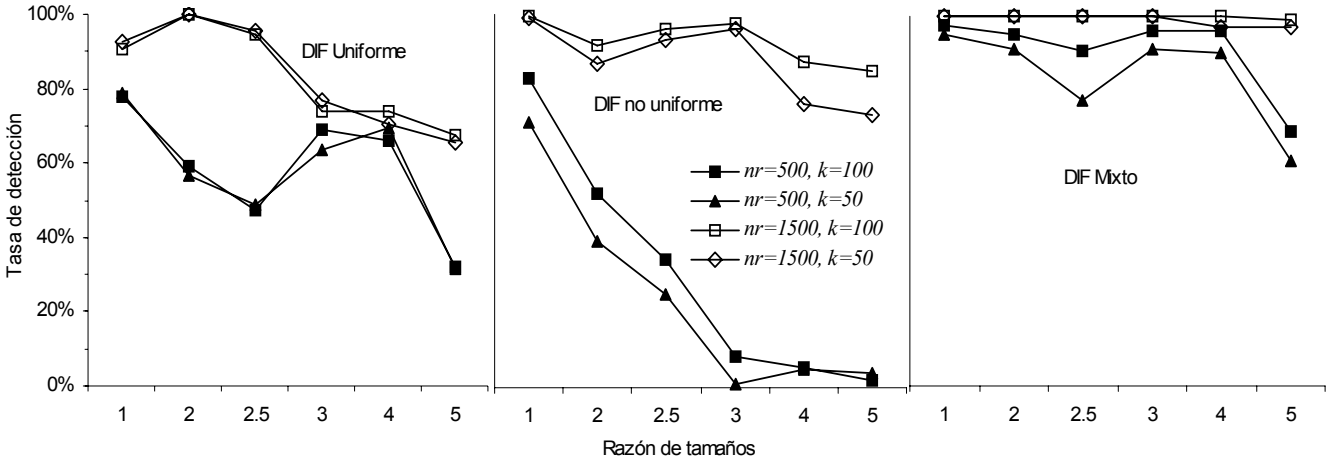


Figura 22. Tasas medias de detección del Ji cuadrado de Lord con $\alpha = 0.05$ para los diferentes tipos de DIF en función de nr, r y k.

Finalmente, no se encontraron asociaciones ni variaciones importantes de las tasas de detección del estadístico relacionadas con los valores de los parámetros de los ítems. En el anexo 11 aparecen las tasas de detección de cada uno de los ítems con DIF, en función de los tamaños de los grupos; se puede observar que el comportamiento del poder del estadístico es similar entre los ítems con el mismo tipo de DIF sin importar los valores de sus parámetros.

Discusión y Conclusiones

En primer lugar, las estadísticas descriptivas de las estimaciones de los parámetros, las diferencias en dichas estimaciones entre los grupos de referencia y focal, sus correlaciones con los respectivos parámetros y los CME, mostraron una satisfactoria recuperación de los parámetros tanto de los ítems como de los individuos. Este resultado no solamente constituye una base adecuada para la investigación sino que resulta interesante teniendo en cuenta el desajuste de los modelos de dos parámetros utilizados aquí.

De otra parte, la longitud de prueba no tuvo efecto significativo sobre el error tipo I calculado como la media de falsas detecciones a través de las 50 réplicas; en promedio, el porcentaje de FP sólo aumento 0,3% cuando la prueba cambió de 50 a 100 ítems. La longitud de prueba se incluyó como factor dentro del presente estudio por cuanto algunos resultados previos (Cohen, Kane, & Kim, 2001) han mostrado que ésta puede afectar la precisión de las estimaciones de los parámetros de los individuos; sin embargo, aunque en efecto parece haber algún cambio en la calidad de las estimaciones de theta con el número de ítems (el promedio de correlación entre theta y sus estimaciones fue ligeramente mayor y el de los CME ligeramente inferior para $k=100$), los resultados fueron satisfactorios para ambas longitudes de prueba y este factor no afectó el funcionamiento del CHI cuadrado de Lord en la detección de DIF.

Por el contrario, el tamaño del grupo de referencia, la razón de tamaños y su interacción afectaron significativamente la tasa de error tipo I del Ji cuadrado. Sin embargo, las tasas medias de falsas detecciones se mantuvieron dentro de los intervalos de Bradley, (1978)[22] y por debajo de los límites nominales en la mayoría de condiciones experimentales; solamente se presentaron importantes inflaciones con grupos iguales de 1500 examinados ($nr=1500$ y $r=1$). Aunque los resultados no son directamente comparables, Kim y Cohen, 1994 y Hidalgo Montesinos & López Pina (2002) habían reportado hallazgos consistentes por cuanto encontraron tasas de error tipo I por debajo de los límites nominales con tamaños hasta de 1000 por grupo; sin embargo, en ninguno de estos estudios se incluyeron tamaños más grandes. Este resultado tiene interés aplicado por cuanto la mayor preocupación del usuario cuando se enfrenta a la opción de utilizar los procedimientos basados en la TRI, es el tamaño de muestra mínimo necesario para obtener estimaciones adecuadas de los parámetros, prestando menos atención a los tamaños de muestra grandes que pueden conllevar a inflaciones importantes del error tipo I con los consecuentes costos que ellos implica en la construcción de pruebas. De acuerdo con los hallazgos de la presente investigación con tamaños de muestra de 125 examinados ya se podrían tener estimaciones aceptables ($CME<0.5$ y $r_{xy}>0.8$) de los parámetros de los ítems, y el error tipo I del Ji cuadrado para la detección de DIF se puede mantener controlado con un grupo hasta de 1500 personas si el otro tiene la mitad de examinados; con tamaños mayores se pueden esperar inflaciones importantes.

De otra parte, un resultado que puede resultar interesante dada la carencia de estudios sobre el tópico, lo constituyen las correlaciones estadísticamente significativas entre la tasa de falsa detección y los parámetros de dificultad y discriminación de los ítems. Sin embargo, tales correlaciones fueron de baja magnitud y, con excepción de dos valores ligeramente altos con grupos iguales de 1500 examinados, no se encontraron inflaciones importantes del error tipo I para diferentes niveles de dificultad (ver anexo 10.1). Por el contrario, con los ítems más discriminativos y los grupos de mayor tamaño, el error tipo I llegó a 24.3%, en esta misma condición experimental la tasa de FP fue elevada para los ítems con discriminación media, y con razones de tamaño de 3 y 4 se presentaron tasas ligeramente altas (ver anexo 10.2). Aunque los valores de los parámetros de los ítems utilizados aqu{i no fueron son muy heterogéneos y el propósito de la presente investigación no se centró en el efecto de los mismos, este resultado llama la atención sobre la necesidad de estudiar el comportamiento del error tipo I del Ji cuadrado, en función de los parámetros de los ítems.

Los resultados sobre el comportamiento del poder del estadístico en función de los factores manipulados fueron bastante similares a los reportados hasta ahora. En primer lugar, la longitud de la prueba no tuvo efecto importante sobre el funcionamiento del estadístico; su poder para de-

[22] Entre 2.5% y 7.5% para $\alpha = 0.05$ y entre 0.5% y 1.5% para $\alpha = 0.01$

tectar DIF uniforme se mantuvo inmodificable cuando cambió la longitud de la prueba y, aunque este factor sí resultó significativo para detectar DIF no uniforme y mixto, las diferencias en las tasas de detección resultan despreciables en la práctica. Debe hacerse notar, sin embargo, que las longitudes de prueba utilizadas en el presente estudio pueden estar un poco por encima de las utilizadas frecuentemente en la práctica y que con 50 ítems ya se pueden obtener estimaciones bastante satisfactorias del parámetro de habilidad. Es necesario realizar mayor investigación sobre este particular, utilizando longitudes de prueba más pequeñas con el fin de establecer los valores por debajo de los cuales el funcionamiento del Ji cuadrado de Lord pueda resultar afectado.

De otra parte, aunque en la literatura no abundan los reportes de investigación sobre el efecto del tamaño de los grupos sobre el Ji cuadrado de Lord, no resulta sorprendente que tanto el tamaño del grupo de referencia como la razón de tamaños y su interacción hayan mostrado efecto significativo sobre la tasa de detección de los diferentes tipos de DIF. Lo que sí resulta interesante es el comportamiento de las tasas de detección de los diferentes tipos de DIF en las diferentes condiciones experimentales. En primer lugar, con un grupo de referencia de 1500 examinados pueden obtenerse resultados satisfactorios independientemente del tipo de DIF, las tasas de detección se mantuvieron por encima del 80% con ligeros descensos para DIF uniforme con razón de tamaños mayor de 2.5. Sin embargo, cuando el grupo de referencia desciende a 500 examinados, el poder del estadístico cambia en función del tipo de DIF y la razón de tamaños: Para DIF uniforme las tasas de detección se mantuvieron altas (por encima del 60% con $\alpha=.05$) con grupos iguales o razón de tamaños de 3 o 4; la tasa de detección de DIF no uniforme fue alta para grupos iguales, descendió drásticamente con $r=2$ y continuó descendiendo regularmente hasta llegar a niveles muy bajos (entre 1% y 3% con $\alpha=.05$) con $r=5$; y para DIF mixto solamente se observó un descenso cuando $r=5$.

En principio estos resultados permitirían concluir que el Ji cuadrado de Lord parece más efectivo para detectar DIF mixto; sin embargo, debe anotarse que la superioridad del estadístico para detectar este tipo de DIF en comparación con el uniforme y el no uniforme puede explicarse por la magnitud, más que por el tipo de DIF. Aunque este factor no se manipuló sistemáticamente en el presente estudio, la estrategia utilizada para simular el DIF tuvo como resultado que el DIF mixto se asociara con mayor magnitud del mismo (ver anexo 1). En consecuencia, estos resultados no admiten este tipo de comparación y la atención debe centrarse en las variaciones en función de los tamaños de los grupos dentro del mismo tipo de DIF.

Para efectos prácticos puede concluirse que con un grupo de 1500 examinados pueden esperarse tasas bastante adecuadas de detección de cualquier tipo de DIF aún si el otro grupo es hasta cinco veces menor (300 examinados). Si el grupo mayoritario tiene 500 examinados, se podrán esperar tasas modestas de detección de DIF uniforme cuando el grupo minoritario sea 2, 2.5 o 5 veces menor; e igualmente la tasa de detección de DIF mixto será moderada cuando la razón de tamaños sea igual a 5; con tamaños de grupo más grandes se pueden esperar tasas de detección

satisfactorias para estos dos tipos de DIF. Con este mismo tamaño del grupo mayoritario, el Ji cuadrado de Lord solamente tendrá una tasa de detección alta de DIF no uniforme si los grupos son iguales, con grupos minoritarios entre 200 y 250 la efectividad del estadístico es bastante modesta y su uso no resulta recomendable con grupos de menor tamaño.

Obviamente, estos resultados son generalizables a situaciones en las que se tengan condiciones similares a las simuladas aquí en términos de longitudes de prueba, porcentaje de ítems con DIF en la misma, distribución de magnitud de atributo de los grupos, valores de los parámetros de los ítems y magnitudes de DIF. Además, teniendo en cuenta la carencia de investigación cuyos resultados sean comparables, al menos aproximadamente, con los hallazgos del presente estudio, hace falta mayor evidencia para evaluar la generalidad de los mismos.

Finalmente, vale la pena mencionar dos líneas de acción que, de acuerdo con los resultados del presente estudio, quedan abiertas para futuras investigaciones. En primer lugar, teniendo en cuenta que las magnitudes de DIF uniforme y no uniforme son similares, resulta interesante el bajo desempeño del estadístico para detectar éste último con grupos pequeños, en comparación con su poder para detectar DIF uniforme. Las estimaciones del parámetro de discriminación fueron bastante precisas a juzgar por los CME, y con grupos de 200 individuos ya se obtuvieron promedios de correlaciones de .9 entre las mismas y los parámetros; sin embargo, las tasas de detección de DIF no uniforme simulado incrementando la discriminación para el grupo de menor tamaño, fueron muy bajas para las condiciones con 500 examinados en el grupo de referencia y 200 o menos en el grupo focal, y la tasa promedio no alcanzó el 50% con grupo minoritario de 250 individuos. Además, las estimaciones del parámetro de dificultad, aunque aceptables, fueron de menor calidad y sin embargo, las tasas de detección de DIF uniforme fueron en general, superiores a las de no uniforme, aún con grupos pequeños. Aunque este resultado no permite atribuir el pobre desempeño del estadístico a posibles deficiencias en las estimaciones de los parámetros y parece respaldar el efecto de la razón de tamaños, es necesaria mayor investigación que manipule y evalúe el efecto de ambos factores.

El segundo tópico que vale la pena investigar es el posible efecto del desajuste de los modelos TRI en la efectividad del Ji cuadrado de Lord para detectar DIF. Lim & Drasgow (1990) encontraron un aceptable funcionamiento del estadístico tanto en sus tasas de FP como de detecciones correctas, cuando se introduce un desajuste en los modelos simulando datos no unidimensionales; Sin embargo, no se ha estudiado el desajuste de los modelos en términos del número de parámetros de los mismos como el que se introdujo en el presente estudio, y la robustez del estadístico para detectar los diferentes tipos de DIF. Vale la pena diseñar estudios que permitan estimar en qué medida los actuales resultados pueden atribuirse a tal desajuste.

DISCUSION, CONCLUSIONES E IMPLICACIONES

Desde que en la jerga psicométrica se adoptaran términos como “sesgo”, “funcionamiento diferencial de los ítems” (popularizado como DIF) y “funcionamiento diferencial de los tests” (DTF), la producción académica sobre el tema ha ido en permanente ascenso. En un estudio bibliométrico que cubrió la producción de artículos científicos publicados en el último cuarto del siglo XX, Gómez Benito, Hidalgo Montesinos, Guilera Ferré & Moreno Torrente (2005) encontraron que más del 80% de tales publicaciones se hicieron en la última década (1991 a 2000) y dentro de éstos casi las tres cuartas partes se publicaron en el último quinquenio. Este ascenso no resulta sorprendente si se entiende desde una perspectiva no sólo académica sino también social y política: la preocupación cada vez más generalizada por garantizar la igualdad de oportunidades y el tratamiento equitativo de los individuos y grupos sociales, dentro de las diferencias individuales y culturales. Lo que puede resultar sorprendente es que todavía algunas entidades responsables de procesos de evaluación, cuyos resultados puedan tener implicaciones en la asignación de cupos u oportunidades educativas o laborales, no hayan incorporado dentro de sus procesos una estrategia para detectar el posible sesgo en los instrumentos que diseña o utiliza. La revisión bibliográfica dentro del presente estudio reveló que además de las condiciones prácticas como el costo computacional y el acceso a la tecnología (equipos y programas) necesaria para tal incorporación, una de las posibles dificultades que debe afrontar el usuario de los métodos es la poca claridad sobre el funcionamiento de algunos de los muchos procedimiento disponibles hoy, en determinadas condiciones de tamaño de grupos, longitud de prueba o distribución de magnitud de atributo, entre otras. Los resultados de los estudios realizados dentro del presente trabajo pueden ser de utilidad a la hora de decidir sobre el uso de estos procedimientos o de elegir alguno de ellos dadas unas condiciones prácticas concretas.

La presente investigación tuvo como objetivo general analizar, a través de tres estudios empíricos, el posible efecto del tamaño del grupo de referencia y la razón de tamaños de los grupos en el funcionamiento de tres procedimientos actualmente utilizados para detectar DIF de ítems dicótomos: el MH, la RL y el Ji cuadrado de Lord. Aunque se pueden identificar algunas limitaciones de los estudios empíricos y por tanto de los hallazgos reportados, una primera conclusión es que el objetivo general se ha cumplido en los términos en que se planteó. Se diseñaron tres estudios en los que se manipularon sistemáticamente los factores de interés, se simularon bases de datos que permitieron ver el efecto de tales factores, se identificaron y midieron las variables respuesta y se adelantaron los análisis de datos que permitieron evaluar la calidad de los mismos, el efecto de los factores de interés y el comportamiento de las tasas de detección correctas (para evaluar la potencia de cada estadístico) y falsas (para evaluar su error tipo I) en diferentes condiciones que se pueden presentar en la práctica profesional. Además, la estrategia metodológica que se siguió a través de los diferentes estudios facilita la comparación entre los métodos evaluados en cada uno

de los estudios, lo que permite extraer conclusiones de utilidad para el potencial usuario de los mismos. Sin embargo, antes de entrar en la discusión de los hallazgos propiamente dichos es necesario enmarcar el ámbito de generalización de los mismos, dadas las estrategias metodológicas utilizadas, el procedimiento de simulación de los datos, el diseño de investigación, las variables manipuladas y los valores o niveles de las mismas.

En primer lugar, la decisión de abordar el problema mediante experimentos de Monte Carlo reporta ventajas pero también impone algunas restricciones. El principal acierto de esta elección, coherente con lo planteado por Gentle, (2003), Ross, (1999) y Spence, (1983), es que esta aproximación metodológica permitió evaluar experimentalmente el efecto de los factores de interés sobre la potencia de los estadísticos y su error tipo I, medidas a través del porcentaje de detecciones ocurridas para cada uno de los ítems a través de 100 réplicas en diferentes condiciones de tamaños de grupos. La manipulación de los valores de los parámetros de los ítems para los dos grupos comparados permitió simular el DIF en un número de ítems previamente identificados a la vez que se ejercía control sobre el tipo de DIF, la magnitud del mismo y el porcentaje de éstos dentro de la prueba; manipulación y control imposibles en estudios con datos empíricos en los que el sesgo se induce experimentalmente. Sin embargo, esta ganancia en control experimental dentro de situaciones simuladas partiendo de procesos aleatorios tiene algunos costos que están representados en las exigencias de recursos, el carácter aproximado de las soluciones encontradas y la limitada validez externa de los estudios, entre otros (Cohen, Kane & Kim, 2001; Serlin, 2000; Skrondal, 2000; Spence, 1983 y Dorn & Greenberg, 1970). Aunque, como se discutirá más adelante, los resultados de los estudios empíricos aquí reportados evidencian un adecuado manejo técnico de los procedimientos de generación de los vectores aleatorios y de reducción de varianza, no puede dejar de mencionarse que los hallazgos solamente tienen generalidad a ámbitos en los cuales las condiciones reales sean similares a las aquí simuladas. Las condiciones que pueden resultar más relevantes a la hora de juzgar la utilidad práctica de estos resultados son las distribuciones de los parámetros de los ítems (sobretodo de a y b), la igualdad en la distribución de la magnitud de atributo entre los grupos focal y de referencia, y la longitud de la prueba, sobretodo en los dos primeros estudios.

De otra parte, aunque algunos análisis se llevaron a cabo separadamente para diferentes tipos de DIF o valores de los parámetros, utilizando, en este último caso, unas categorizaciones arbitrarias, el diseño experimental en todos los estudios fue factorial completamente cruzado con dos factores sistemáticamente manipulados y analizados: tamaño del grupo de referencia y razón de tamaños de los grupos definida como el número de examinados del grupo de referencia por cada uno del grupo focal ($r = {nr}/{nf}$); en el tercer estudio se manipuló, además, la longitud de la prueba. Los análisis en los que se consideraron algunas de las variables como los valores de los parámetros de los ítems y el tipo o magnitud de DIF, tuvo como único propósito controlar su potencial

efecto considerando los resultados de estudios previos sobre el tema, a la vez que permitieron analizar relaciones entre las mismas y las variables dependientes y, en algunos casos, formular hipótesis sobre su posible efecto. En consecuencia, sin desconocer la potencial importancia de muchos otros factores, como bien puede evidenciarse a partir de la revisión bibliográfica, y respondiendo a los objetivos del presente trabajo, los hallazgos de los estudios empíricos permiten dos tipos de conclusiones: a) evaluar el efecto de los factores manipulados en cada estudio sobre el comportamiento del estadístico correspondiente comparando las tasas de detección en las diferentes condiciones experimentales, y b) comparar el efecto de los dos factores (nr y r) y el comportamiento de las tasas de detección entre los diferentes estadísticos estudiados. La mayoría de conclusiones del primer tipo se presentaron en el capítulo correspondiente a cada estudio empírico y aquí se hace énfasis en las del segundo tipo.

Finalmente, los valores fijados para las variables manipuladas también imponen ciertas limitaciones a la hora de generalizar los resultados de los estudios realizados. Se fijaron dos tamaños para el grupo de referencia (500 y 1500) y seis para la razón de tamaños (1, 2, 2.5, 3, 4, y 5); esta elección tuvo dos ventajas y una limitación. El cruce de los dos factores resultó en 12 condiciones experimentales que permitieron analizar una amplia gama de tamaños del grupo focal, desde un grupo tan pequeño como 100 examinados hasta uno de tamaño considerable como el de 1500, al tiempo que se analizaban situaciones también diversas en cuanto a la razón de tamaños: desde grupos iguales ($r=1$) hasta un grupo focal considerablemente minoritario ($r=5$). Todas estas condiciones pueden presentarse en la práctica y en ese sentido los hallazgos reportan utilidad aplicada. Sin embargo, como consecuencia de este mismo diseño la razón de tamaños quedó confundida con el tamaño del grupo focal, lo que hace necesario que los hallazgos de los estudios se repliquen independizando los dos factores y evaluando separadamente sus efectos.

Dentro de este marco general, un primer resultado de interés está en la evaluación de la calidad de las estimaciones de los parámetros a través de sus CME y correlaciones con los parámetros. Los datos fueron simulados siguiendo un modelo TRI de tres parámetros con $c=0.2$ pero las estimaciones se realizaron ajustando un modelo 3P, en los dos primeros estudios, y un modelo 2P en el tercero (ver anexos 2.1 y 8). La decisión de elegir esta estrategia se justificó en el capítulo anterior y representó algunas ventajas importantes pero introdujo un desajuste en los datos analizados en el tercer estudio, factor que debe tenerse en cuenta a la hora de analizar los resultados. Y es en este contexto que resultan interesantes los resultados de la evaluación de la calidad de los datos. Aunque todos fueron suficientemente satisfactorios para justificar el uso de las bases de datos generadas dentro de los estudios, cuando se ajustaron modelos 2P los CME de las estimaciones de dificultad crecieron y las correlaciones disminuyeron en comparación con los de las estimaciones ajustando modelos 3P. En el anexo 12 se muestran las medias de los CME y de las correlaciones en función del tamaño del grupo para los dos tipos de modelos ajustados; como es

de esperarse a la luz de estudios anteriores (Cervantes-Botero & Herrera, 2005; Tejada, 2002; Cohen & Kim, 2001) en todos los casos se observó una pérdida de precisión de las estimaciones con la disminución del tamaño del grupo, sin que afectara el potencial uso de los datos en la práctica. Además, la calidad de las estimaciones del parámetro de discriminación y de habilidad fue comparable cuando se ajustaron los dos tipos de modelos; aunque se observaron diferencias en los CME y las correlaciones, éstas fueron muy pequeñas en el caso de a y prácticamente despreciables para el parámetro θ. Sin embargo, tales diferencias fueron de mayor magnitud para el parámetro de dificultad; como puede verse en el anexo 12, el desajuste introducido al suponer falsamente $c=0$ parece tener mayor impacto en la estimación del parámetro de dificultad en comparación con los otros dos parámetros y en comparación con la situación en la cual no se supone un valor c constante.

En segundo lugar, los resultados referentes al comportamiento del error tipo I de los tres estadísticos evaluados mostraron que el tamaño del grupo de referencia, al menos en los niveles estudiados aquí, afectó significativamente el error tipo I del MH y el Ji cuadrado, y no lo hizo con la RL. Para el primer estadístico este efecto resultó significativo con ítems de dificultad baja o media y para el segundo, en todas las categorías de ítems. El efecto de la razón de tamaños, sin embargo, resultó significativo para los tres estadísticos y en casi todas las categorías de ítems, exceptuando dos categorías de ítems con baja discriminación en el MH y los ítems con discriminación media en la RL. Resultados similares se observaron en relación con el efecto de interacción entre estos dos factores, éste resultó significativo para los ítems con dificultad media en el MH, las categorías de ítems de alta dificultad para la RL y todas las categorías para el Ji cuadrado de Lord.

Esta significación estadística debe analizarse, sin embargo, teniendo en cuenta su impacto en la práctica (Rudas & Zwick, 1997; Stark, Chernyshenko & Drasgow, 2004). En el anexo 13 se presentan las tasas de detección falsas (anexo 13.1) y correctas (anexos 13.2 a 13.4) con $\alpha=0.05$ de los tres estadísticos para los diferentes tamaños de muestra y razones de tamaños. Si se acepta el límite superior de los intervalos propuestos por Bradley, (1978) como el valor máximo admisible para identificar condiciones en las cuales puede haber una inflación importante del error tipo I, se puede concluir que solamente cuando los grupos son iguales a 1500 ($nr=1500$ y $r=1$) examinados el MH y el Ji cuadrado pueden presentar alguna inflación, mientras que para la RL el error tipo I se mantiene controlado aunque ligeramente superior a los valores nominales. Con 500 examinados en el grupo de referencia las mayores tasas de FP se observaron para la RL y las menores para el Ji^2, en todas las razones de tamaños. Sin embargo, cuando el grupo de referencia tenía 1500 examinados las tasas de FP fueron similares para el MH y la RL con valores ligeramente superiores a los nominales, mientras que para el Ji^2 las tasas se mantuvieron por debajo de los valores nominales para todas las razones y presentaron un drástico incremento con grupos de igual tama-

ño, llegando a 8.8% con $\alpha = 0.05$. De acuerdo con estos resultados si el usuario de los métodos tiene un grupo mayoritario de hasta 1500 examinados y un grupo minoritario de tamaño similar a los estudiados aquí, puede tener la tranquilidad de que cualquiera de los tres estadísticos mantendrá controlado el error tipo I, sin embargo si los grupos comparados son de igual de tamaño (1500 examinados), resulta preferible la RL.

Pero no todos los ítems tienen similar probabilidad de resultar falsamente detectados como sesgados. Para todos los estadísticos se observaron algunas relaciones entre las tasas de FP y los valores de los parámetros de los ítems. La tasa de FP del MH se asoció positivamente con la discriminación y la fuerza de tal asociación aumentó con el tamaño de los grupos; sin embargo sólo se presentaron tasas ligeramente superiores a 7.5 para los ítems más discriminativos y grupos de máximo tamaño (1500 y 1500 o 1500 y 750). En lo que se refiere a la RL, su tasa de FP se asoció significativamente con la dificultad del ítem; tal correlación fue inversa y modesta, y sólo ocurrió con grupos de igual tamaño; en estas condiciones experimentales (grupos iguales de 1500 y grupos iguales de 500) se presentaron ligeras elevaciones por encima de los límites de Bradley (1978) para ítems con baja dificultad. Finalmente, la tasa de FP del Ji^2 se asoció positivamente con la discriminación y negativamente con la dificultad de los ítems, a pesar de su significación estadística, ambas correlaciones fueron modestas; sin embargo, en el caso de la dificultad sólo se observó una ligera elevación para los ítems más fáciles y con grupos de máximo tamaño, mientras que en los ítems de mayor discriminación se observó la inflación importante con grupos iguales de 1500 examinados. Aunque en general estas correlaciones fueron modestas o bajas apoyan la hipótesis del posible efecto de los valores de los parámetros de los ítems sobre el error tipo I de los estadísticos, que ya ha sido formulada a partir de los hallazgos de estudios anteriores como los de Rogers & Swaminathan (1993), Uttaro & Millsap (1994), Narayanan & Swaminathan (1994), Clauser, Mazor & Hambleton (1994) y Narayanan & Swaminathan (1996) entre otros. Pero además, el hecho de que las asociaciones sólo se presenten para algunas condiciones experimentales sugiere una posible interacción entre los valores de los parámetros y los factores manipulados en los estudios; interacción que no se ha explorado aún y que debe evaluarse en futuros estudios.

De otra parte, los resultados referentes al comportamiento de la potencia de los estadísticos para detectar diferentes tipos de DIF mostraron que en todos los casos tanto el tamaño del grupo mayoritario como la razón de tamaños y su interacción, tienen efecto significativo sobre la tasa de detecciones correctas; la única excepción se presentó para el efecto de interacción sobre la potencia de MH con DIF no uniforme (ver tablas 13, 19 y 28). Nuevamente, este resultado exige un examen de las tasas de detección en las diferentes condiciones experimentales para cada estadístico y tipo de DIF, con el fin de establecer la importancia práctica de esta significación estadística. En los anexos 13.2 a 13.4 se muestran las tasas medias de detecciones correctas con $\alpha=.05$, de los tres estadísticos estudiados para los tres tipos de DIF.

Un primer resultado de implicaciones prácticas está en la diferencia en el comportamiento de los tres estadísticos para detectar DIF uniforme. Mientras la potencia del MH y del Ji^2 se mantuvo en valores altos o aceptables, las tasas de detección de la RL estuvieron por debajo del 25% y sólo llegó al 40% con grupos iguales de 1500 examinados cada uno. Pero además, también se observaron diferencias entre los dos primeros estadísticos. La potencia del MH descendió sistemáticamente con el aumento de la razón de tamaños y este descenso fue más importante con *nr=500*, mientras que las variaciones de la potencia del Ji cuadrado en función de *r* fueron más drásticas y menos sistemáticas. Cuando el grupo de referencia tenía 1500 examinados la potencia del MH se mantuvo por encima del 95% independientemente de *r*, y la del Ji cuadrado presentó un descenso por debajo de este límite cuando la razón de tamaños paso de 2.5 a 3, manteniéndose entre el 67% y el 76 en las condiciones con $r \geq 3$. Con 500 examinados en el grupo mayoritario la potencia del MH cayó de 96% al 35% cuando *r* pasó de 1 a 5 y sólo se mantuvo por encima de 80% cuando *r<3*; el Ji cuadrado por su parte tuvo tasas de detección inferiores a las del MH en estas últimas condiciones experimentales, y muy similares cuando $r \geq 3$. Esta comparación favorece ampliamente el uso del MH sobre la RL cuando las condiciones prácticas sean similares a las simuladas aquí, y sobre el Ji cuadrado en algunas condiciones. Si el grupo de referencia tiene alrededor de 1500 examinados y la razón de tamaños es menor de 3, cualquiera de los dos estadísticos (MH o Ji cuadrado) resulta recomendable, con grupos de tamaño pequeño (*nr* alrededor de 500 y *r* de 3 o más) los dos estadísticos muestran potencia bastante modesta, en las demás condiciones resulta preferible usar el MH.

Por el contrario las tasas de detección de DIF no uniforme desfavorecen el uso del MH en comparación con los otros dos métodos en casi todas las condiciones experimentales. Con grupos grandes (*nr=1500*) tanto la RL como el Ji cuadrado mantuvieron tasas de detección satisfactoriamente altas (80% o más) mientras que las del MH estuvieron entre 38% y 74%. Con menores tamaños de grupo de referencia la RL mostró las mayores tasas de detección pero sólo fueron moderadas cuando *r>2*; el Ji cuadrado tuvo tasas de detección mayores que las del MH con *r>3* y muy similares en las demás condiciones. De acuerdo con estos resultados la RL es el procedimiento más recomendable para detectar DIF no uniforme en comparación con los otros dos procedimientos, sin embargo, con grupo mayoritario de 500 y razón de tamaños mayores de 2, solo cabe esperar tasas de detección moderadas.

El comportamiento más similar entre los tres estadísticos se observó en las tasas de detección de DIF mixto. En general la potencia de los procedimientos fue bastante satisfactoria en la mayoría de condiciones experimentales. Con 1500 examinados en el grupo de referencia las tasas de detección fueron mayores al 80% con excepción del Ji cuadrado en las condiciones con *r=5*; aún con grupos de menor tamaño (*nr=500*) estas tasas se mantuvieron bastante altas y solo descen-

dieron a niveles apenas aceptables con $r=5$. Estos resultados no son sorprendentes teniendo en cuenta que la magnitud de DIF mixto se simuló aumentando por igual los parámetros de dificultad y de discriminación y en consecuencia el área entre las CCI es considerablemente mayor que la de los ítems con DIF uniforme y no uniforme. Así, cuando la magnitud de DIF sea tan alta como la simulada aquí para el DIF mixto (áreas entre 0.63 y 0.79, ver anexo 1) cualquiera de los tres procedimientos mostrará tasas bastante satisfactorias incluso con un grupo mayoritario de 500 examinados y una razón de tamaños menor de 5.

En síntesis, tanto el tamaño del grupo de referencia como la razón de tamaño tuvieron efecto importante sobre el error tipo I y la potencia de los tres estadísticos estudiados; sin embargo, solamente se presentaron inflaciones de las tasas de FP con grupos iguales de máximo tamaño para el MH y el Ji cuadrado. Además, la RL fue prácticamente incapaz de detectar DIF uniforme mientras que el MH mostró muy bajo poder para detectar el no uniforme y no se observaron diferencias en la potencia de los estadísticos para detectar DIF mixto. La diferencia entre la RL y el MH había sido reportada y explicada por Swaminathan & Rogers (1990). “La RL está diseñada para detectar DIF no uniforme y en consecuencia puede no ser efectiva para detectar DIF uniforme. Inversamente, el MH es diseñado para detectar DIF uniforme y puede no ser efectivo para detectar el No Uniforme” (pag. 366). Si se acepta que en la práctica “el DIF no uniforme se presenta con mucho menos frecuencia que el uniforme” (Jodoin & Gierl, 2001, p. 332), estos resultados parecen favorecer el uso del MH sobre los otros dos estadísticos en la mayoría de condiciones similares a las simuladas aquí. Aunque el comportamiento del Ji cuadrado fue también satisfactorio en la mayoría de las condiciones, tiene un comportamiento más inconsistente a lo largo de los diferentes valores de r y sus costos computacionales son mayores que los del MH.

Finalmente, los análisis de los efectos de los factores sistemáticamente manipulados en los diferentes trabajos permiten concluir que la razón de tamaños tiene un efecto importante que puede tener implicaciones prácticas sobretodo cuando ésta es mayor de 3, lo que hace suponer que con valores más extremos éste efecto puede ser dramático. Los diferentes hallazgos han permitido identificar condiciones en las cuales los estadísticos estudiados tienen un mejor desempeño y, desde este punto de vista se espera aportar en el cuerpo de conocimientos sobre el comportamiento de tales procedimientos a la vez que se sientan las bases para formular nuevas hipótesis de trabajo sobre todo en lo referente a condiciones más extremas y a la interacción entre los tamaños de los grupos y los valores de los parámetros de los ítems.

REFERENCIAS

Ackerman, T. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. *Journal of Educational Measurement. 29*(1), 67-91.

Anastasi, A. (1974). *Test psicológicos*. Madrid: Morata.

Anastasi, A. & Urbina, S. (1998). *Tests psicológicos*. (7 ed.) México: Prentice Hall.

Andrich, D. A. (1998). Theory precedes measurement. *Rasch Measurement Transactions. 12*(2).

Angoff, W. H. (1972). *A technique for the investigation of cultural differences*. Documento presentado en el Annual meeting of the American Psychological Association: Honolulu.

Angoff, W. H. (1982). Use of Difficulty and Discrimination Indices for Detecting Item Bias. En R.A.Berck (Ed.), *Handbook of Methods for Detecting Test Bias*, pp. 96-116. Baltimore: The Jons Hopkins University Press.

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, N. J.: Educational Testing Service.

Angoff, W. H. (1988). Validity: An evolving concept. En H.Wainer & H. I. Braun (Eds.), *Test Validity*, pp. 19-32. New Yersey: Lawrence Erlbaum Associates Inc.

Angoff, W. H. (1993). Perspectives on Differential Item Functioning Methodology. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, New Jersey: Lawrence Erlbaum Associates, Inc.

Angoff, W. H. & Ford, S. F. (1973). Item race interaction on a test of scholastic aptitude. *Journal of Educational Measurement. 10*, 95-105.

Baker, F. B. (1996). An investigation of the sampling distributions of equating coefficients. *Applied Psychological Measurement. 20*, 45-57.

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement. 28*(147), 162.

Baker, F. B. (1981). A Criticism of Scheuneman´s Item Bias Techinique. *Journal of Educational Measurement. 18*(1), 59-62.

Baker, F. B. (1995). *Equate computer Program Version 2.1*. Madison, Wisconsin: Laboratory of Experimental Design. Department of Eductinal Psychology. University of Wisconsin.

Bennet, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped group. *Journal of Educational Measurement. 24*, 41-55.

Binet, A. & Simon, T. (1916). *The development of intelligence in children*. New York: Amo.

Birnbaum, A. (1958). *Further considerations of efficiency in test of a mental ability* (Rep. No. 17). Rabdolph Air Force, Texas: USAF School of Aviation Medicine.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a examinee's ability. En F.M.Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*, Reading, Mass: Addison-Wesley.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge: MIT Press.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika. 46*, 443-449.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement. 25*, 275-285.

Bolt, D. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika. 23*(1), 67-95.

Bond, L. (1987). The Golden Rule settlement: A minority perspective. *Educational Measurement: Issues and Practice. 6*, 18-20.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (Ed) (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement. 26* (4), 433-450.

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

Brown, F. (1980). *Fundamentos de Medición y Evaluación en Psicología y Educación*. México: El Manual Moderno.

Camilli, G. (1992). A Conceptual Analysis of Differential Item Functioning in Terms of a Multidimensional Item Response Model. *Applied Psychological Measurement. 16*(2), 12-147.

Camilli, G. (1993). The Case against Item Bias Detection Techniques Based on Internal Criteria: Do Item Bias Procedures Obscure Test Fairness Issues? En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, pp. 397-417. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Camilli, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. 4. United States: SAGE Publications.

Candell, G. & Drasgow, F. (1988). An Iterative Procedure for Linking Metrics and Assessing Item Bias in Item Response Theory. *Applied Psychological Measurement. 12*(3), 253-260.

Cervantes-Botero, V. H. & Herrera, A. N. (2005). Un estudio de Monte Carlo sobre la precision del alfa de Cronbach. *Escrito enviado para publicación.*

Christensen, R. (1997). *Log-Linear Models and Logistic regression*. (2 ed.) New York: Springe.

Clauser, B. E. (1993). *Factors influencing the performance of the Mantel-Haenszel procedure in identifying differential item functioning*. Ambherst. MA: University of Massachusetts.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effect of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education. 6*(4), 269-279.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1994). The Effects of Score Group Width on the Mantel-Haenszel Procedure. *Journal of Educational Measurement. 31*(1), 67-78.

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

Cohen, A. S., Kane, M. T., & Kim, S.-H. (2001). The Precision of Simulation Study Results. *Applied Psychological Measurement. 25*(2), 136-145.

Cohen, A. S. & Kim, S.-H. (1993). A Comparison of Lord´s X^2 and Raju´s Area Measures in Detection of DIF. *Applied Psychological Measurement. 17*(1), 39-52.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of Differential Item Functioning in the Graded Response Model. *Applied Psychological Measurement. 17*(4), 335-350.

Cohen, A. S., Kim, S.-H., & Subkoviak, M. J. (1991). Influence of prior distributions on detection of DIF. *Journal of Educational Measurement. 28*, 49-59.

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

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1980). Valididty on parole: How can we go straight? En W.B.Schrader (Ed.), *New directions for testing and measurement -Measuring achievement: Progress over a decade.*, pp. 99-108. San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five pespectives on validity argument. En H.Wainer & H. I. Braun (Eds.), *Test validity*, pp. 3-17. New Jersey: Lawrence Erlbaum Associates, Inc.

Cuesta, M. (1996). Unidimensionalidad. En J.Muñiz (Ed.), *Psicometría*, pp. 239-291. Madrid: Editorial Universitas S.A.

Cuesta, M. & Muñiz, J. (1994). Utilización de los modelos de teoría de respuesta a los ítems con datos multifactoriales. *Psicothema. 6*(2), 283-296.

Cuesta, M. & Muñiz, J. (1995). Efectos de la multidimensionalidad en la estimación de parámetros desde modelos unidimensionales de teoría de respuesta a los ítems. *Psicológica. 16*(1), 65-86.

Divgi, D. R. (1985). A minimum chi-square methods for developing a common metric in IRT. *Applied Psychological Measurement. 9*(4), 413-415.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo Study of Factor That affect the Mantel-Haenszel and Standardization Measures of Differential Item Functioning. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, pp. 137-166. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publisher.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education. 2*, 217-233.

Dorans, N. J. & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, New Jersey: Lawrence Erlbaum Associates, Inc.

Dorans, N. J. & 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 (Research Rep. No 83-9)*. Princeton, NJ: Educational Testing Service.

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

Dorn, W. S. & Greenberg, H. J. (1970a). *Matematicas y computacion con programacion FORTRAN*. Trabajo original publicado en 1967. Mexico D.F., Mexico: Editorial Limusa Wiley, S.A.

Dorn, W. S. & Greenberg, H. J. (1970b). *Matematicas y computacion: con programacion FORTRAN*. ((Trabajo original publicado en 1967) ed.) Mexico D.F., Mexico: Editorial Limusa Wiley, S.A.

Drasgow, F. & Parsons, C. K. (1993). Aplication of unidimensional item response theory model to multidimensional data. *Applied Psychological Measurement. 7*(2), 189-199.

Drasgow, F. (1987). Study of the Measurement Bias of Two Standardized Psychological Tests. *Journal of Applied Psychology. 72*(1), 19-29.

Eelles, K., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. (1951). *Intelligence and cultural differences*. Chicago: University of Chicago Press.

Faggen, J. (1987). Golden Rule revisited: Introduction. *Educational Measurement: Issues and Practice. 6*, 5-8.

Ferreres, T. (1998). *Funcionamiento diferencial de los items de una prueba de aptitud intelectual en función de la lengua familiar y la lengua de escolarización..* (Tesis doctoral inédita ed.) Valencia: Universidad de Valencia.

Fidalgo, Á. M. (1996a). Funcionamiento Diferencial de los Items. En J.Muñiz (Ed.), *Psicometría*, pp. 371-455. Madrid: Editorial Universitas, S.A.

Fidalgo, Á. M. (1996b). *Funcionamiento diferencial de los ítems. Procedimiento Mantel-Haenszel y modelos loglineales.* Tesis doctoral no publicada. Oviedo: Universidad de Oviedo.

Fidalgo, Á. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel ´procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement. 64*(6), 925-936.

Fidalgo, Á. M. & Mellenbergh, G. J. (1995). *Evaluación del procedimiento Mantel-Haenszel frente al método logit iterativo en la detección del funcionamiento diferencial de los ítems uniforme y no uniforme.* Comunicación presentada al IV Simposio de Metodología de las Ciencias del Comportamiento. La Manga del Mar Menor.

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

Fidalgo, Á. M. & Paz Caballero, M. D. (1995). Modelos lineales logarítmicos y funcionamiento diferencial de los ítems. *Anuario de Psicología. 1995*(64), 57-66.

Flanagan, J. C. (1951). Units, scores, and norms. En E.F.Lindquist (Ed.), *Educational measurement*, pp. 695-763. Washington, DC: American Council on Education.

French, A. W. & Miller, T. (1996). Logistic Regression and its Use in Detecting Differential Item Functioning in Polytomous Items. *Journal of Educational Measurement. 33*(3), 315-333.

García Cueto, E. (1993). *Introducción a la Psicometría*. México: Siglo veintiuno editores.

Gentle, J. E. (2003). *Random number generation and Monte Carlo methods*. New York: Springer-Verlag.

Ghiselli, E. E. (1964). *Theory of psychological measurement*. New York: McGraw Hill.

Gómez Benito, J. & Hidalgo Montesinos, M. D. (1997a). *A Comparison of Two Procedures of Ability Purification on the Detection of Differential Item Functioning Using Multinomial Logistic Regression*. Poster presented at the 10th European Meeting of the Psychometric Society. Santiago de Compostela, Spain.

Gómez Benito, J. & Hidalgo Montesinos, M. D. (1997b). Evaluación del funcionamiento diferencial en items dicotómicos: una revisión metodológica. *Anuario de Psicología.*(74), 3-32.

Gómez Benito, J., Hidalgo Montesinos, M. D., Guilera Ferré, G., & Moreno Torrente, M. (2005). A bibliometric study of differential item functioning. *Scientometrics. 64*(1), 3-16.

Gómez Benito, J. & Navas-Ara, M. J. (1996). Detección del funcionamiento diferencial de los ítems mediante regresión logística: Purificación paso a paso de la habilidad. *Psicológica. 17*, 397-411.

Gómez Benito, J. & Navas-Ara, M. J. (2000). A Comparison of $X^{2, RFA}$ and IRT Based Procedures in the Detection of DIF. *Quality & Quantity. 34*, 17-31.

Gottfredson, L. S. (Ed) (1997). Intelligence and social policy. *Intelligence. 24* (Special issue), 1-320.

Green, D. R. (1975). What does it mean to say a test is biased? *Education and Urban Society. 8*, 33-52.

Green, D. R. & Draper, J. F. (1972). *Exploratory studies of bias in achievement test*. Paper presented at the Annual meeting of the American Psychological Association: Honolulu.

Gruijter, D. N. & Van der Kamp, L. J. (1984). *Statistical Models in psychological and educational testing*. Lisse: Swets and Zeitlinger.

Gulliksen, H. (1950). *Theory of mental test*. New York: John Wiley & Sons.

Haebara, T. (1980). Equatin logistic ability scales by weighted least squares methods. *Japanese psychological research. 22*, 144-149.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. (1993). Advances in the Detection of Differentially Functioning Test Items. *European Journal of Psychological Assessment. 9*(1), 1-18.

Hambleton, R. K. & Rogers, H. J. (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education. 2*(4), 313-334.

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

Harvey, R. & Hammer, A. (1999). Item Response Theory. *Counseling Psychologist. 27*(3), 353-383.

Herrera, A. N., Sánchez Pedraza, R., & Gómez Benito, J. (2001). Funcionamiento diferencial de los Items: Una revisión conceptual y metodológica. *Acta Colombiana de Psicología. 5*, 41-61.

Herrera, A. N., Sánchez, N., & Jiménez, H. (2001). De la Teoría Clásica de los Test a la Teoría de Respuesta a los Items. En J.I.Ruiz Pérez, E. Ponce de León Díaz, A. N. Herrera, N. Sánchez, H. Jiménez, & E. Medellín Lozano (Eds.), *Avances en Medición y evaluación en Psicología y Educación: cinco lecturas selectas*, pp. 293-332. Bogotá: Universidad El Bosque.

Hidalgo Montesinos, M. D. & Gómez Benito, J. (1996). *The Effect of Ability Purification on the Evaluation of Differential Item Funtioning with the Technique of Multinomial Logistic Regression*. Paper presented at the 20th biennial conference of the Society for Multivariate Analysis in the Behavioral Sciences. Barcelona, Spain.

Hidalgo Montesinos, M. D. & Gómez Benito, J. (2003). Test Purification and the Evaluation of Differential Item Functioning with Multinomial Logistic Regression. *European Journal of Psychological Assessment. 19*(1), 1-11.

Hidalgo Montesinos, M. D. & López Pina, J. A. (1997). Comparación entre las medidas de área, el estadístico de Lord y el análisis de regresión logística en la evaluación del funcionamiento diferencial de los ítems. *Psicothema. 9*, 417-431.

Hidalgo Montesinos, M. D. & López Pina, J. A. (2002). Two-stage equating differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement. 62*, 32-44.

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

Hills, J. (1990). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice. 8*, 5-11.

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

Holland, P. W. & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure (Tecnical report N° 86-89)*. Princeton, NJ: Educational Testing Service.

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

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

Holland, P. W. (1989). A note on the covariance of the Mantel-Haenszel log-odds estimator and the sample marginal rates. *Biometrics. 45*, 1009-1015.

Hong, S. (2001). An investigation of the influence of internal test bias on regresion slope. *Applied Measurement in Education. 14*(4), 351-368.

Horst, P. (1966). *Psychological measurement and prediction*. Belmont, California: Cole Publishing Company.

Ironson, G. (1982). Use of Chi-square and Latent Trait Approaches for Detectin Item Bias. En *Handbook of Methods for Detecting Test Bias*, pp. 117-160.

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

Jensen, A. R. (1974). How biased are culture-loaded test? *Genetic Psychology Monographs. 90*, 185-244.

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

Jensen, A. R. (1998). The *g* factor in the design of education. En R.J.Sternberg & W. M. Williams (Eds.), *Intelligence, instruction, and assessment*, pp. 111-131. Mahwah, NJ: Erlbaum.

Jensen, A. R. (2000). TESTING. The Dilemma of Group Differences. *Psychology, Public Policy, and Law. 6*(1), 121-127.

Jodoin, M. G. & Gierl, M. J. (Ed) (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.

Jodoin, M. G. & Huff, K. L. (2001). *Examining type I error and power rates when ability distribution are unequal with the logistic regresion procedure for DIF detection*. Paper presented at Annual metting of the National Council on Measurement in Education. Seattle.

Kaskowitz, G. S. & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function methods of linking. *Applied Psychological Measurement. 25*(1), 39-52.

Kelderman, H. (1989). Item Bias Detection Using Loglinear IRT. *Psychometrika. 54*(4), 681-697.

Kim, S.-H. & 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. & Cohen, A. S. (1992). Effects on linking methods on detection of DIF. *Journal of Educational Measurement. 29*, 51-66.

Kim, S.-H. & Cohen, A. S. (1994). An Investigation of Lord´s Procedure for the Detection of Diferential Item Functioning. *Applied Psychological Measurement. 18*(3), 217-228.

Kim, S.-H. & Cohen, A. S. (1995). A Comparison of Lord´s Chi-Square, Raju´s Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning. *Applied Measurement in Education. 8*(4), 291-312.

Kleinbaum, D. G. (1994). *Logistic regression. A self learning text*. New York: Springer.

Knoke, D. & Burke, P. J. (2000). *Log-linear models*. Beverly Hills: SAGE.

Kok, F. (1988). Item bias and test multidimensionality. En R.Langeheine & J. Rost (Eds.), *Latent trait and latent class models*, pp. 263-274. New York: Plenum.

Kok, F. G., Mellenbergh, G. J., & Van Der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement. 22*, 295-303.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer.

Kolen, M. J. (2004). Linking Assessments: Concept and History. *Applied Psychological Measurement. 28*(4), 219-226.

Kulick, E. & Dorans, N. J. (1983a). *Assessing the unexpected differential item functioning of oriental candidates on SAT form CSAG and TSWE Form E33 (Statictical Rep. No 83-106)*. Princeton, NJ: Educational Testing Service.

Kulick, E. & Dorans, N. J. (1983b). *Assessing the unexpected differential item performance of candidates reporting different levels of father's education on SAT form CSA2 and TSWE Form E29 (Statictical Rep. No 83-27)*. Princeton, NJ: Educational Testing Service.

Lautenschlager, G., Flaherty, V. L., & Park, D.-G. (1994). IRT Differential Item Functioning: An Examination of Ability Scale Purifications. *Educational and Psychological Measurement. 54*(1), 21-31.

Lautenschlager, G. & Park, D.-G. (1988). IRT Item Bias Detection Procedures; Issues of Model Misspecification, Robustness, and Parameter Linking. *Applied Psychological Measurement. 12*(4), 365-376.

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

Lim, R. G. & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement: Issues and Practice. 6*, 13-17.

Lim, R. G. & Drasgow, F. (1990). Evaluation of Two Methods for Estimating Item Response Theory Parameters When Assessing Differential Item Functioning. *Journal of Applied Psychology. 75*(2), 164-174.

Linn, R. L. & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement. 18*, 109-118.

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

López Pina, J. A., Hidalgo, M. D., & Sánchez Meca, J. (1993). Error tipo I de las pruebas Chi-cuadrado en el estudio del sesgo de los items. En C.Arce & G. Seoane (Eds.), *III Simposium de metodología de las ciencias sociales y del comportamiento*, pp. 521-529.

Lord, F. M. (Ed) (1952). A theory of test scores. *Psychometric Monographs.* (7).

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

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

Lord, F. M. & Novick, M. R. (1966). *Statistical theories of mental test scores*. Reading, MA: Adison Wesley.

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

Marco, G. L. (1988). Does the use of item assembly procedures proposed in legislation make any difference in test properties and test performance of black and white test takers? *Applied Measurement in Education. 1*, 109-133.

Marco, G. L. (1977). Item Characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement. 14*(139), 160.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic. *Educational and Psychological Measurement. 52*, 443-451.

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

McCauley, C. D. & Mendoza, J. (1985). A Simulation Study of Item Bias Using a Two-parameters Item Response Model. *Applied Psychological Measurement. 9*(4), 389-400.

McDonald, R. P. (1997). Normal-Ojiva multidimensional model. En W.J.van der Linder & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, pp. 258-269. New York: Springer.

McDonald, R. P. (2000). A basic for multidimensional item response theory. *Applied Psychological Measurement. 24*(2), 99-114.

McLaughlin, M. E. & Drasgow, F. (1987). Lord´s chi-square test of item bias with estimated and with know person parameters. *Applied Psychological Measurement. 11*, 161-173.

Mellenbergh, G. J. (1982a). Contingency table models for assesing item bias. *Journal of Educational Statistics. 7*, 105-118.

Mellenbergh, G. J. (1982b). Contingency table models for assessing item bias. *Journal of Educational Statistics. 7*, 105-118.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research. 13*, 127-143.

Meredith, W. (Ed) (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika. 58*, 525-543.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. En H.Wainer & H. I. Braun (Eds.), *Test validity*, pp. 33-45. New Jersey: Lawrence Erlbaum Associates, Inc.

Miller, M. D. & Oshima, T. C. (1992). Effect of Sample Size, Number of Biased Items, and Magnitude of Bias on a Two-Stage Item Bias Estimation Method. *Applied Psychological Measurement. 16*(4), 381-388.

Millsap, R. & Everson, H. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement. 17*(4), 297-334.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods. 2*, 248-260.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika. 51*, 177-195.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models [Computer program]*. Mooresville IN: Scientific Software.

Mosher, D. L. (1966). The development and multitrait-multimethod matrix analysis of three measures of three aspects of guilt. *Journal of consulting and clinical psychology. 30*(25), 29.

Mosher, D. L. (1968). Measures of guilt in females by self-report inventories. *Journal of consulting and clinical psychology. 32*(690), 695.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Items*. Madrid: Ediciones Pirámide.

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

Muñiz, J. & Hambleton, R. K. (1992). Medio siglo de Teoría de Respuesta a los ítems. *Anuario de Psicología. 52*, 41-66.

Muthén, B. & Lehman, J. (1985). Multiple group IRT modeling: Application to item bias analysis. *Journal of Educational Statistics. 10*(133), 141.

Nandakumar, R. (Ed) (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement. 30*, 293-311.

Narayanan, P. & 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. & Swaminathan, H. (1996). Identification of Items that Show Nonuniform DIF. *Applied Psychological Measurement. 20*, 257-274.

Navas Ara, M. J. (1996). Equiparación de puntuaciones. En J.Muñiz (Ed.), *Psicometría*, pp. 293-369. Madrid: Editorial Universitas, S. A.

Navas-Ara, M. J. & Gómez Benito, J. (1994). *Comparison of several bias detection techniques.* (Paper presented at the 23rd. International Congress of Applied Psychology ed.) Madrid.

Navas-Ara, M. J. & Gómez Benito, J. (2002a). Effects of Ability Scale Purification on the Identification of dif. *European Journal of Psychological Assessment. 18*(1), 9-15.

Navas-Ara, M. J. & Gómez Benito, J. (2002b). Effects of ability scale purification on the idnetification of DIF. *European Journal of Psychological Assessment. 18*(1), 9-15.

Nerlove, M. (1997). *Notes on monte carlo, booststrapping and estimation by simulation*. (Recuperado el 1 de Marzo de 2005, de http://www.arec.umd.edu/montecarlo/ ed.).

Núñez Núñez, R. M., Hidalgo Montesinos, M. D., & López Pina, J. A. (2000). Influencia de la igualación iterativa en la detección del funcionamiento diferencial del ítem mediante medidas de áres de Raju y el estadístico de Lord. *Psicothema. 12*(3), 495-502.

Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement. 25*(4), 373-383.

Park, D.-G. & Lautenschlager, G. (1990). Improving IRT Item Bias Detection with Iterative Linking and Ability Scale Purification. *Applied Psychological Measurement. 14*(2), 163-173.

Petersen, N. S. & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement. 13*, 3-29.

Phillips, A. & Holland, P. W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds ratio estimate. *Biometrics. 43*, 425-431.

Powers, D. A. & Xie, Y. (2000). *Statistical methods for catagorical data analysis*. San Diego: Academic Press.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in C: The art of scientic computing* . (2da ed.) Cambridge, Inglaterra: Cambridge University Press.

Prieto Marañón, P., Barbero Garcia, M. I., & San Luis Costas, C. (1997). Identification of Nonuniform Differential Item Functioning: A Comparison of Mantel-Haenszel and Item Response Theory Analysis Procedures. *Educational and Psychological Measurement. 57*(4), 559-568.

Raju, N. (1988). The Area Between Two Item Characteristic Curves. *Psychometrika. 53*(4), 495-502.

Raju, N. (1990). Determining the significance of estimated signed and unsigned areas between two item response function. *Applied Psychological Measurement. 14*, 197-207.

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

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. En W.J.van der Linder & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, pp. 271-286. New York: Springer.

Reynolds, C. R. (1982a). Methods for Detecting Construct and Predictive Bias. En R.A.Berck (Ed.), *Handbook of Methods for Detecting Test Bias*, pp. 199-227. Baltimore: Johns Hopkins University Press.

Reynolds, C. R. (1982b). The problem of bias in psychological assessment. En C.R.Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology*, pp. 178-208. New York: Wiley.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics. 42*, 311-323.

Rogers, H. J. & Hambleton, R. K. (1989). Evaluation of Computer Simulated Baseline Statistics for Use in Item Bias Studies. *Educational and Psychological Measurement. 49*, 355-369.

Rogers, H. J. & Swaminathan, H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement. 17*(2), 105-116.

Ross, S. M. (1999). *Simulacion*. Mexico D.F., Mexico: Prentice Hall Hispanoamericana.

Roussos, L. A. & Stout, W. F. (1996). Simulation Studies of the Effects of Small Sample Size an Studied Item Parameters on SIBTEST and Mantel-Haenszel Type I Error Performance. *Journal of Educational Measurement. 33*(2), 215-230.

Rudas, T. & Zwick, R. (1997). Estimating the Importance of Differential Item Functioning. *Journal of Educational and Behavioral Statistics. 22*(1), 31-45.

Rudner, L. M. (1977). *An approach to biased item identification using latent trait measurement theory*. Paper presented at Annual Meeting of the American Educational Research Association. New York.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980a). A Monte Carlo Comparison of Seven Biased Item Detection Techniques. *Journal of Educational Measurement. 17*(1), 1-10.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980b). Biased item detection techniques. *Journal of Educational Statistics. 5*, 213-233.

Scheuneman, J. D. (1979). A new method for assesing bias in test items. *Journal of Educational Measurement. 16*, 143-152.

Scheuneman, J. D. (1981). A response to Baker's criticism. *Journal of Educational Measurement. 18*, 63-66.

Segal, D. O. (1983). *Test characteristic curves, item bias and transformation to a common metric in IRT: A methodological artifact with serious consequences and a simple solution*. (Unpublished manuscript ed.) Illinois: University of Illinois.

Serlin, R. C. (2000). Testing for robustness in monte carlo studies. *Psychological Methods. 5*, 230-240.

Shealy, R. & Stout, W. F. (1989). *A procedure to detect test bias present simultaneously in several items*. San Francisco: Paper presented at Annual Meeting of the American Educational Research Association.

Shealy, R. & Stout, W. F. (1993a). 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.

Shealy, R. & Stout, W. F. (1993b). An item response theory model for test bias an differential test functioning. En P.W.Holland & H. Wainer (Eds.), *Differential item functioning*, Hilisdale, N. J.: LEA.

Shepard, L. A. (1982). Definitions of Bias. En R.A.Berck (Ed.), *Handbook of Methods for Detecting Test Bias*, pp. 9-30. Baltimore: Johns Hopkins University Press.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics. 6*, 317-375.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics. 9*, 93-128.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of Approximation Techniques for Detecting Item Bias. *Journal of Educational Measurement. 22*(2), 77-105.

Skrondal, A. (2000). Design and analysis of monte carlo experiments: attacking the conventional wisdom. *Multivariate Behavioral Research. 35*, 137-167.

Sobol, I. M. & Myshetskaya, E. E. (2003). Modelling correlated random variables. *MonteCarlo Methods and Applications. 9*, 67-76.

Spence, I. (1983). Monte Carlo Simulation Studies. *Applied Psychological Measurement. 7*(4), 405-425.

Spray, J. A. & Carlson, J. E. (1986). *Comparison of loglinear and logistic regression models for detecting changes in proportions*. Paper presented at Annual meeting of the American Educational Research Association. San Francisco.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of Differential Item (Functioning and Differential) Test Functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology. 89*(3), 497-508.

Stern, W. (1914). *The psychological methods of testing intelligence*. Baltimore: Warwick y York.

Stocking, M. & Lord, F. M. (1983). Developing a common metric in IRT. *Applied Psychological Measurement. 7*(2), 201-210.

Swaminathan, H. & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement. 27*(4), 361-370.

Tejada, J. (2002). *Desarrollo de un programa de computador para simular datos de ítemes que presenten funcionamiento diferencial.* Tesis de pregrado no publicada. Universidad Nacional de Colombia.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond Group-Mean Differences: The Concept of Item Bias. *Psychological Bulletin. 99*(1), 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use ot item response theory in the study of group differences in trace lines. En H.Wainer & H. I. Braun (Eds.), *Test validity*, pp. 147-169. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameter of item response models. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, pp. 67-113. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika. 47*, 397-412.

Thorndike, R. L. (1995). *Psicometría aplicada*. México: Limusa.

Uttaro, T. & Millsap, R. (1994). Factors Influencing the Mantel-Haenszel Procedure in the Detection of Differential Item Functioning. *Applied Psychological Measurement. 18*(1), 15-25.

van de Vijver, F. & Leung, K. (1997). *Methods and data analysis for Cross-cultural research.* London: SAGE Publication.

Van Der Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An iterative item bias detection methods. *Journal of Educational Measurement. 21*(2), 131-145.

Waller, N. G. (1998). EZDIF: Detection of Uniform and Nonuniform Differential Item Functioning with the Mantel-Haenszel and Logistic Regression Procedures. *Applied Psychological Measurement. 22*(4), 391.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning of Rach models. *The Journal of Experimental Education. 72*(3), 221-261.

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

Zieky, M. (1993). Practical Questions in the Use of DIF Statistics in Test Development. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, pp. 337-347. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

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

Zumbo, B. D. & Thomas, D. R. (1996). *A measure of DIF effect size using logistic regression procedure*. (Paper presented at the National Board of Medical Examiners ed.) Philadelphia.

Zwick, R. & Thayer, D. T. (2002). Application of an Empirical Bayes Enhancement of Mantel-Haenszel Differential Item Functioning Analysis to a Computerized Adaptive Test. *Applied Psychological Measurement. 26*(1), 57-76.

Zwick, R., Thayer, D. T., & Lewis, C. (1997). *An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis*. (ETS Research Report No. 97-21 ed.) Princeton, N.J: Educational Testing Service.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement. 36*(1), 1-28.

ANEXOS

ANEXO 1:

Parámetros de los ítems de la prueba simulada

Item	a_r	b_r	c_r	a_f	b_f	c_f	Dif^I	Area
1	0.629	1.18	0.2	0.629	1.18	0.2	0	0
2	0.535	1.468	0.2	0.535	1.468	0.2	0	0
3	0.672	-1.165	0.2	0.672	-1.165	0.2	0	0
4	0.86	0.587	0.2	0.86	1.087	0.2	1	0.40
5	0.32	-0.663	0.2	0.32	-0.663	0.2	0	0
6	0.813	1.229	0.2	0.813	1.229	0.2	0	0
7	0.119	0.697	0.2	0.119	0.697	0.2	0	0
8	0.448	-0.422	0.2	0.448	-0.422	0.2	0	0
9	0.802	-0.413	0.2	1.302	-0.413	0.2	2	0.31
10	0.00	0.999	0.2	0.00	0.999	0.2	0	0
11	0.499	-1.529	0.2	0.499	-1.529	0.2	0	0
12	0.588	-1.192	0.2	0.588	-1.192	0.2	0	0
13	0.427	0.007	0.2	0.427	0.007	0.2	0	0
14	0.601	-1.118	0.2	0.601	-1.118	0.2	0	0
15	0.27	0.297	0.2	0.27	0.297	0.2	0	0
16	0.113	0.928	0.2	0.113	0.928	0.2	0	0
17	0.618	0.244	0.2	0.618	0.244	0.2	0	0
18	0.701	0.48	0.2	1.201	0.98	0.2	3	0.79
19	0.798	1.885	0.2	0.798	1.885	0.2	0	0
20	0.579	0.297	0.2	0.579	0.297	0.2	0	0
21	0.079	0.527	0.2	0.079	0.527	0.2	0	0
22	0.424	-0.386	0.2	0.424	-0.386	0.2	0	0
23	0.786	0.828	0.2	0.786	0.828	0.2	0	0
24	0.858	-0.834	0.2	0.858	-0.334	0.2	1	0.40
25	0.782	-1.158	0.2	0.782	-1.158	0.2	0	0
26	0.461	-1.456	0.2	0.461	-1.456	0.2	0	0
27	0.541	-0.09	0.2	0.541	-0.09	0.2	0	0
28	0.525	-0.025	0.2	0.525	-0.025	0.2	0	0
29	0.253	-0.56	0.2	0.253	-0.56	0.2	0	0
30	0.119	-0.696	0.2	0.119	-0.696	0.2	0	0
31	0.022	0.362	0.2	0.022	0.362	0.2	0	0
32	0.287	-0.614	0.2	0.287	-0.614	0.2	0	0
33	0.225	-1.087	0.2	0.225	-1.087	0.2	0	0
34	0.081	1.725	0.2	0.081	1.725	0.2	0	0
35	0.887	-0.788	0.2	0.887	-0.788	0.2	0	0

Item	a_r	b_r	c_r	a_f	b_f	c_f	Dif^I	Area
36	0.095	2.294	0.2	0.095	2.294	0.2	0	0
37	0.568	-0.467	0.2	0.568	-0.467	0.2	0	0
38	0.726	-1.621	0.2	0.726	-1.621	0.2	0	0
39	0.369	2.795	0.2	0.369	2.795	0.2	0	0
40	0.597	-1.529	0.2	0.597	-1.529	0.2	0	0
41	0.428	1.498	0.2	0.428	1.498	0.2	0	0
42	0.704	-0.671	0.2	1.204	-0.171	0.2	3	0.78
43	0.761	0.395	0.2	0.761	0.895	0.2	1	0.40
44	0.47	0.526	0.2	0.47	0.526	0.2	0	0
45	0.107	0.481	0.2	0.107	0.481	0.2	0	0
46	0.632	-0.315	0.2	0.632	-0.315	0.2	0	0
47	0.555	0.2	0.2	0.555	0.2	0.2	0	0
48	0.291	-0.599	0.2	0.291	-0.599	0.2	0	0
49	0.841	-0.483	0.2	1.341	-0.483	0.2	2	0.30
50	0.442	0.61	0.2	0.442	0.61	0.2	0	0
51	0.374	-0.877	0.2	0.374	-0.877	0.2	0	0
52	0.824	-0.967	0.2	0.824	-0.967	0.2	0	0
53	0.861	0.474	0.2	1.361	0.974	0.2	3	0.68
54	0.402	-1.771	0.2	0.402	-1.771	0.2	0	0
55	0.261	0.367	0.2	0.261	0.367	0.2	0	0
56	0.245	-0.181	0.2	0.245	-0.181	0.2	0	0
57	0.339	1.435	0.2	0.339	1.435	0.2	0	0
58	0.339	1.102	0.2	0.339	1.102	0.2	0	0
59	0.537	-0.992	0.2	0.537	-0.992	0.2	0	0
60	0.622	0.458	0.2	0.622	0.458	0.2	0	0
61	0.512	-0.843	0.2	0.512	-0.843	0.2	0	0
62	0.806	1.667	0.2	0.806	1.667	0.2	0	0
63	0.189	-0.495	0.2	0.189	-0.495	0.2	0	0
64	0.588	-0.793	0.2	0.588	-0.793	0.2	0	0
65	0.429	-0.06	0.2	0.429	-0.06	0.2	0	0
66	0.608	-0.749	0.2	0.608	-0.749	0.2	0	0
67	0.576	-0.92	0.2	0.576	-0.92	0.2	0	0
68	0.433	0.265	0.2	0.433	0.265	0.2	0	0
69	0.466	-1.517	0.2	0.466	-1.517	0.2	0	0
70	0.658	-0.564	0.2	1.158	-0.564	0.2	2	0.43
71	0.562	-0.147	0.2	0.562	-0.147	0.2	0	0
72	0.417	-1.348	0.2	0.417	-1.348	0.2	0	0
73	0.755	-0.428	0.2	0.755	0.072	0.2	1	0.40
74	0.512	-2.078	0.2	0.512	-2.078	0.2	0	0

Item	a_r	b_r	c_r	a_f	b_f	c_f	Dif[1]	Area
75	0.292	-0.165	0.2	0.292	-0.165	0.2	0	0
76	0.096	-1.751	0.2	0.096	-1.751	0.2	0	0
77	0.736	-1.95	0.2	0.736	-1.95	0.2	0	0
78	0.579	-1.18	0.2	0.579	-1.18	0.2	0	0
79	0.625	0.629	0.2	0.625	0.629	0.2	0	0
80	0.612	0.59	0.2	0.612	0.59	0.2	0	0
81	0.338	0.54	0.2	0.338	0.54	0.2	0	0
82	0.815	0.012	0.2	1.315	0.012	0.2	2	0.30
83	0.966	-0.404	0.2	1.466	0.096	0.2	3	0.63
84	0.112	-0.358	0.2	0.112	-0.358	0.2	0	0
85	0.33	-0.634	0.2	0.33	-0.634	0.2	0	0
86	0.506	-0.116	0.2	0.506	-0.116	0.2	0	0
87	0.575	-0.235	0.2	0.575	-0.235	0.2	0	0
88	0.385	0.032	0.2	0.385	0.032	0.2	0	0
89	0.682	-0.859	0.2	0.682	-0.859	0.2	0	0
90	0.601	1.703	0.2	0.601	1.703	0.2	0	0
91	0.387	-1.56	0.2	0.387	-1.56	0.2	0	0
92	0.362	-0.885	0.2	0.362	-0.885	0.2	0	0
93	0.274	0.06	0.2	0.274	0.06	0.2	0	0
94	0.923	1.045	0.2	0.923	1.045	0.2	0	0
95	0.468	0.686	0.2	0.468	0.686	0.2	0	0
96	0.401	0.425	0.2	0.401	0.425	0.2	0	0
97	0.32	0.906	0.2	0.32	0.906	0.2	0	0
98	0.148	-0.69	0.2	0.148	-0.69	0.2	0	0
99	0.51	-0.376	0.2	0.51	-0.376	0.2	0	0
100	0.407	0.605	0.2	0.407	0.605	0.2	0	0

[1] 0: No DIF 1: Uniforme, 2: no uniforme; 3: Mixto

ANEXO 2:

Muestra de los archivos de comandos utilizados en el estudio sobre MH

2.1. Archivo de comandos del BILOG, para estimación de los parámetros

```
Diseño factorial completamente cruzado MH en función de tamaño muestral (2 niveles) r/f (5 niveles) con datos simulados
>COMMENTS  Este es el programa maestro para hacer las estimaciones de los parámetros en la primera condición experimental del estudio MH en DIF;

>GLOBAL   NPArm=3,  DFName='c:\bilog\datos\grF1si00.dat',  SAVe;
>SAVe   PARm='c:\bilog\datos\grF1si00.PAR', score='c:\bilog\datos\grF1si00.scr';
>LENGTH   NITems= 100;
>INPUT   NTOtal=100,  NALt=5, NIDch=4, CODe='10';
(104A1)
>TEST    TNAme=FOCAL;
>CALIB   float, Tprior, Case=2, cycles=50, Plot=0.1;
>SCORE   Method=1, RSCtype=3;
```

2.2. Archivo de comandos del EZDIF para identificación de ítems DIF

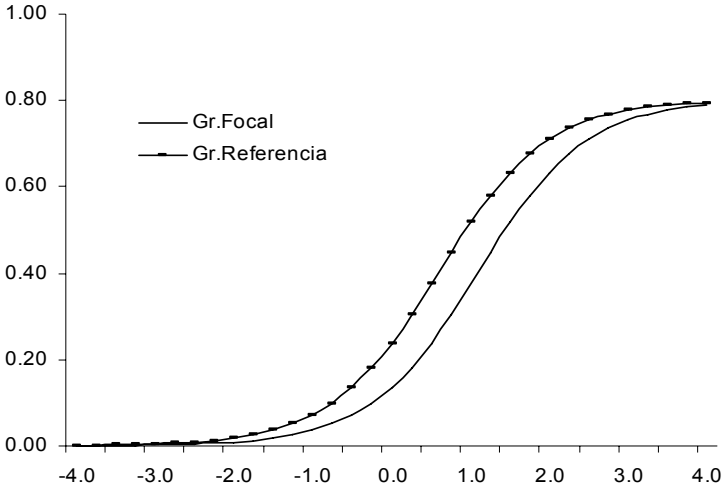
```
>TITLE:  DIF analisis 4unmsin 500 y 100 replica 100
>MODEL: 2
>REFERENCE:  c:\progdos\dif\datos\grr1si00.dat  500 100
(4X, 100F1.0)
>FOCAL:      c:\progdos\dif\datos\grf1si00.dat  100 100
(4X, 100F1.0)
>OUTPUT:     c:\progdos\dif\datos\resdif00.OUT
>LEVELS:     10
    0   28   35   42   46   51   56   61   67   74
    27  34   41   45   50   55   60   66   73  100
>LABELS:  N
```

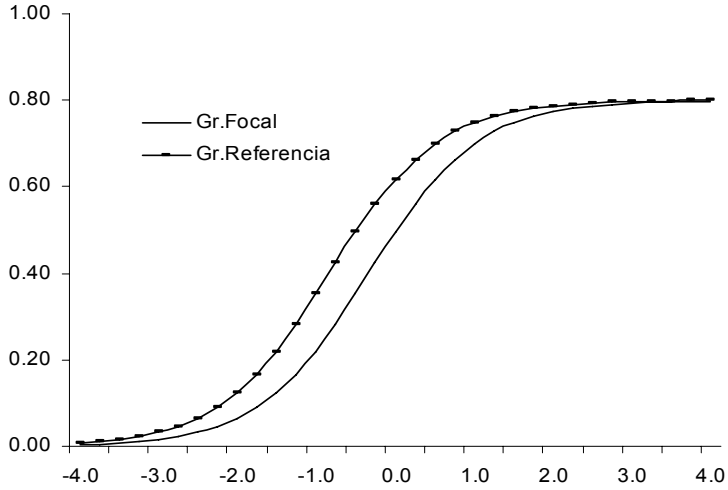
ANEXO 3:

Tasas de detección del Mantel-Haenszel

3.1. CCI, parámetros y tasas de detección de los cuatro ítems con DIF uniforme

Ítem 4	
a_r =	0.860
a_f =	0.860
b_r =	0.587
b_f =	1.087
$Max(T_{MH})$ =	1.000
$Min(T_{MH})$ =	0.400
$Media(T_{MH})$ =	0.880

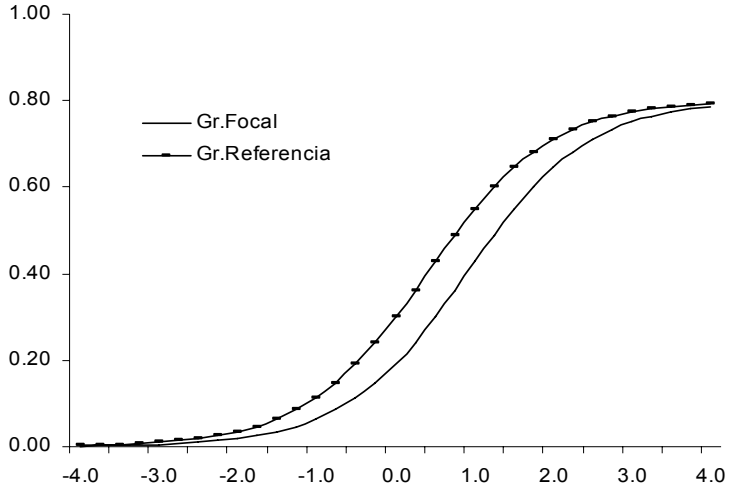


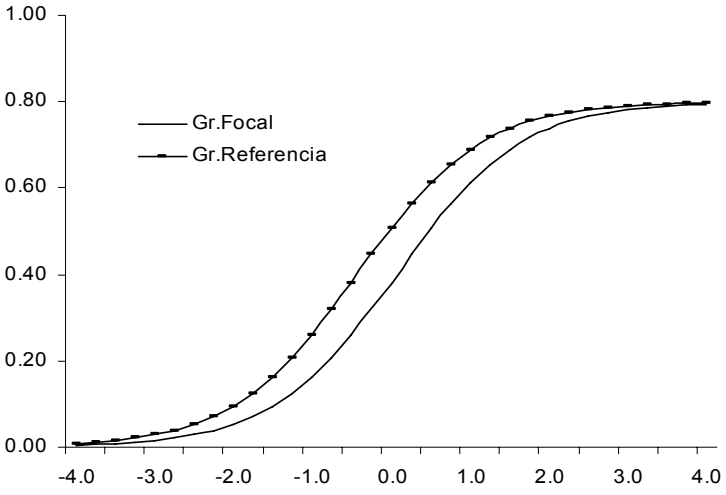


Ítem 24	
a_r =	0.858
a_f =	0.858
b_r =	-0.834
b_f =	-0.334
$Max(T_{MH})$ =	1.000
$Min(T_{MH})$ =	0.380
$Media(T_{MH})$ =	0.877

3.1 CCI, parámetros y tasas de detección de los cuatro ítems con DIF uniforme (*Continuación*)

Ítem 43	
a_r =	0.761
a_f =	0.761
b_r =	0.395
b_f =	0.895
$Max(T_{MH})$ =	1.000
$Min(T_{MH})$ =	0.320
$Media(T_{MH})$ =	0.818

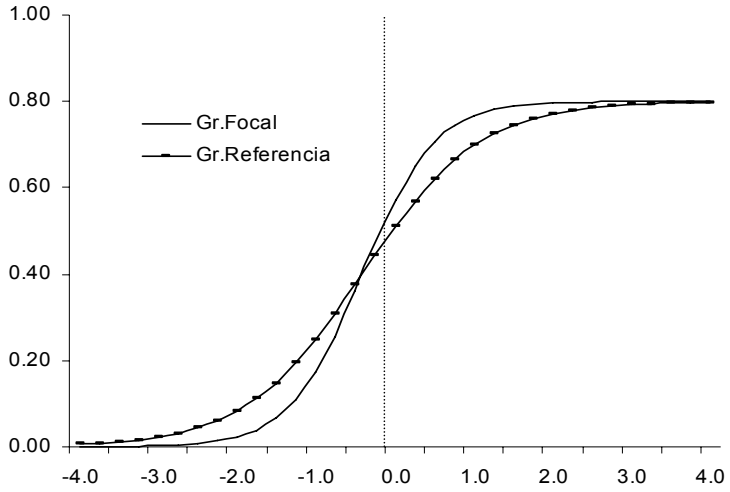


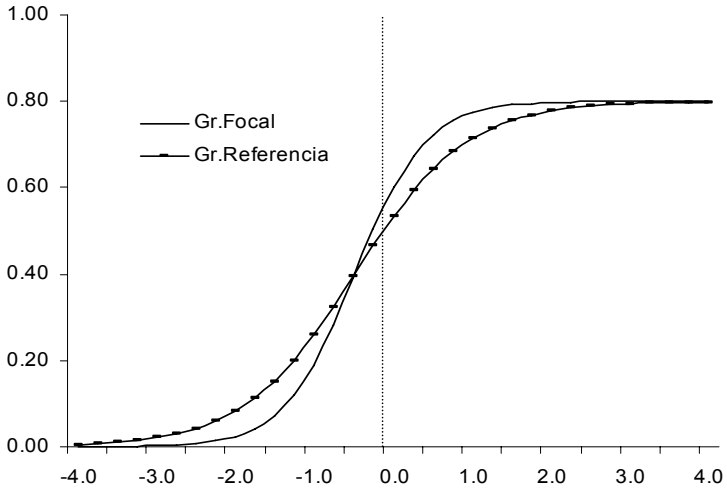


Ítem 73	
a_r =	0.755
a_f =	0.755
b_r =	-0.428
b_f =	0.072
$Max(T_{MH})$ =	1.000
$Min(T_{MH})$ =	0.280
$Media(T_{MH})$ =	0.843

3.2. CCI, parámetros y tasas de detección de los cuatro ítems con DIF no uniforme

Ítem 9	
a_r =	0.802
a_f =	1.302
b_r =	-0.413
b_f =	-0.413
$Max(T_{MH})$ =	0.870
$Min(T_{MH})$ =	0.050
$Media(T_{MH})$ =	0.371

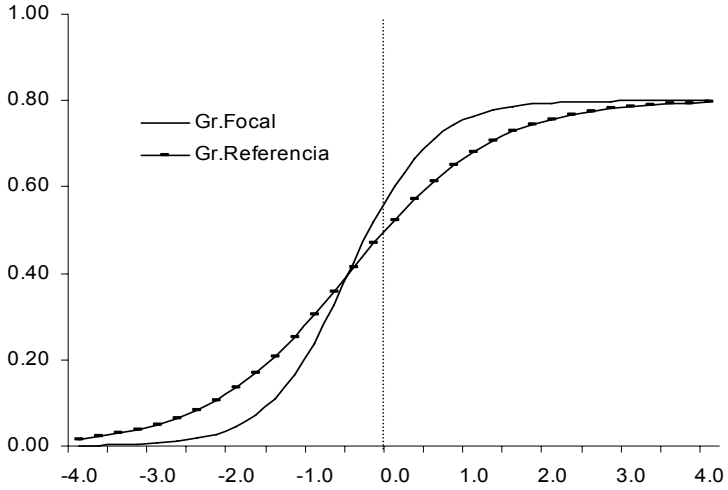


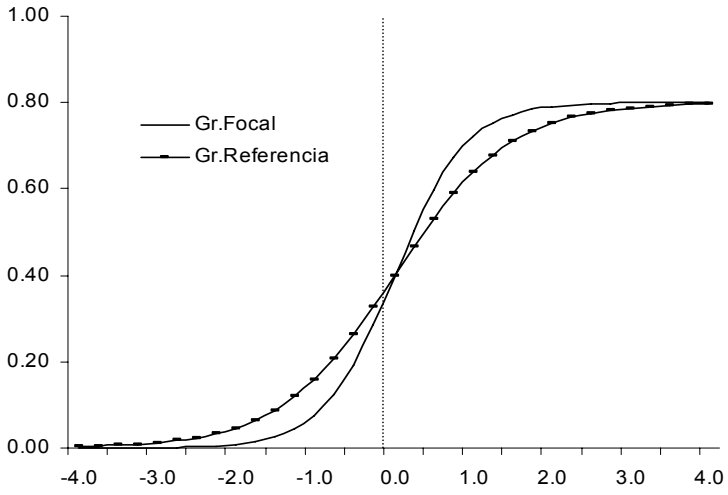


Ítem 49	
a_r =	0.841
a_f =	1.341
b_r =	-0.483
b_f =	-0.483
$Max(T_{MH})$ =	0.860
$Min(T_{MH})$ =	0.030
$Media(T_{MH})$ =	0.441

3.2. CCI, parámetros y tasas de detección de los cuatro ítems con DIF no uniforme (*Continuación*)

Ítem 70	
a_r =	0.658
a_f =	1.158
b_r =	-0.564
b_f =	-0.564
$Max(T_{MH})$ =	0.960
$Min(T_{MH})$ =	0.080
$Media(T_{MH})$ =	0.533

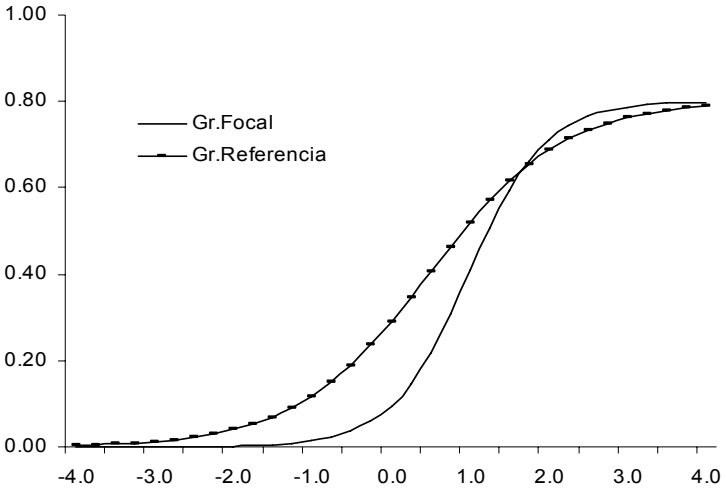


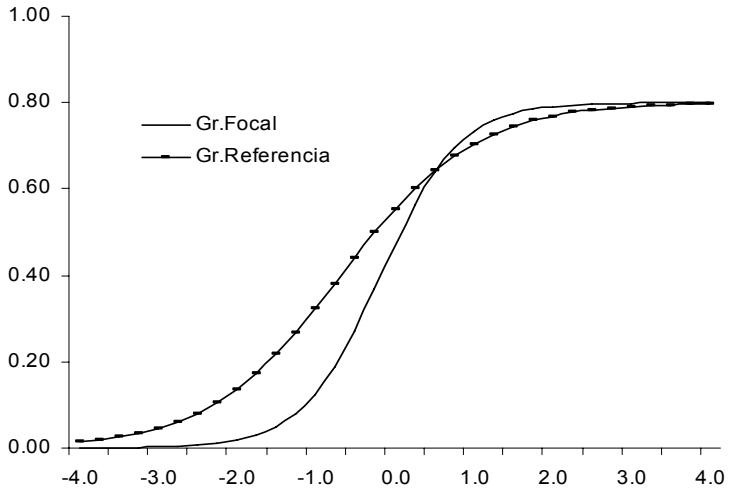


Ítem 82	
a_r =	0.815
a_f =	1.315
b_r =	0.012
b_f =	0.012
$Max(T_{MH})$ =	0.250
$Min(T_{MH})$ =	0.010
$Media(T_{MH})$ =	0.081

3.3. CCI, parámetros y tasas de detección de los cuatro ítems con DIF mixto

Ítem 18	
a_r =	0.701
a_f =	1.201
b_r =	0.480
b_f =	0.980
$Max(T_{MH})$ =	1.000
$Min(T_{MH})$ =	0.510
$Media(T_{MH})$ =	0.948

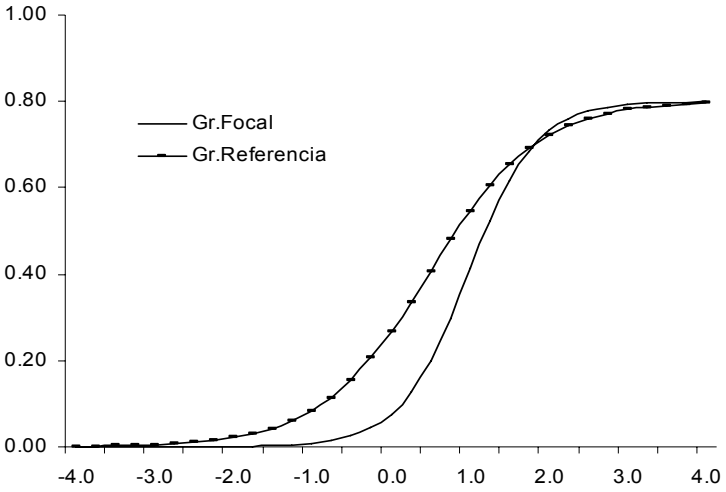


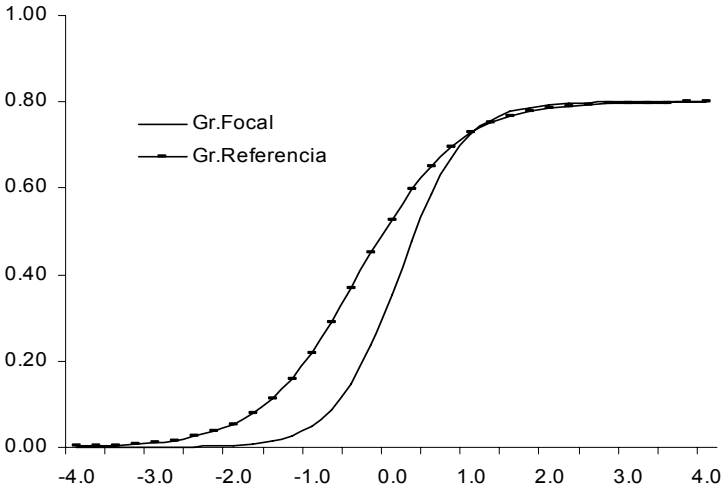


Ítem 42	
a_r =	0.704
a_f =	1.204
b_r =	-0.671
b_f =	-0.171
$Max(T_{MH})$ =	1.000
$Min(T_{MH})$ =	0.200
$Media(T_{MH})$ =	0.768

3.3. CCI, parámetros y tasas de detección de los cuatro ítems con DIF mixto (*Continuación*)

Ítem 53	
a_r =	0.861
a_f =	1.361
b_r =	0.474
b_f =	0.974
$Max(T_{MH})$ =	1.000
$Min(T_{MH})$ =	0.650
$Media(T_{MH})$ =	0.961





Ítem 83	
a_r =	0.966
a_f =	1.466
b_r =	-0.404
b_f =	0.096
$Max(T_{MH})$ =	1.000
$Min(T_{MH})$ =	0.390
$Media(T_{MH})$ =	0.923

3.4 Tasas de detección (%) del MH con $\alpha = .05$ para cada ítem DIF en cada condición experimental

		Nr=500							*Nr=1500*							
Tipo de DIF	Razón N° Item	1	2	2.5	3	4	5	Total	1	2	2.5	3	4	5	Total	Total general
Uniforme	4	96	91	91	79	66	40	77	99	99	99	99	100	97	99	88
	24	94	91	86	74	73	38	76	100	99	100	100	100	97	99	88
	43	97	82	73	63	46	32	66	100	99	99	100	98	93	98	82
	73	97	89	80	72	55	28	70	100	100	99	98	98	96	99	84
	Total	96	88	83	72	60	35	72	100	99	99	99	99	96	99	85
No Uni-forme	9	45	20	19	14	5	8	19	87	57	65	57	35	33	56	37
	49	48	32	26	18	11	3	23	86	71	68	63	53	50	65	44
	70	66	41	29	22	16	8	30	96	80	79	81	62	60	76	53
	82	12	5	7	1	3	2	5	25	6	5	12	10	9	11	8
	Total	43	25	20	14	9	5	19	74	54	54	53	40	38	52	36
Mixto	18	99	100	99	99	90	51	90	100	99	100	100	100	100	100	95
	42	83	67	59	64	58	20	59	100	100	98	95	92	85	95	77
	53	100	100	100	99	89	65	92	100	100	100	100	100	100	100	96
	83	99	98	97	95	83	39	85	99	99	100	100	100	99	100	92
	Total	95	91	89	89	80	44	81	100	100	100	99	98	96	99	90
Total general		78	68	64	58	50	28	58	91	84	84	84	79	77	83	70

ANEXO 4:

Tasas de FP (%) de la RL para cada condición experimental y cada categoría de ítems

4.1. Tasas de falsos positivos con $\alpha = .05$

Categoría de ítem	Tamaño *nr*	Razón de tamaños						Total general
		1	2	2.5	3	4	5	
a baja *b* baja (5 ítems)	*500*	**8.2**	7.2	7.0	6.0	4.6	6.8	6.6
	1500	**8.8**	4.4	5.6	5.4	6.2	**7.8**	6.4
	Total	**8.5**	5.8	6.3	5.7	5.4	7.3	6.5
a media *b* baja (4 ítems)	*500*	**8.0**	5.3	6.3	6.5	6.8	4.5	6.2
	1500	**8.0**	6.0	5.5	6.0	6.3	**8.0**	6.6
	Total	**8.0**	5.6	5.9	6.3	6.5	6.3	6.4
a baja *b* media (41 ítems)	*500*	6.3	6.9	6.7	6.5	5.7	5.9	6.3
	1500	6.6	6.4	6.0	6.5	5.5	6.1	6.2
	Total	6.5	6.7	6.3	6.5	5.6	6.0	6.3
a media *b* media (32 ítems)	*500*	6.5	6.0	5.8	5.8	5.5	5.8	5.9
	1500	6.3	5.6	6.3	6.3	6.2	6.7	6.2
	Total	6.4	5.8	6.1	6.0	5.9	6.2	6.1
a baja *b* alta (3 ítems)	*500*	3.7	4.7	2.0	7.3	7.0	5.0	4.9
	1500	7.3	6.7	7.3	4.7	5.7	3.7	5.9
	Total	5.5	5.7	4.7	6.0	6.3	4.3	5.4
a media *b* alta (3 ítems)	*500*	5.0	**9.0**	7.3	5.0	6.7	5.0	6.3
	1500	**8.3**	6.3	6.3	5.7	6.0	**8.7**	6.9
	Total	6.7	7.7	6.8	5.3	6.3	6.8	6.6
Total general		6.6	6.2	6.2	6.2	5.8	6.1	6.2

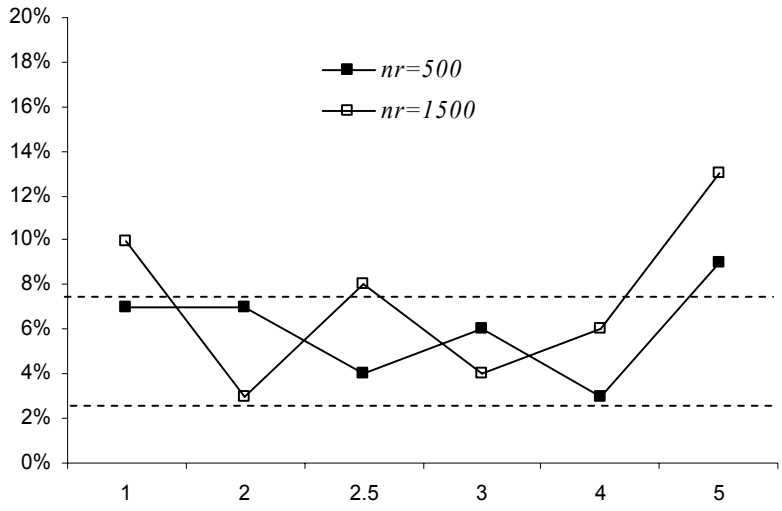
4.2. Tasas de falsos positivos con $\alpha = .01$

Categoría de ítem	Tamaño *nr*	Razón de tamaños						Total general
		1	2	2.5	3	4	5	
a baja *b* baja (5 ítems)	*500*	1.4	2.0	0.2	1.0	0.6	0.6	1.0
	1500	2.8	1.4	1.2	0.6	1.8	1.4	1.5
	Total	2.1	1.7	0.7	0.8	1.2	1.0	1.3
a media *b* baja (4 ítems)	*500*	2.3	1.5	1.3	0.8	1.0	0.8	1.3
	1500	2.8	2.5	1.3	0.3	1.3	1.8	1.6
	Total	2.5	2.0	1.3	0.5	1.1	1.3	1.4
a baja *b* media (41 ítems)	*500*	1.5	1.4	1.1	1.4	1.2	1.2	1.3
	1500	2.2	1.4	1.1	1.4	1.3	1.4	1.5
	Total	1.9	1.4	1.1	1.4	1.3	1.3	1.4
a media *b* media (32 ítems)	*500*	1.2	1.4	1.6	1.3	1.3	1.2	1.3
	1500	2.3	1.4	1.1	1.2	1.1	1.2	1.4
	Total	1.8	1.4	1.3	1.3	1.2	1.2	1.3
a baja *b* alta (3 ítems)	*500*	0.3	1.7	0.0	0.3	1.3	0.7	0.7
	1500	2.0	1.7	1.3	0.7	1.0	1.3	1.3
	Total	1.2	1.7	0.7	0.5	1.2	1.0	1.0
a media *b* alta (3 ítems)	*500*	1.0	0.7	1.7	0.7	1.3	0.7	1.0
	1500	2.3	1.3	2.3	1.3	1.0	3.0	1.9
	Total	1.7	1.0	2.0	1.0	1.2	1.8	1.4
Total general		1.8%	1.4	1.2	1.2	1.2	1.3	1.4

ANEXO 5:

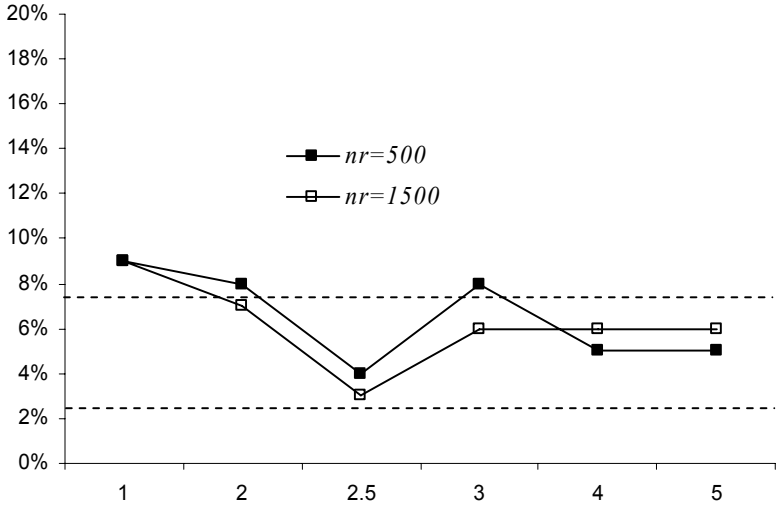
Tasas de FP de la RL con $\alpha = .05$ para las categorías de ítems que presentaron alguna elevación del error tipo I por encima del intervalo de Bradley (1978)

5.1. Ítems de baja dificultad y baja discriminación

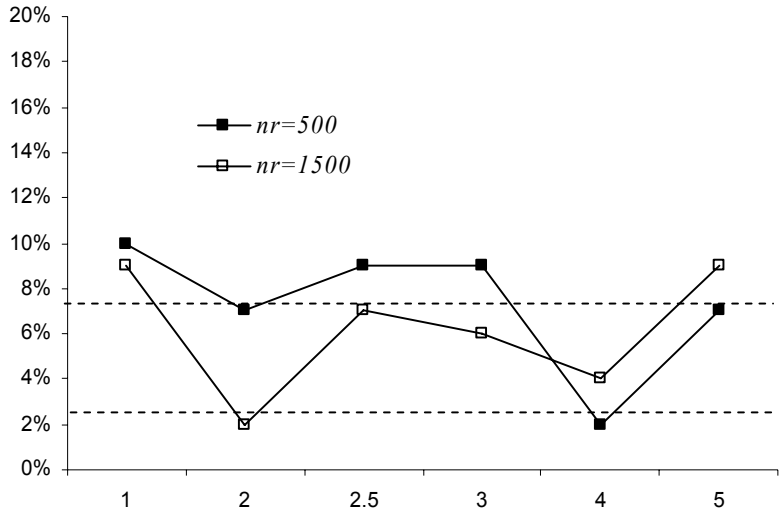


ítem 11	
a=	*0.499*
b=	*-1.529*
c=	*0.200*
T_{RL}=	*6.67*

ítem 54	
a=	*0.402*
b=	*-1.771*
c=	*0.2*
T_{RL}=	*6.33*

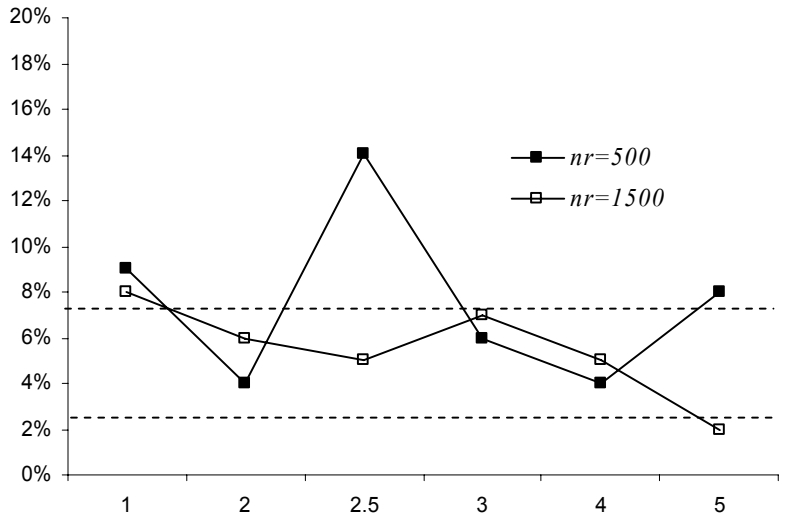


5.1. Ítems de baja dificultad y baja discriminación (*Continuación*)

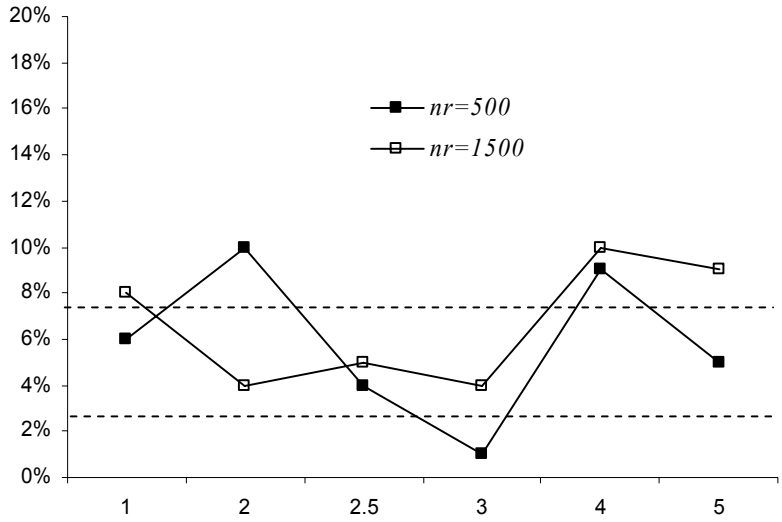


ítem 69	
a=	*0.466*
b=	*-1.517*
c=	*0.2*
T_{RL}=	*6.75*

ítem 76	
a=	*0.096*
b=	*-1.751*
c=	*0.2*
T_{RL}=	*6.5*



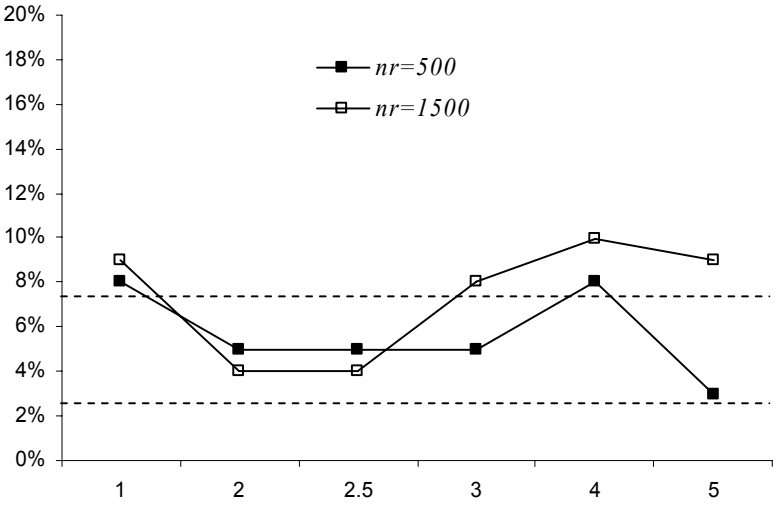
5.1. Ítems de baja dificultad y baja discriminación (*Continuación*)



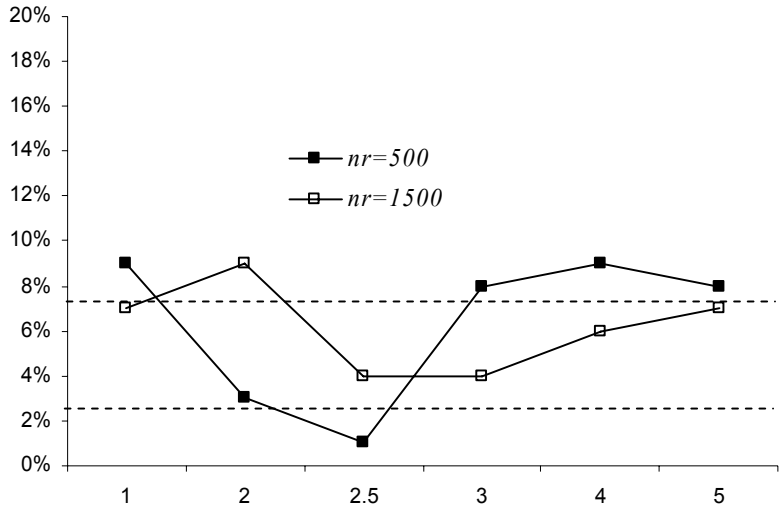
ítem 91	
a=	*0.387*
b=	*-1.560*
c=	*0.2*
T_{RL}=	*6.25*

5.2. Ítems de baja dificultad y discriminación media

Ítem 38	
a=	*0.726*
b=	*-1.621*
c=	*0.2*
T_{RL}=	*6.5*

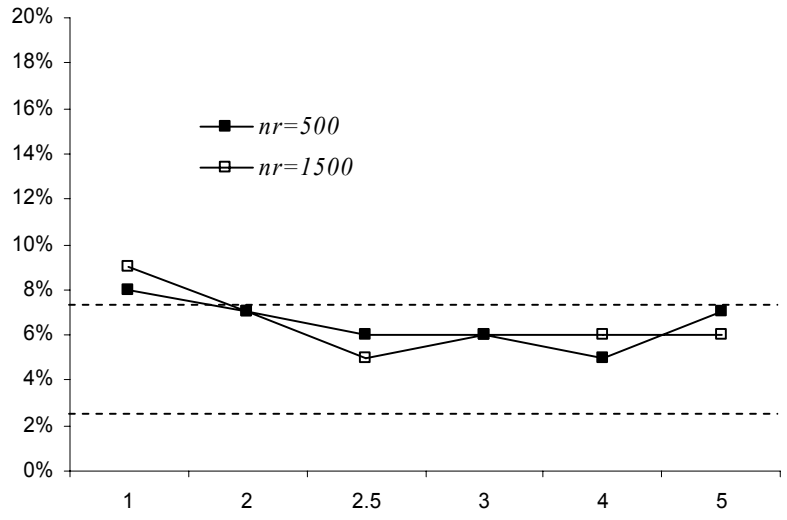


5.2. Ítems de baja dificultad y discriminación media (*Continuación*)

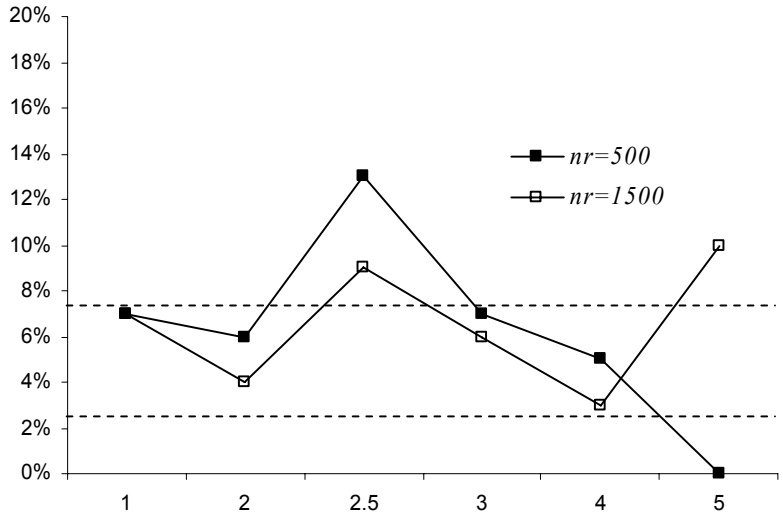


Ítem 40	
a=	0.597
b=	-1.529
c=	0.2
T_{RL}=	6.25

Ítem 74	
a=	0.512
b=	-2.078
c=	0.2
T_{RL}=	6.5



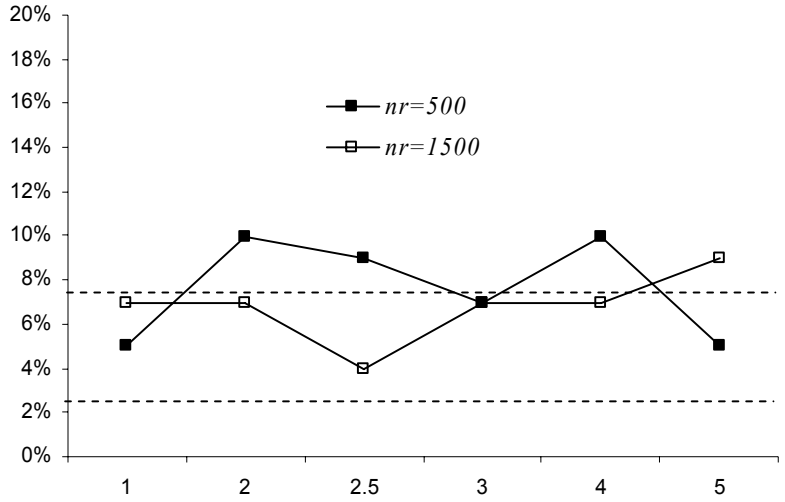
5.2. Ítems de baja dificultad y discriminación media (*Continuación*)



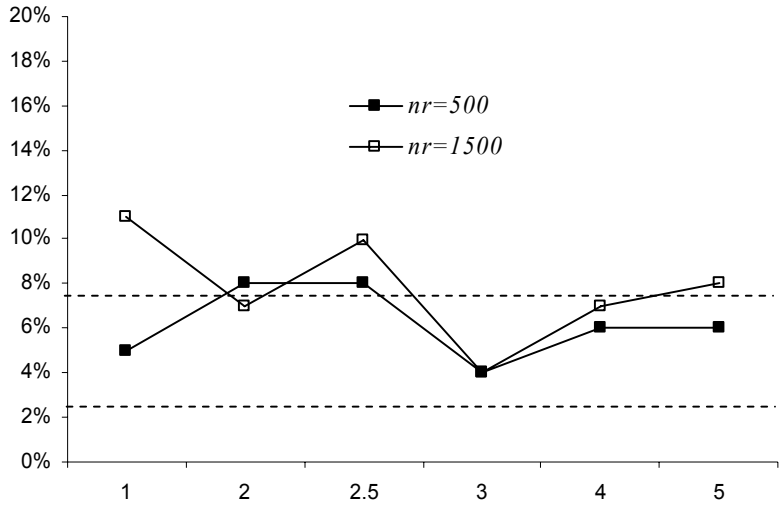
Ítem 77
a= 0.736
b= -1.950
c= 0.2
T_{RL}= 6.417

5.3. Ítems de alta dificultad y discriminación media

Ítem 19
a= 0.798
b= 1.885
c= 0.2
T_{RL}= 7.25

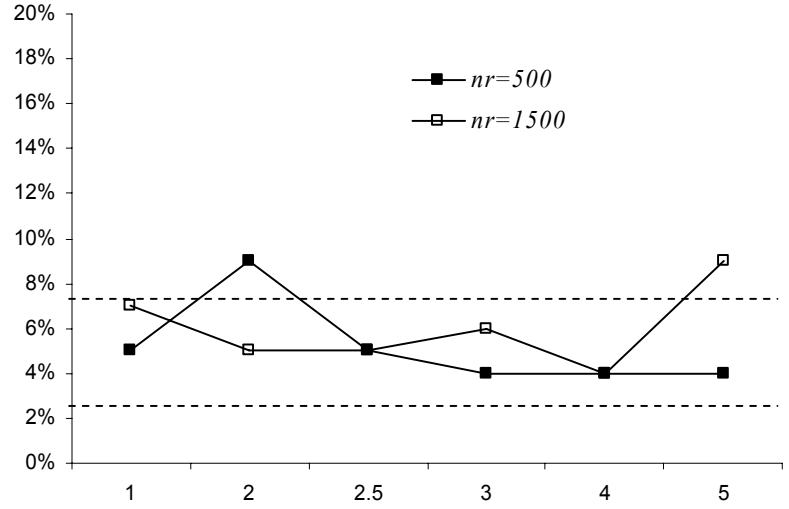


5.3. Ítems de alta dificultad y discriminación media (*Continuación*)



Ítem 62	
a=	*0.806*
b=	*1.667*
c=	*0.2*
T_{RL}=	*7.0*

Ítem 90	
a=	*0.601*
b=	*1.703*
c=	*0.2*
T_{RL}=	*5.58*



ANEXO 6:

Correlaciones bivariadas y parciales entre los parámetros de los ítems y las tasas de FP de la regresión logística

6.1. Correlación de Pearson entre los parámetros de dificultad y discriminación con la tasa de FP de la RL dentro de cada condición experimental

	Discriminación		Dificultad	
Tamaños de grupos (razón)	**r**	**Significación**	**r**	**Significación**
500 y 500 (1)	0.06	0.56	**-0.22**	**0.04**
500 y 250 (2)	-0.03	0.78	0.09	0.43
500 y 200 (2.5)	-0.05	0.67	-0.19	0.07
500 y 167 (3)	-0.12	0.25	0.12	0.25
500 y 125 (49	-0.11	0.33	0.11	0.32
500 y 100 (5)	-0.10	0.33	-0.06	0.56
1500 y 1500 (1)	-0.16	0.13	**-0.25**	**0.02**
1500 y 750 (2)	-0.11	0.29	0.20	0.06
1500 y 600 (2,5)	0.08	0.47	0.03	0.75
1500 y 500 (3)	0.03	0.79	0.03	0.76
1500 y 375 (4)	0.08	0.48	-0.09	0.39
1500 y 300 (5)	0.21	0.05	0.04	0.74

6.2. Correlación parcial entre los parámetros de dificultad y discriminación con la tasa de FP de la RL

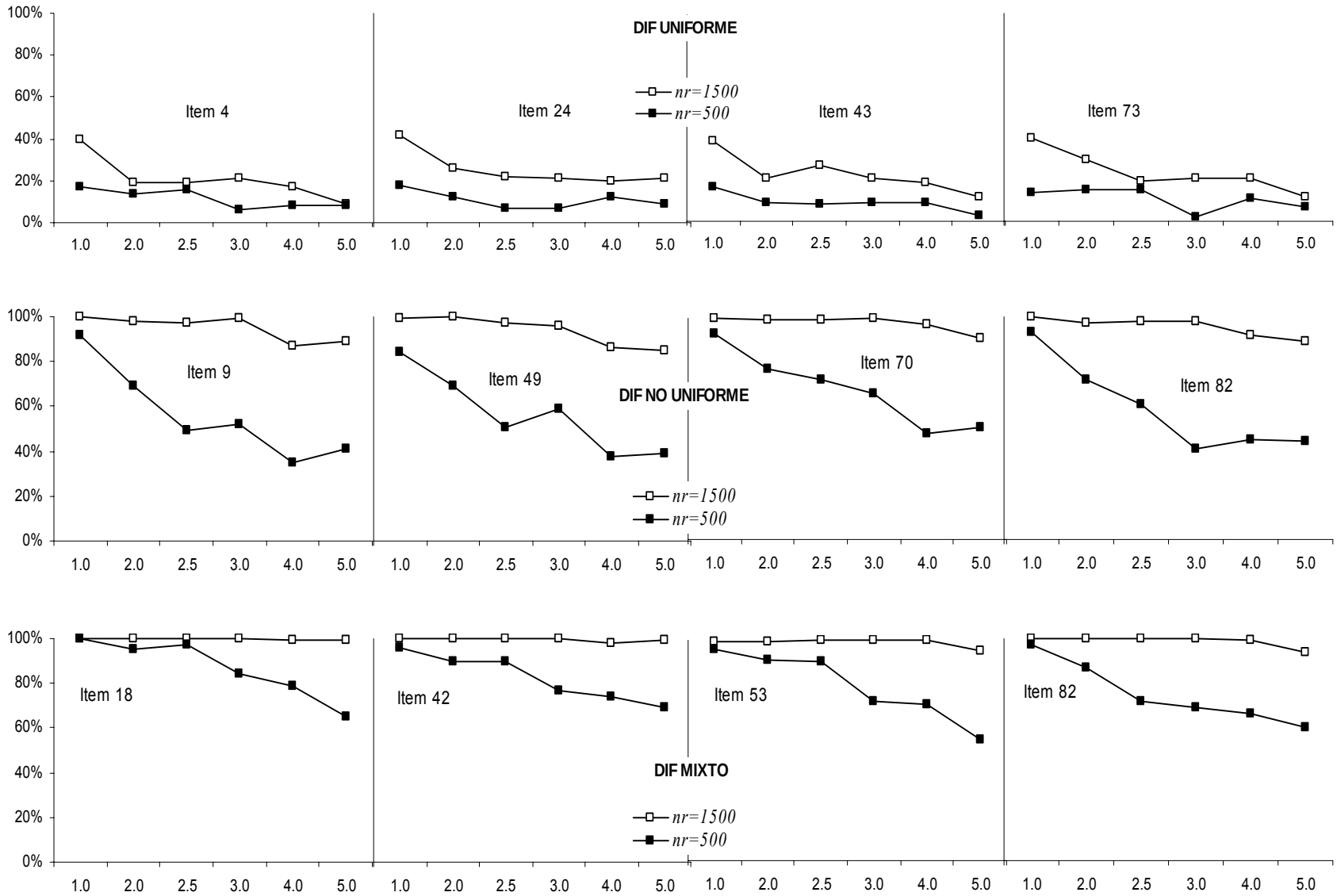
	Discriminación		Dificultad	
	r	**Significación**	***r***	**Significación**
Para todas las condiciones experimentales				
Controlando *nr*	-0.02	0.57	-0.02	0.49
Controlando *r*	-0.02	0.57	-0.02	0.49
Controlando *nr* y *r*	-0.02	0.57	-0.02	0.49
Para cada *r* controlando *nr*				
r=1	-0.04	0.61	**-0.24**	**0.00**
r=2	-0.07	0.35	0.14	0.06
r=2.5	0.01	0.92	-0.09	0.23
r=3	-0.05	0.50	0.08	0.29
r=4	-0.01	0.87	0.00	0.96
r=5	0.05	0.51	-0.01	0.86

ANEXO 7:

Tasas de detecciones correctas de la Regresión Logística

7.1. Tasas de detección (%) con $\alpha = .05$ para cada ítem DIF en cada condición experimental

		Nr=500							*Nr=1500*							
Tipo de DIF	Razón N° Item	1	2	2.5	3	4	5	Total	1	2	2.5	3	4	5	Total	Total general
Uniforme	4	17	14	16	6	8	8	12	40	19	19	21	17	9	21	16
	24	18	12	7	7	12	9	11	42	26	22	21	20	21	25	18
	43	17	9	8	9	9	3	9	39	21	27	21	19	12	23	16
	73	14	15	15	2	11	7	11	40	30	19	21	21	12	24	17
	Total	17	13	12	6	10	7	11	40	24	22	21	19	14	23	17
No Uni-forme	9	92	69	49	52	35	41	56	100	98	97	99	87	89	95	76
	49	84	69	51	59	38	39	57	99	100	97	96	86	85	94	75
	70	93	77	72	66	48	51	68	100	99	99	100	97	91	98	83
	82	93	72	61	41	45	44	59	100	97	98	98	92	89	96	78
	Total	91	72	58	55	42	44	60	100	99	98	98	91	89	96	78
Mixto	18	100	95	97	84	79	65	87	100	100	100	100	99	99	100	93
	42	96	90	90	77	74	69	83	100	100	100	100	98	99	100	91
	53	96	91	90	72	71	55	79	99	99	100	100	100	95	99	89
	83	97	87	72	69	66	60	75	100	100	100	100	99	94	99	87
	Total	97	91	87	76	73	62	81	100	100	100	100	99	97	99	90
Total general		68	58	52	45	41	38	51	80	74	73	73	70	66	73	62


DIF UNIFORME
Item 4
Item 24
Item 43
Item 73
nr=1500
nr=500
DIF NO UNIFORME
Item 9
Item 49
Item 70
Item 82
nr=1500
nr=500
DIF MIXTO
Item 18
Item 42
Item 53
Item 82
nr=1500
nr=500
100%
80%
60%
40%
20%
0%
1.0
2.0
2.5
3.0
4.0
5.0

7.2. Tasas de detección (%) con $\alpha = .01$ para cada ítem DIF en cada condición experimental

		Nr=500							*Nr=1500*							
Tipo de DIF	Razón N° Item	1	2	2.5	3	4	5	Total	1	2	2.5	3	4	5	Total	Total general
Uniforme	4	4	4	0	2	1	0	2	15	6	8	4	6	2	7	4
	24	5	6	4	4	2	0	4	15	7	10	8	3	5	8	6
	43	4	2	1	2	2	2	2	15	8	8	6	2	4	7	5
	73	2	7	4	1	1	1	3	19	11	9	6	4	4	9	6
	Total	4	5	2	2	2	1	3	16	8	9	6	4	4	8	5
No Uniforme	9	72	39	21	17	17	11	30	97	95	91	88	65	73	85	57
	49	67	31	24	27	8	12	28	99	92	87	82	67	65	82	55
	70	83	51	24	28	19	13	36	99	99	96	93	84	77	91	64
	82	74	44	27	13	17	14	32	99	92	88	87	78	62	84	58
	Total	74	41	24	21	15	13	31	99	95	91	88	74	69	86	59
Mixto	18	94	78	74	58	41	20	61	99	100	100	100	98	95	99	80
	42	93	78	55	44	42	35	58	100	99	99	100	94	90	97	77
	53	90	67	68	46	38	9	53	99	99	99	98	98	88	97	75
	83	80	66	40	28	26	28	45	100	98	98	98	95	85	96	70
	Total	89	72	59	44	37	23	54	100	99	99	99	96	90	97	76
Total general		56	39	29	23	18	12	29	71	67	66	64	58	54	63	46

7.3. Parámetros de los ítems y estadísticas descriptivas de las tasas de detección por tipo de DIF

Tipo de DIF	N° del Ítem	Parámetros de los ítems*				Área entre CCI	Tasa de detección de la RL					
		Grupo Referencia		Grupo focal			$\alpha = .05$			$\alpha = .01$		
		a_r	b_r	a_f	b_f		Máximo	Mínimo	Media	Máximo	Mínimo	Media
Uniforme	4	0.86	0.587	0.86	1.087	0.4	40	6	16	15	0	4
	24	0.858	-0.834	0.858	-0.334	0.4	42	7	18	15	0	6
	43	0.761	0.395	0.761	0.895	0.4	39	3	16	15	1	5
	73	0.755	-0.428	0.755	0.072	0.4	40	2	17	19	1	6
No uni-forme	9	0.802	-0.413	1.302	-0.413	0.31	100	35	76	97	11	57
	49	0.841	-0.483	1.341	-0.483	0.30	100	38	75	99	08	55
	70	0.658	-0.564	1.158	-0.564	0.43	100	48	83	99	13	64
	82	0.815	0.012	1.315	0.012	0.30	100	41	78	99	13	58
Mixto	18	0.701	0.48	1.201	0.98	0.79	100	65	93	100	20	80
	42	0.704	-0.671	1.204	-0.171	0.78	100	69	91	100	35	77
	53	0.861	0.474	1.361	0.974	0.68	100	55	89	99	9	75
	83	0.966	-0.404	1.466	0.096	0.63	100	60	87	100	26	70

* Para todos los ítems $c=0.2$

ANEXO 8:

Muestra de los archivos de comando de BILOG creados para el estudio de Ji cuadrado de Lord.

```
Factorial completamente cruzado JI en función de Tamano muestral (2 niveles) r/f (5 niveles)
y longitud de prueba (2 niveles) con datos simulados

>COMMENTS
Este es el programa maestro para hacer las estimacines de los parametros en la primera condición
experimental del estudio JI en DIF con float, Case=2 y Plot=0.1. Replica 01

>GLOBAL   NPArm=2,  DFName='c:\bilog\datos\1F5001.dat',
  SAVe;
>SAVe   PARm='c:\bilog\datos\1F5001.PAR', score='c:\bilog\datos\1F5001.scr' ,
          COvariance='c:\bilog\datos\1F5001.cov';
>LENGTH   NITems= 50;
>INPUT   NTOtal=50,  NALt=5, NIDch=4, CODe='10';
(54A1)
>TEST    TNAme=FOCAL;
>CALIB   float, Tprior, Case=2, cycles=50, Plot=0.1;
>SCORE   Method=1, RSCtype=3;
```

ANEXO 9:

Tasas (en %) de FP y de detecciones correctas del Ji cuadrado en cada condición experimental por tipo de DIF

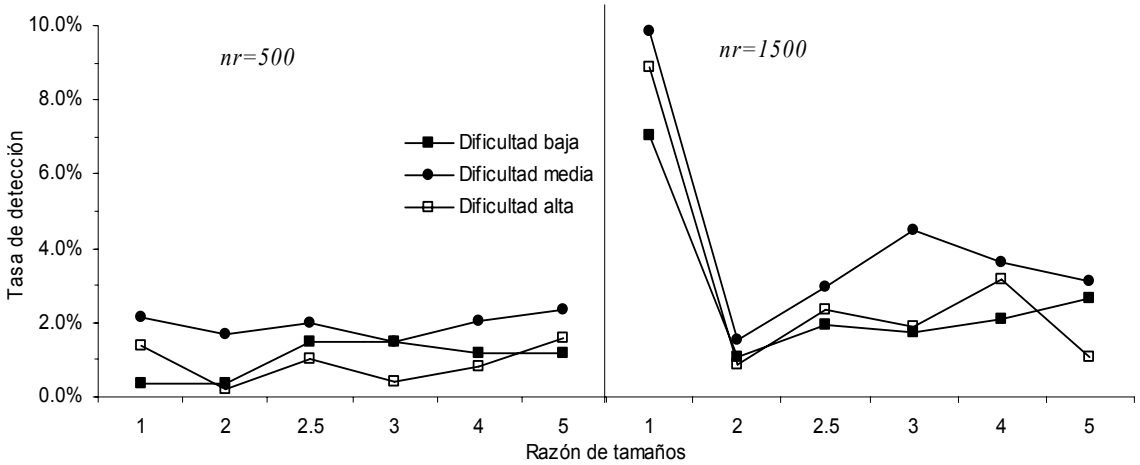
Condición	Tamaño grupo referencia	Razón de tamaños	N de ítems	Tasa de falsos positivos			DIF uniforme			DIF no uniforme			DIF mixto		
				$\alpha = .01$	$\alpha = .05$	$\alpha = .1$	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
1	500	5	50	0.3	2.1	5.2	17	31	40	1	4	11	39	61	76
2	500	5	100	0.3	2.2	4.8	13	32	43	0	2	10	44	69	78
3	500	4	50	0.2	1.8	3.7	43	69	79	0	5	9	74	90	96
4	500	4	100	0.3	1.9	4.0	36	66	77	2	6	12	85	96	99
5	500	3	50	0.2	1.3	3.3	39	63	74	0	1	7	78	91	93
6	500	3	100	0.2	1.5	3.5	40	69	78	3	9	15	92	96	98
7	500	2.5	50	0.3	2.0	4.0	22	49	59	8	25	35	58	77	87
8	500	2.5	100	0.2	1.8	3.9	24	48	64	14	35	47	73	91	94
9	500	2	50	0.4	1.3	3.3	26	57	71	13	39	59	76	91	93
10	500	2	100	0.3	1.5	3.3	32	59	73	20	52	67	86	95	99
11	500	1	50	0.4	2.0	4.0	65	79	83	52	71	87	87	95	98
12	500	1	100	0.4	1.9	4.3	62	78	85	60	83	90	94	98	99
13	1500	5	50	0.6	2.2	5.6	44	65	74	38	73	89	89	97	97
14	1500	5	100	0.6	3.0	6.1	44	68	75	60	85	90	96	99	99
15	1500	4	50	0.5	3.3	7.3	51	71	83	49	76	83	89	97	99
16	1500	4	100	0.7	3.2	7.0	53	74	86	70	88	92	96	100	100
17	1500	3	50	0.5	3.7	7.5	57	77	85	90	96	100	98	100	100
18	1500	3	100	0.8	3.9	7.8	54	74	86	92	98	99	99	100	100
19	1500	2.5	50	0.4	2.3	5.4	88	95	98	80	93	95	100	100	100
20	1500	2.5	100	0.4	2.9	5.7	86	95	97	89	96	97	100	100	100
21	1500	2	50	0.2	1.5	3.3	99	100	100	71	87	94	100	100	100
22	1500	2	100	0.1	1.3	3.3	99	100	100	78	92	96	100	100	100
23	1500	1	50	**2.0**	**8.3**	**14.3**	82	93	95	98	99	99	100	100	100
24	1500	1	100	**2.3**	**9.3**	**15.2**	82	91	95	99	100	100	100	100	100

ANEXO 10:

Detecciones incorrectas del Ji cuadrado de Lord según los parámetros de los ítems

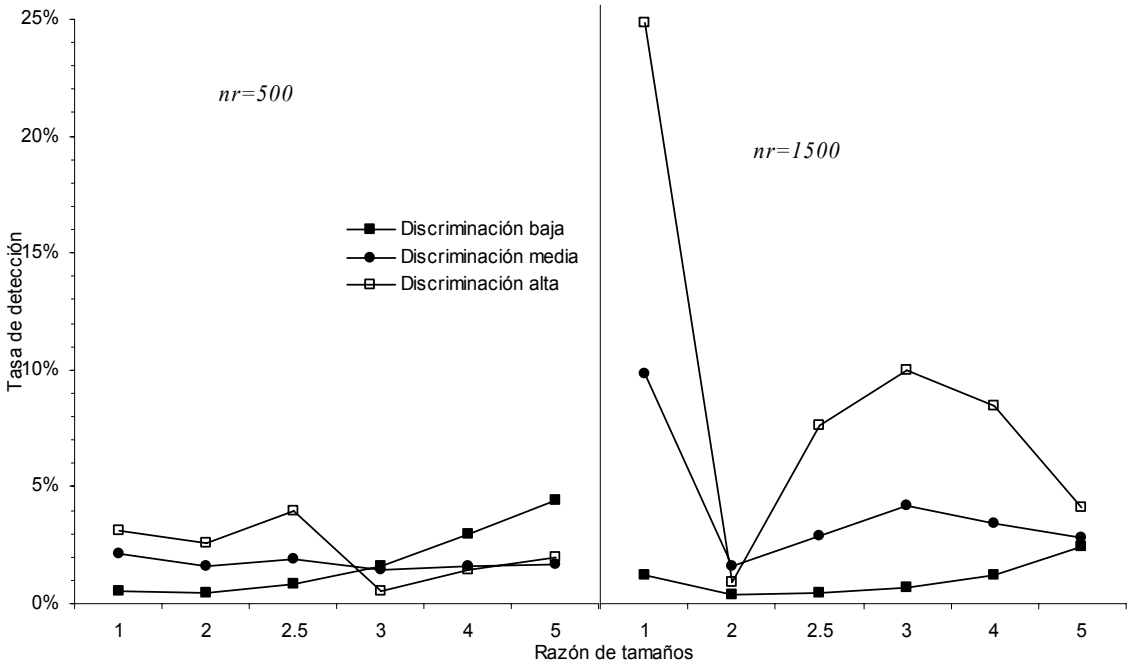
10.1. Promedio de detecciones incorrectas (%)con α =0.05 según la dificultad del ítem

Dificultad del ítem	Tamaño del grupo de referencia	Razón de tamaños						
		1	2	2.5	3	4	5	Total
Baja (b_i<-1.5)	500	0.3	0.3	1.5	1.5	1.2	1.2	1.0
	1500	6.7	1.0	1.8	1.7	2.0	2.5	2.6
	Total	3.5	0.7	1.7	1.6	1.6	1.8	1.8
Media (-1.5≤ b_i ≤1.5)	500	2.2	1.7	2.0	1.5	2.0	2.3	2.0
	1500	**9.3**	1.4	2.8	4.2	3.4	2.9	4.0
	Total	5.7	1.6	2.4	2.9	2.7	2.6	3.0
Alta (b_i>1.5)	500	1.4	0.2	1.0	0.4	0.8	1.6	0.9
	1500	**8.4**	0.8	2.2	1.8	3.0	1.0	2.9
	Total	4.9	0.5	1.6	1.1	1.9	1.3	1.9
Total		5.5	1.4	2.3	2.6	2.5	2.5	2.8



10.2. Promedio de detecciones incorrectas (%) con α=0.05 según la discriminación del ítem

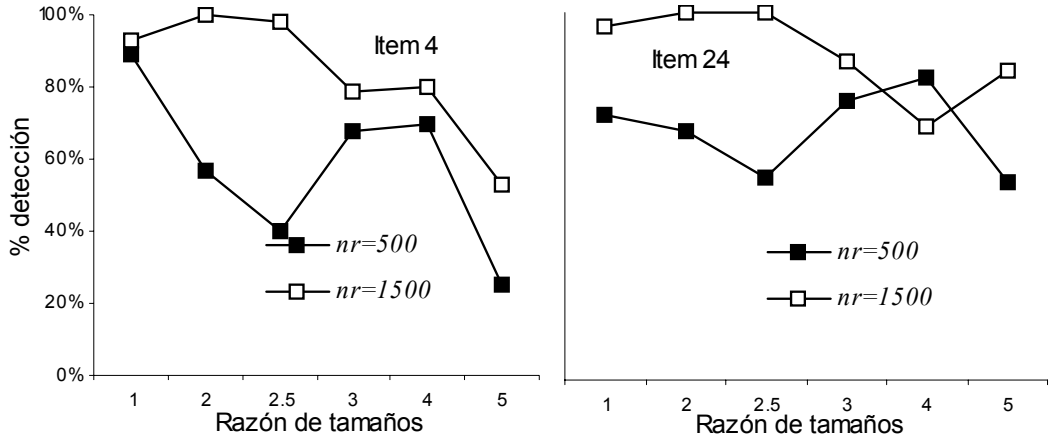
Discriminación del ítem	Tamaño del grupo de referencia	Razón de tamaños 1	2	2.5	3	4	5	Total
Baja (a_i<.2)	500	0.5	0.5	0.8	1.6	3.0	4.5	1.8
	1500	1.2	0.4	0.5	0.6	1.2	2.4	1.0
	Total	0.9	0.4	0.6	1.1	2.1	3.4	1.4
Media (.2≤ a_i ≤.8)	500	2.2	1.6	1.9	1.4	1.6	1.7	1.7
	1500	**9.6**	1.6	2.8	4.1	3.4	2.7	4.0
	Total	5.9	1.6	2.4	2.8	2.5	2.2	2.9
Alta (a_i>.8)	500	3.1	2.6	4.0	0.6	1.4	2.0	2.3
	1500	**24.3**	0.9	7.4	**9.7**	**8.3**	4.0	9.1
	Total	13.7	1.7	5.7	5.1	4.9	3.0	5.7
Total		5.5	1.4	2.3	2.6	2.5	2.5	2.8

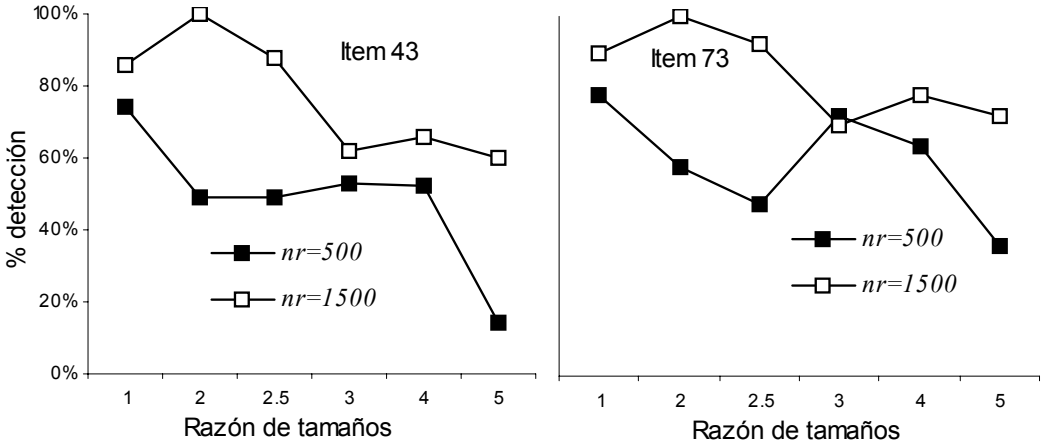


ANEXO 11:

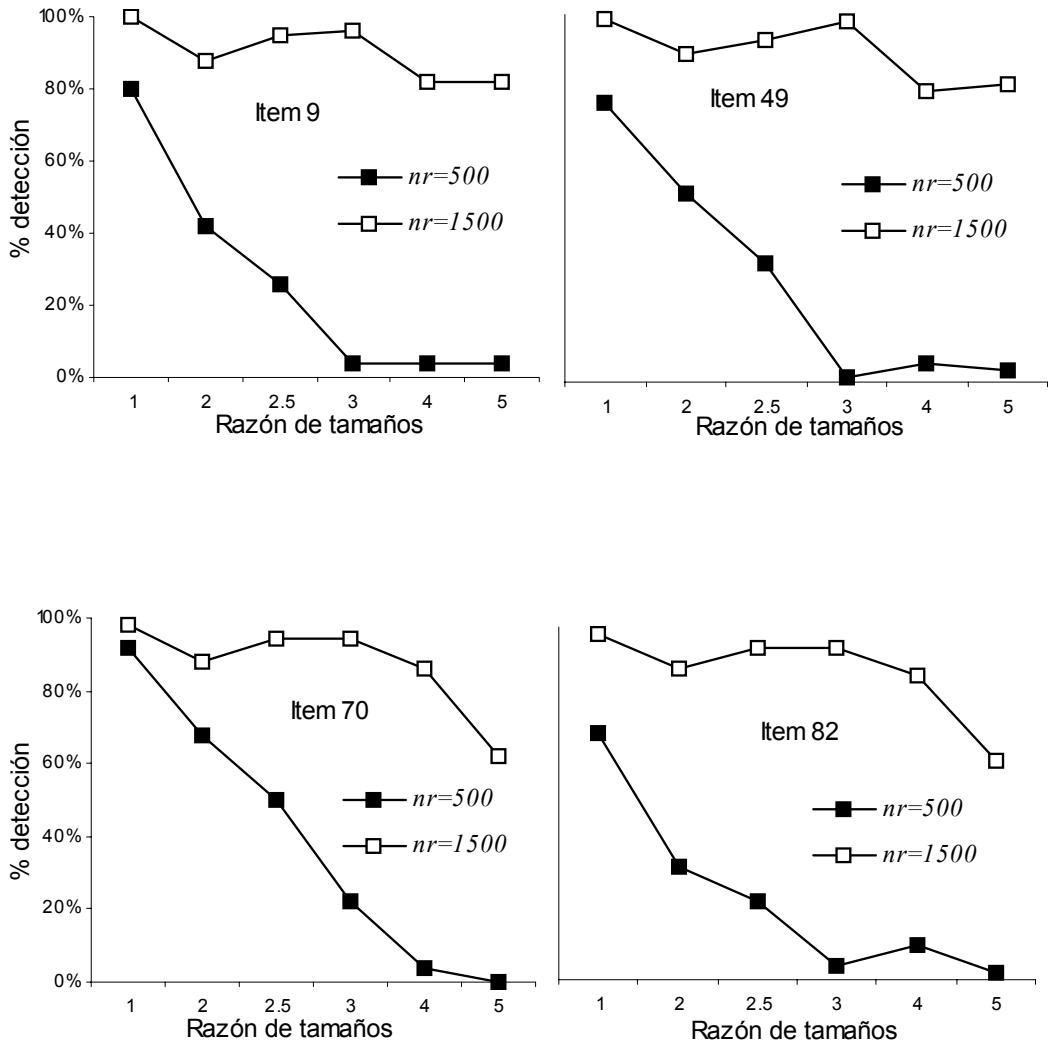
Tasas de detecciones correctas del Ji cuadrado de Lord

11.1. Tasas de detección de DIF uniforme con α =.05 en función de los tamaños de los grupos

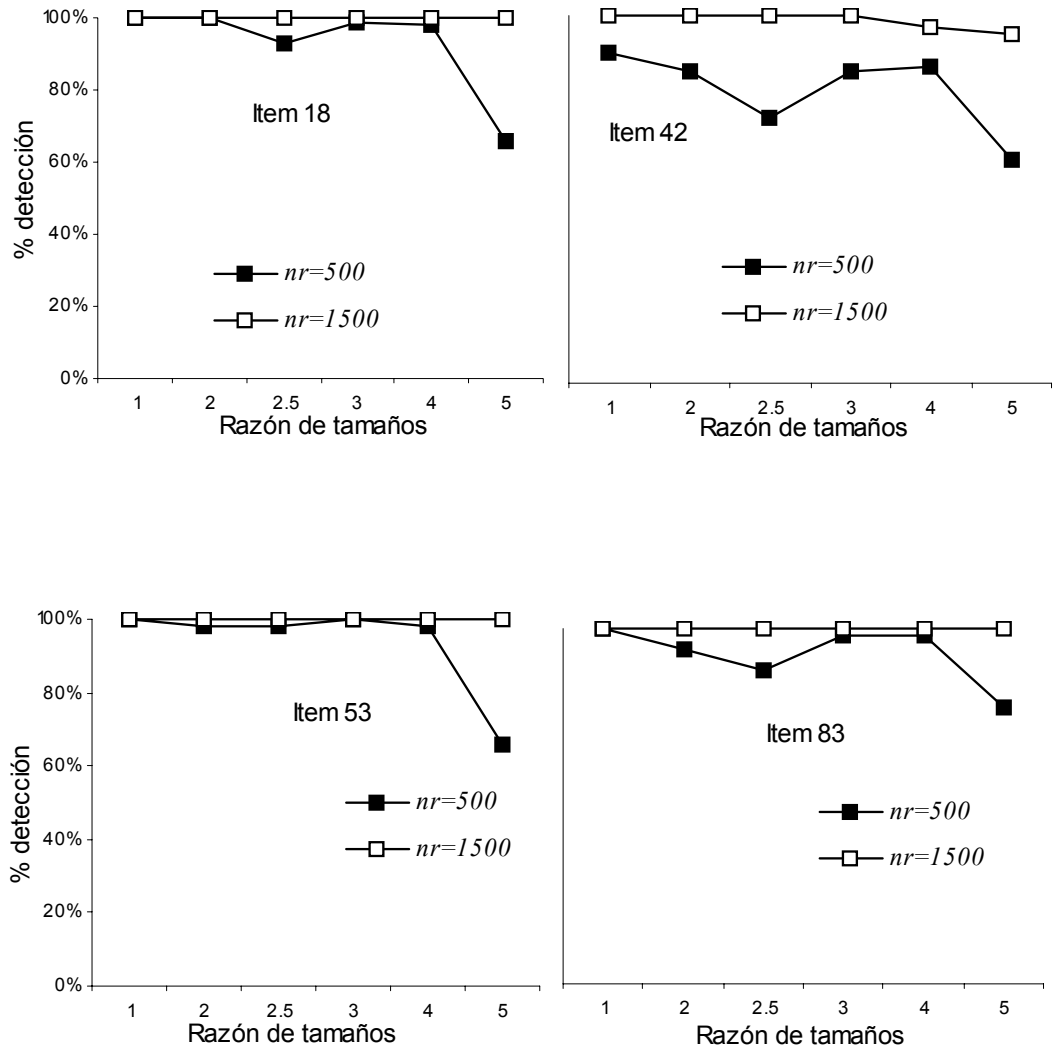




11.2. Tasas de detección de DIF no uniforme con α =.05 en función de los tamaños de los grupos



11.3. Tasas de detección de DIF mixto con $\alpha = .05$ en función de los tamaños de los grupos

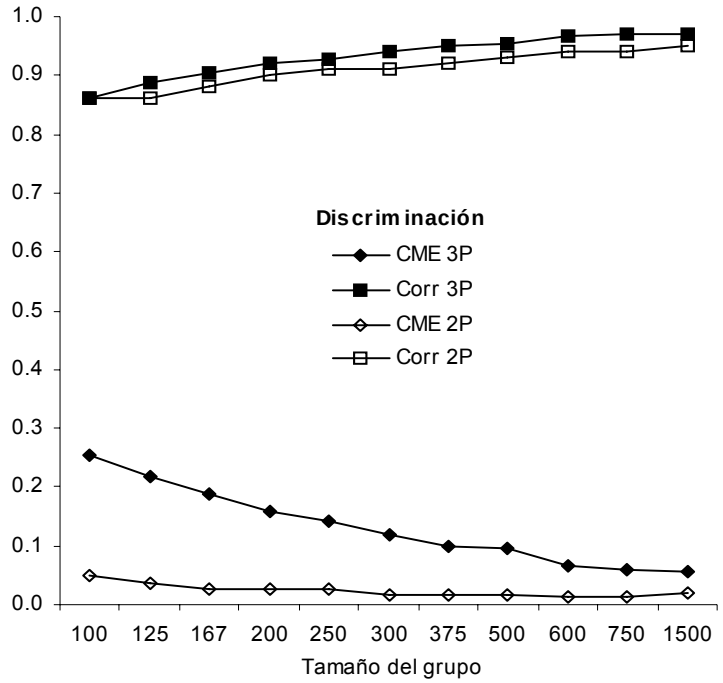


ANEXO 12:

Calidad de las estimaciones de los parámetros IRT cuando se ajustaron modelos de 2 y 3 parámetros

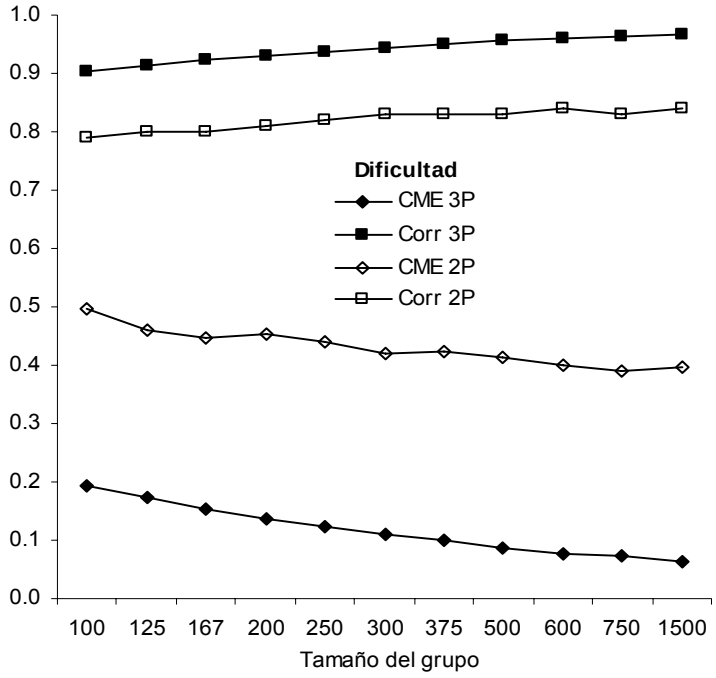
12.1. Medias de los CME y las correlaciones de las estimaciones del parámetro de discriminación en función del tamaño del grupo

Tamaño del grupo	Modelo de 3 parámetros		Modelo de 2 parámetros	
	CME	Correlación	CME	Correlación
100	0.254	0.862	0.049	0.860
125	0.219	0.889	0.037	0.860
167	0.189	0.905	0.028	0.880
200	0.157	0.921	0.027	0.900
250	0.142	0.928	0.026	0.910
300	0.118	0.940	0.018	0.910
375	0.099	0.950	0.016	0.920
500	0.095	0.953	0.015	0.930
600	0.067	0.966	0.014	0.940
750	0.060	0.970	0.014	0.940
1500	0.056	0.972	0.019	0.950



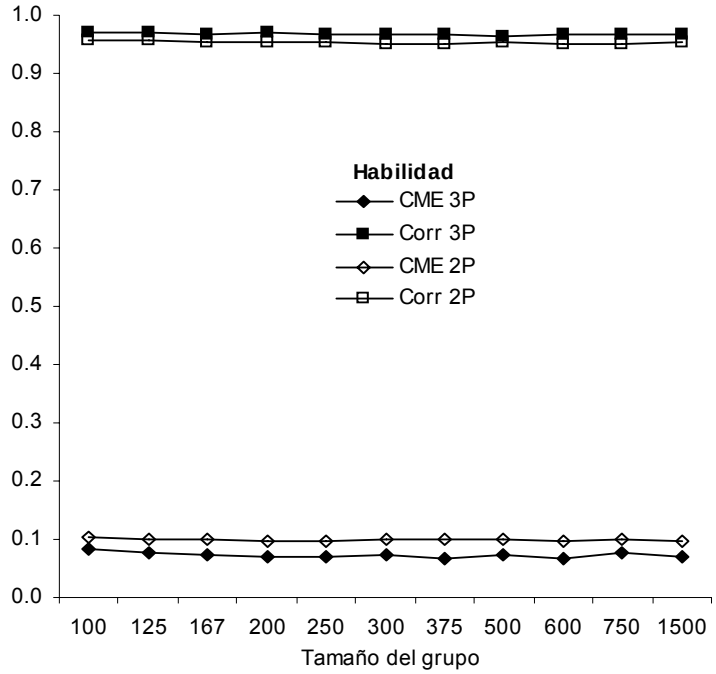
12.2. Medias de los CME y las correlaciones de las estimaciones del parámetro de dificultad en función del tamaño del grupo

Tamaño del grupo	Modelo de 3 parámetros		Modelo de 2 parámetros	
	CME	Correlación	CME	Correlación
100	0.193	0.903	0.496	0.790
125	0.175	0.912	0.460	0.800
167	0.153	0.923	0.447	0.800
200	0.136	0.931	0.452	0.810
250	0.124	0.937	0.440	0.820
300	0.111	0.944	0.421	0.830
375	0.099	0.950	0.424	0.830
500	0.087	0.956	0.414	0.830
600	0.078	0.960	0.399	0.840
750	0.073	0.963	0.390	0.830
1500	0.065	0.967	0.398	0.840



12.3. Medias de los CME y las correlaciones de las estimaciones del parámetro de habilidad en función del tamaño del grupo

Tamaño del grupo	Modelo de 3 parámetros		Modelo de 2 parámetros	
	CME	Correlación	CME	Correlación
100	0.084	0.969	0.102	0.957
125	0.076	0.970	0.101	0.956
167	0.075	0.967	0.101	0.953
200	0.070	0.969	0.098	0.955
250	0.069	0.967	0.098	0.954
300	0.072	0.966	0.100	0.950
375	0.068	0.966	0.100	0.951
500	0.074	0.965	0.099	0.952
600	0.066	0.966	0.097	0.951
750	0.076	0.966	0.100	0.951
1500	0.069	0.967	0.097	0.954

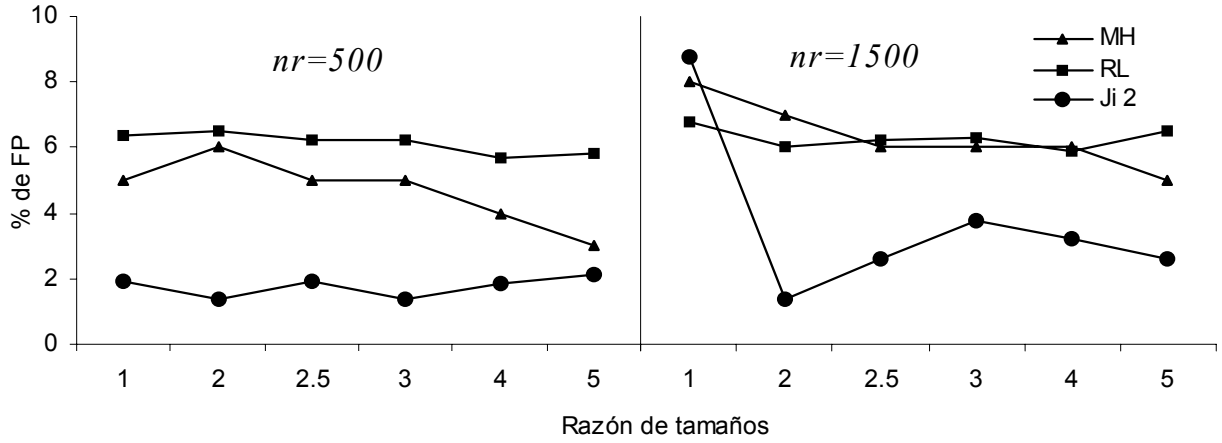


ANEXO 13:

Tasas de detección de los tres estadísticos estudiados

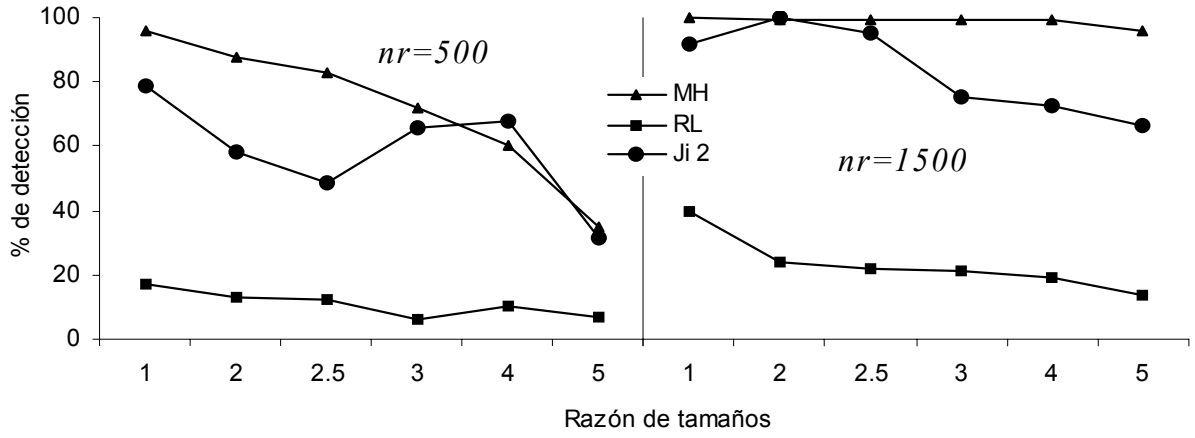
13.1. Medias de las tasas de FP con $\alpha = 0.05$ para los tres estadísticos

Razón de tamaños	*n=500*			*n=1500*		
	MH	RL	Ji^2	MH	RL	Ji^2
1	5.0	6.4	2.0	8.0	6.8	8.8
2	6.0	6.5	1.4	7.0	6.0	1.4
2.5	5.0	6.2	1.9	6.0	6.2	2.6
3	5.0	6.2	1.4	6.0	6.3	3.8
4	4.0	5.7	1.9	6.0	5.9	3.3
5	3.0	5.8	2.2	5.0	6.5	2.6



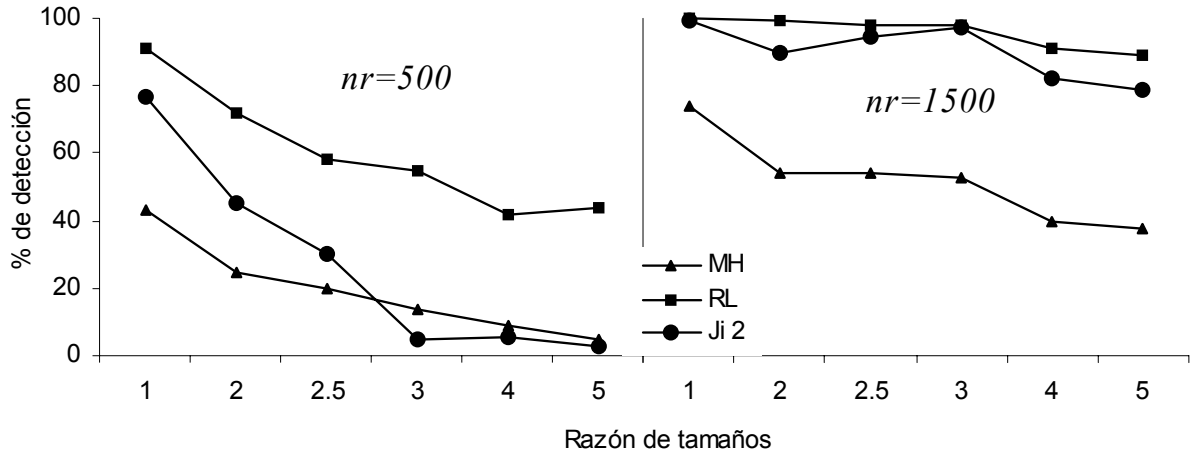
13.2. Medias de las tasas de detección de DIF uniforme con $\alpha = 0.05$ para los tres estadísticos

Razón de tamaños	*n =500*			*n =1500*		
	MH	RL	Ji^2	MH	RL	Ji^2
1	96	17	79	100	40	92
2	88	13	58	99	24	100
2.5	83	12	49	99	22	95
3	72	6	66	99	21	76
4	60	10	68	99	19	73
5	35	7	32	96	14	67



13.3. Medias de las tasas de detección de DIF no uniforme con $\alpha = 0.05$ para los tres estadísticos

Razón de tamaños	*n =500*			*n =1500*		
	MH	RL	Ji^2	MH	RL	Ji^2
1	43	91	77	74	100	100
2	25	72	46	54	99	90
2.5	20	58	30	54	98	95
3	14	55	5	53	98	97
4	9	42	6	40	91	82
5	5	44	3	38	89	79



13.4. Medias de las tasas de detección de DIF mixto con $\alpha = 0.05$ para los tres estadísticos

Razón de tamaños	*n =500*			*n =1500*		
	MH	RL	Ji^2	MH	RL	Ji^2
1	95	97	97	100	100	97
2	91	91	93	100	100	93
2.5	89	87	84	100	100	84
3	89	76	94	99	100	94
4	80	73	93	98	99	93
5	44	62	65	96	97	65