EFECTO DE LA RAZÓN DE TAMAÑOS SOBRE LA DETECCIÓN DEL FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM MEDIANTE REGRESIÓN LOGÍSTICA

ANA CRISTINA SANTANA ESPITIA

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS HUMANAS
DEPARTAMENTO DE PSICOLOGIA
BOGOTÁ
2009

EFECTO DE LA RAZÓN DE TAMAÑOS SOBRE LA DETECCIÓN DEL FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM MEDIANTE REGRESIÓN LOGÍSTICA

ANA CRISTINA SANTANA ESPITIA

Tesis para optar al título de Magíster en Psicología con énfasis en Psicología de la Salud

Director: Aura Nidia Herrera Rojas, Ph. D.

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS HUMANAS
DEPARTAMENTO DE PSICOLOGÍA
BOGOTÁ
2009

Nota de aceptación

_____

_____

_____

_____

Presidente del Jurado

_____

Jurado

_____

Jurado

Bogotá, _____

(día, mes, año)

A Dios y a la Virgen María, por la oportunidad de haber cursado la Maestría y culminarla con éxito.

A mis padres, hermanos, y demás familia, por su confianza y apoyo incondicional.

A Jesús, por creer en mí y por su confianza en la ciencia.

Agradecimientos

A la Universidad Nacional de Colombia, por el apoyo académico y económico brindado en el transcurso de la Maestría.

A Aura Nidia Herrera, por la enseñanza que me proporcionó en medición y evaluación, la cual ha contribuido no sólo a mi orientación académica sino a la formulación de nuevas preguntas y al incremento de la curiosidad en ciencia.

A Víctor Cervantes, por su ayuda en la elaboración de los scripts de simulación y ejecución de la regresión logística.

A Jesús Fajardo y Martha Cuevas, por sus agudas observaciones realizadas a lo largo del presente trabajo.

A Juana Gómez, Constanza Quintero y Olga Rodríguez, jurados de tesis, por sus observaciones y sugerencias en la revisión preliminar de la tesis, así como por las observaciones que puedan surgir posteriores a la sustentación de la misma.

Resumen

La investigación de procedimientos que permitan estimar y consecuentemente reducir la presentación de funcionamiento diferencial del ítem (DIF) en pruebas psicométricas es una de las áreas de estudio más importantes en la medición contemporánea. Swaminathan y Rogers (1990) señalaron que la regresión logística es eficaz en la detección de DIF de tipo uniforme y no uniforme, así mismo se ha reclamado que la regresión provee herramientas para predecir la probabilidad de acierto a un ítem con arreglo a un conjunto de variables (habilidad, pertenencia a un grupo, habilidad*grupo) coordinadas en el modelo de la regresión. El presente estudio se propone establecer cuáles son las variables que inciden en la potencia y error tipo I de la regresión logística, cuando se emplea para detectar DIF. Esta investigación se enmarca en la línea de investigación sobre DIF y se inserta dentro del proyecto en curso titulado "Identificación de ítems con sesgo cultural en las pruebas de Estado ICFES en Colombia". Se usaron los datos de los trabajos anteriores que hacen parte del proyecto de investigación (Arias, 2008; Berrío, 2008) y se obtuvo un diseño experimental completamente cruzado, con 36 condiciones experimentales, producto de manipular las variables razón de tamaños, impacto, porcentaje de DIF y modelo de simulación. Las condiciones se replicaron 500 veces, con un tamaño de muestra n= 65000, que replica las condiciones que se presentan en los escenarios masivos de aplicación de una prueba unidimensional de 30 ítems en el Examen de Estado colombiano (ICFES). Los resultados muestran que una razón de tamaños pequeña (menor diferencia entre el grupo focal y el de referencia), el impacto (diferencias de habilidad no dependientes del sesgo cultural) y el porcentaje de DIF inciden en la potencia del procedimiento de regresión logística. Del mismo modo, se encontró que la incidencia de error tipo I es mayor en las condiciones en las cuales la razón de tamaños es menor y el impacto consiste en diferencias en la media. Finalmente, se discuten algunas implicaciones prácticas y la pertinencia de la regresión logística como procedimiento de detección; igualmente se propone la posibilidad de extender el análisis a los parámetros de los ítems de la prueba, esto es, investigar la influencia de la dificultad y la discriminación de un ítem en la probabilidad de que sea detectado por la regresión logística como un ítem DIF.

Palabras clave: DIF, Regresión Logística, Razón de Tamaños, Diferencias en las medias.

Abstract

Research about procedures to allow estimate and consequently reduce the presentation of differential item functioning (DIF) in psychometrics tests is an area of major interest in contemporary measurements studies. Swaminathan and Rogers (1990) has showed that logistic regression is useful in DIF detection in both uniform and non-uniform type of DIF, also they has claimed that regression offer tools to predict the probability of success to one item with arrangement to a variables set (abilities, membership for a group, abilities*group) coordinate in the regression model. The present study has been proposed to establish what are the variables involved in power and Type I error of Logistic Regression procedure, when is used for detecting DIF. This research is derived of research emphasis about DIF and is part of the project entitled “Identificación de ítems con sesgo cultural en las pruebas de Estado ICFES en Colombia”. The data used were providing by previous works that are part of the project (Arias, 2008; Berrío, 2008) and was obtained a completely cross experimental design with 36 experimental conditions, as result of to manipulate the variables sample size ratio, impact, DIF percentage, and simulation models. This conditions were replicated 500 times, with a sample size of n= 65000, that replies the conditions that are showed in massive applications of a one-dimensional testing in the official Colombian tests (ICFES). The results show that a small size sample ratio (less difference between focal group and reference group), the impact (differences of abilities that not depend of cultural bias) and DIF percentage are involved in the power of the logistic regression procedure. Similarly, the occurrence of type I error is higher in the conditions in that size sample ratio is smaller and there are differences in the mean. Finally, I discuss some practical implications and relevance of the use of logistic regression procedure to detect DIF, also I propose to extend the analysis to the parameters of items, that is to say, research how difficulty and discrimination of an item are involved in the probability of detect it by logistic regression as an DIF item.

Key words: DIF, Logistic Regression, Sample size ratio, Mean differences.

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

La medición de atributos psicológicos a través de instrumentos como pruebas, escalas e inventarios, constituye una de las principales tareas del psicólogo, con el fin de: (a) Obtener una estimación del nivel de habilidad o conocimiento que los individuos poseen en un dominio dado, y (b) realizar inferencias en el marco de procesos de toma de decisiones propios de los escenarios en donde tiene lugar la aplicación de los instrumentos (ej. selección y promoción en ambientes laborales).

El empleo de pruebas psicológicas en entornos educativos y laborales, por citar sólo algunos ejemplos, requiere instrumentos *válidos* cuyos ítems midan adecuadamente el atributo objeto de interés, de manera que sea posible clasificar a los individuos de acuerdo con su nivel de atributo, y tal clasificación no se vea influenciada por variables irrelevantes para la comprensión del atributo, que pueden afectar la calidad de las mediciones y por ende, conducir a inferencias erróneas sobre el desempeño de las personas. Este aspecto resulta de gran importancia cuando se aplican pruebas objetivas a diversos subgrupos poblacionales de un país, escenarios en los cuales algunas variables "intrusas" como género o grupo étnico pueden comprometer la validez del instrumento.

Uno de los escenarios de medición más importante en Colombia es el de la aplicación de las Pruebas de Estado. En el contexto nacional, el Instituto Colombiano para el Fomento de la Educación Superior (ICFES) aplica el Examen de Estado para el Ingreso a la Educación Superior a los estudiantes de último año de educación secundaria del país[1]. El examen tiene como propósitos:

> Servir como un criterio para el Ingreso a la Educación Superior; informar a los estudiantes acerca de sus competencias en cada una de las áreas evaluadas, con el ánimo de aportar elementos para la orientación de su opción profesional; apoyar los procesos de autoevaluación y mejoramiento permanente de las instituciones

[1] Para el año 2008 presentaron el Examen de Estado 148117 egresados y 505714 estudiantes (Datos tomados de la página web del ICFES www.icfes.gov.co).

escolares; constituirse en base e instrumento para el desarrollo de investigaciones y estudios de carácter cultural, social y educativo; y servir de criterio para otorgar beneficios educativos. (ICFES, 2006).

Dada la pluralidad étnica y cultural presente en el país, y al ser el Examen de Estado un reflejo de dicha pluralidad, su aplicación incluye estudiantes de diversas regiones del país, y de grupos étnicos variados (indígenas, afrocolombianos)[2], etnias cuya distribución numérica a lo largo del territorio nacional presenta fuertes contrastes.

En términos de la validez del Examen de Estado, se esperaría que los conocimientos y competencias que pretende medir el examen sean abarcados por los ítems que componen la prueba, de modo que las puntuaciones obtenidas por los estudiantes den cuenta del grado de atributo que poseen, y conduzcan a su vez a inferencias válidas sobre la ejecución de los alumnos y su nivel de competencias requeridas para el acceso a la educación superior.

Si dos estudiantes pertenecientes a grupos étnicos diferentes tienen el mismo nivel de atributo o habilidad, se esperaría que la probabilidad de acertar un ítem fuese igual para ambos. Esto garantizaría el propósito del examen, que es determinar de una manera justa el nivel de habilidad de los estudiantes en el territorio nacional. Si, por otra parte, los estudiantes, aun cuando están igualados por nivel de atributo, difieren en la probabilidad de acierto al ítem, podría pensarse que el ítem se ve afectado por consideraciones irrelevantes para la medición del atributo (ej. pertenencia a un grupo étnico particular), lo que contribuye a crear desventajas para uno de los grupos en cuestión. A este fenómeno se le conoce como *Funcionamiento Diferencial de los Ítems* (DIF por sus siglas en inglés), que alude al estudio de las *propiedades estadísticas* de los ítems que componen los instrumentos de medida. Un ítem presenta funcionamiento diferencial cuando la probabilidad de responderlo correctamente no es igual para sujetos pertenecientes a diferentes grupos, y que a su vez *poseen el mismo nivel de*

---

[2] De acuerdo con el documento *Colombia: Una nación multicultural. Su diversidad étnica* publicado por el Departamento Administrativo Nacional de Estadística (DANE), el 3,43% de la población del país pertenece a grupos indígenas; los afrocolombianos representan un 10,62%, y el pueblo gitano está conformado por 4.858 personas que representan el 0,01% de la población; esto significa que el 14,06% de la población colombiana se identifica como perteneciente a algún grupo étnico (DANE, 2007).

habilidad (Camilli & Shepard, 1994). Hablar de DIF implica que la puntuación obtenida por la persona se ve afectada no sólo por el nivel que los individuos tienen en la variable medida, sino también por otras características *irrelevantes* para el atributo que se está midiendo, como género, etnia, estrato socioeconómico (Camilli & Shepard, 1994).

El estudio estadístico del DIF ha permitido precisar que las personas a las que se le aplica un instrumento de medición, se ubican en uno de estos grupos: Grupo de referencia y grupo focal. Por *grupo de referencia* se conoce al grupo cuyas respuestas se emplean para analizar el funcionamiento del instrumento durante su construcción (Hidalgo, Gómez & Padilla, 2005). Este grupo es generalmente mayoritario y se toma como estándar de comparación del grupo focal (Herrera, Gómez & Hidalgo, 2005; Jodoin & Gierl, 2001). El *grupo focal* es aquel que se considera desfavorecido y generalmente es minoritario (Jodoin & Gierl, 2001), en donde se sospecha que las propiedades psicométricas de los ítems pueden tener valores distintos (Hidalgo et al., 2005).

En todas las situaciones de aplicación masiva de pruebas, un aspecto esencial a considerar es la *razón de tamaños*, que se define como el número de individuos que hay en el grupo de referencia por cada miembro del grupo focal, y se expresa como $r = {nr}/{nf}$, donde *nr* es el número de individuos del grupo de referencia y *nf* el número de individuos del grupo focal. Esta variable ha de tenerse en cuenta en los análisis de DIF, ya que las marcadas disparidades en cuanto a la distribución de los grupos es un factor que podría afectar de manera negativa las puntuaciones de los individuos pertenecientes al grupo focal.

La investigación en DIF en los últimos cincuenta años se ha enfocado en el análisis de las condiciones bajo las cuales se da una correcta identificación de ítems con DIF (potencia), y se controla la identificación errónea de ítems con DIF (Error tipo I). Estos esfuerzos han dado lugar al *análisis de variables, bien de construcción del instrumento, o de la conformación de los grupos de aplicación, que influyen en la detección de DIF*. Entre las primeras se encuentran la longitud de la prueba, el porcentaje de ítems con DIF y la magnitud de DIF; entre las segundas puede

mencionarse la razón de tamaños, siendo ésta una de las variables menos abordadas en estudios con datos simulados y reales.

Otro aporte proveniente de la investigación en DIF es el *desarrollo de técnicas estadísticas* para la detección de funcionamiento diferencial, que se caracterizan por establecer conceptos fundamentales para comprender el DIF, así como por la formulación matemática de la (s) variables a tener en cuenta en los estudios sobre DIF. Una de tales técnicas es la *regresión logística,* propuesta por Swaminathan y Rogers (1990), que tiene como objetivo construir un modelo para predecir la probabilidad de acertar un ítem en función del grado de habilidad de la persona, su pertenencia o no a un grupo particular, y la interacción entre la habilidad de la persona y la pertenencia o no al grupo de interés.

Debido a las características del Examen del Estado, la diversidad de usuarios del instrumento y las inferencias que del mismo se realizan, el Grupo de Investigación "Métodos e Instrumentos para la Investigación en Salud", adscrito al Departamento de Psicología de la Universidad Nacional de Colombia, se encuentra desarrollando desde el 2006 una investigación sobre *Identificación de ítems con sesgo cultural en las pruebas de los Exámenes de Estado en Colombia.* Uno de sus objetivos consiste en identificar los procedimientos de análisis que puedan ser más adecuados dadas las condiciones reales de las pruebas utilizadas dentro del Examen de Estado, referidas a número de examinados por grupo, longitud de las pruebas, tipos de preguntas utilizados, probable proporción de ítems con DIF y parámetros de los ítems.

Los procedimientos de análisis que se incluyeron en la formulación de la investigación fueron la prueba Mantel-Haenszel (MH), la prueba de la diferencia de la dificultad, y la regresión logística, examinando primordialmente el efecto de la razón de tamaños sobre la potencia y el error tipo I de cada uno de estos procedimientos, aplicados a la detección de DIF. Como resultados parciales de la investigación se han producido dos trabajos tendientes a examinar el efecto de la razón de tamaños en el MH (Arias, 2008), y la diferencia de la dificultad (Berrío, 2008).

Este estudio hace parte del proyecto de investigación en curso, y constituye una aproximación alternativa al análisis de la potencia y error tipo I en DIF al emplear una técnica como la regresión logística, que permite elaborar un modelo matemático de la

probabilidad de acierto en un ítem en virtud de ciertas variables y las interacciones entre ellas.

El presente trabajo tiene como objetivo general *evaluar el efecto de las razones de tamaño extremas en el funcionamiento del procedimiento de regresión logística para la detección del funcionamiento diferencial del ítem, a partir de la realización de un estudio de simulación.*

Como objetivos específicos se plantean los siguientes:

1. Establecer las razones de tamaño que proporcionen una adecuada potencia y control del error tipo I de la RL, cuando ésta se emplea para detectar funcionamiento diferencial del ítem.
2. Examinar las relaciones entre la razón de tamaño con diferentes distribuciones de habilidad de los grupos, diferentes porcentajes de ítems con DIF en una misma prueba y diferentes modelos de simulación, y su efecto sobre la potencia y el error tipo I de la regresión logística en la detección de DIF.

Capítulo 1:

REVISION BIBLIOGRÁFICA

Sesgo, impacto y funcionamiento diferencial de los ítems

El estudio de lo que hoy en día se conoce como funcionamiento diferencial de los ítems puede rastrearse desde inicios del siglo XX. Autores como Binet y Stern comenzaron a señalar que algunos ítems de las pruebas de inteligencia eran sensibles al bagaje cultural en entornos de socialización específicos como el hogar y la escuela, afectando la medición de las capacidades intelectuales, para lo cual habían sido originalmente construidos. Sin embargo, es a finales de la década de 1950 y comienzos de los sesenta, bajo el rótulo de *sesgo de los ítems,* que aparecen los trabajos pioneros sobre sesgo en las pruebas. Eells, Davis, Havighurst, Herrick y Tyler (1951) publican un estudio sobre sesgo cultural en los ítems, cuyo propósito fue examinar el grado de familiaridad de los ítems para ciertos grupos socioeconómicos en función de características como la forma o el contenido de los ítems (Camilli & Shepard, 1994). De acuerdo con Cromwell (2006):

> Muchos de los estudios en sesgo fueron conducidos para entender mejor las diferencias culturales entre negros e hispanos y, específicamente, para demostrar que la disparidad en los puntajes de las pruebas tenía que ver más con el sesgo encontrado en los ítems de la prueba que con el nivel de habilidad de una u otra subpoblación (p. 9).

Adicional a lo anterior, Jensen publicó en 1969 un trabajo en torno a examinar la heredabilidad de la inteligencia, y su principal conclusión consistió en señalar que el CI está más determinado por diferencias genéticas que por influencias ambientales. Los estudios de sesgo en los instrumentos de inteligencia acrecentaron la polémica en torno al sesgo de las pruebas.

Desde una perspectiva histórica, la década de los sesenta en los Estados Unidos se caracterizó por el movimiento de la defensa de los derechos civiles (Fidalgo, 1996), que cuestionaba el empleo de las pruebas psicológicas en los procesos de toma de decisiones en escenarios educativos y laborales, al considerar que los resultados de las

pruebas desfavorecían a los grupos minoritarios de población, generando con ello problemas en el acceso a servicios laborales y educativos. Desde entonces, el sesgo se asemejó a cuestiones sociales de *equidad,* de igualdad de oportunidades, en donde las pruebas psicológicas, como criterio de selección o promoción, no eran garante suficiente para asegurar que individuos de diferentes grupos sociales tuvieran las mismas oportunidades de acceso a trabajo o a educación. Cole (1993) señaló frente a esta cuestión que la comunidad psicométrica considera las pruebas como instrumentos neutrales de medición, que no son buenas o malas en sí mismas; sino que las pruebas una vez disponibles para ser aplicadas en escenarios sociales, podrían ser utilizadas de buena o mala manera, lo cual incide de forma importante en la inferencias derivadas de la aplicación de los instrumentos. La perspectiva de sesgo en términos de equidad dio lugar a la aparición de múltiples artículos y métodos estadísticos para analizar el sesgo en las pruebas, a partir de la década de los setenta y posteriores (Figura 1).

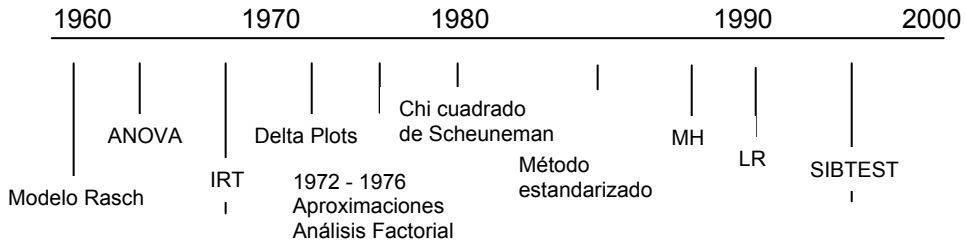


Figura 1. Panorama histórico de la Metodología en DIF 1960-2000 (Adaptado de Cromwell, 2006, p. 10).

Uno de los problemas que se encuentra en la base del estudio sobre sesgo es precisamente lo que se entiende por él, las acepciones que tiene y cuál acepción debe orientar el trabajo de los constructores de pruebas psicológicas. El sesgo es definido por Camilli y Shepard (1994) como "invalidez o *error sistemático* en cómo una prueba mide a miembros de un grupo particular. El sesgo es sistemático en el sentido que crea distorsión en los resultados para miembros de un grupo específico" (p. 8). Jensen (1980, citado por Fidalgo, 1996), proporcionó una definición de sesgo: "En psicometría, el "sesgo" se refiere a *errores sistemáticos en la validez predictiva o en la validez de constructo de las puntuaciones en el test* de individuos que pertenecen a grupos diferentes" (1980, p. 375).

El sesgo, por consiguiente, constituye una amenaza a la validez de constructo, en la medida que algunos ítems no están reflejando adecuadamente el atributo psicológico que se busca medir, afectando con ello las inferencias que se realicen a partir de los resultados de la prueba, con las implicaciones posteriores en el proceso de toma de decisiones en ámbitos laborales y educativos, por citar sólo algunos.

La pregunta por el sesgo en las pruebas, de acuerdo con Camilli y Shepard (1994) implica considerar las *causas* por las cuales los ítems pueden estar funcionando de manera diferente para personas de grupos distintos: “El sesgo es un asunto en el estudio de diferencias de género, étnicas y raciales, o de subgrupos identificados por clase social, edad, región, ambiente urbano vs. rural” (p. 8).

Estudiar el sesgo de las pruebas desde una perspectiva estadística, por su parte, requiere atender a los *efectos* de los resultados obtenidos en las pruebas, por individuos miembros de diferentes grupos. El análisis de estos efectos se apoya en el empleo de estadísticos, que a partir de pruebas de significancia y/o medidas de tamaño del efecto, permiten dar un primer indicio de si los ítems de una prueba presentan resultados dispares para los individuos que contestaron la prueba.

A partir de la controversia entre las implicaciones sociales y de medida del sesgo, se comenzó a desarrollar desde la psicometría un cuerpo de conocimientos tendiente a analizar el sesgo como un problema técnico de construcción de instrumentos, sin las connotaciones políticas, sociales y éticas que se le atribuían al término (Fidalgo, 1996). Fruto de este trabajo, Holland y Thayer (1988) acuñaron el término *Funcionamiento diferencial del ítem*, en aras de una mayor precisión conceptual, que se ha constituido en punto de partida para el desarrollo de un amplio tema de investigación en psicometría (Gómez, Hidalgo, Guilera & Moreno, 2005). Ya en 1982, Angoff (citado por Camilli & Shepard, 1994, p. 16) había señalado la necesidad de identificar aquellas situaciones en las que se intenta esclarecer las propiedades psicométricas de los ítems en distintos grupos, sin entrar en juicios de valor.

El funcionamiento diferencial del ítem (DIF) se ha definido en general haciendo énfasis en la diferencia de las *propiedades estadísticas del ítem cuando se aplica a dos o más subpoblaciones con el mismo nivel de habilidad* (Holland & Wainer, 1993), o haciendo alusión específicamente a la diferencia en la probabilidad de acertar por parte

de individuos que tienen la misma habilidad (Hambleton, Swaminathan & Rogers, 1991). Es decir, la probabilidad de acertar un ítem es diferente para sujetos pertenecientes a grupos diversos, partiendo de la consideración de que todos los individuos poseen el mismo nivel de habilidad en el atributo psicológico a evaluar. En los estudios sobre DIF, la puntuación total de la prueba se ha empleado tradicionalmente como criterio de igualación entre los grupos a comparar, puesto que esta medida proporciona un estimativo de la magnitud de atributo de los individuos en una métrica común.

Si se supone que los individuos están igualados en cuanto a magnitud de atributo, ¿por qué se presentan diferencias en las puntuaciones? Una explicación a ello es que variables ajenas al atributo (ej. raza, género, estrato socioeconómico) están siendo consideradas en la prueba, y por ende, afectan las puntuaciones del instrumento para los individuos de uno u otro grupo. Ackerman (1992) señala como una de las características de DIF las diferencias en la distribución de las variables irrelevantes o "espúreas" (Fidalgo, 1996) para los miembros de dos o más grupos, objeto de evaluación.

Al examinar los conceptos de DIF y sesgo, puede seguirse que la detección de DIF constituye uno de los pasos dentro del proceso general de detección de sesgo, es decir, identificar un ítem que presenta funcionamiento diferencial no implica automáticamente que el ítem esté sesgado para un grupo específico en particular; es preciso realizar análisis posteriores para determinar las causas de sesgo; por ende, la detección de DIF constituye evidencia necesaria pero no suficiente para realizar inferencias sobre sesgo. No obstante, las diferencias entre grupos socioculturales o con características específicas en cuanto al desempeño de un ítem no se deben exclusivamente a variables irrelevantes que incidan en las puntuaciones, sino que estas diferencias pueden ser producto de *diferencias reales* en el nivel de atributo, a estas diferencias se les conoce como *Impacto* (Ackerman, 1992). El objeto de análisis del presente trabajo se centrará en la *identificación de DIF* como primer paso para la posterior evaluación de sesgo cultural en los Exámenes de Estado.

Aspectos generales en el estudio de DIF

La evaluación de DIF en las pruebas psicológicas comprende los siguientes aspectos: (a) Identificación de los grupos a analizar, (b) consideración del tipo de DIF, y (c) elección del método estadístico adecuado para identificar los ítems con DIF.

Con respecto al primer punto, las investigaciones sobre DIF han evaluado como grupos focales las minorías étnicas, particularmente afrodescendientes o inmigrantes, al asumir que estos grupos tradicionalmente han sido subvalorados en cuanto a su desempeño en las pruebas psicológicas. Esto no quiere decir, sin embargo que el grupo focal sea necesariamente minoritario en cuanto a composición demográfica con respecto a la población general, antes bien, puede ser un grupo mayoritario en términos demográficos pero subvalorado en algunas esferas sociales (ej. mujeres).

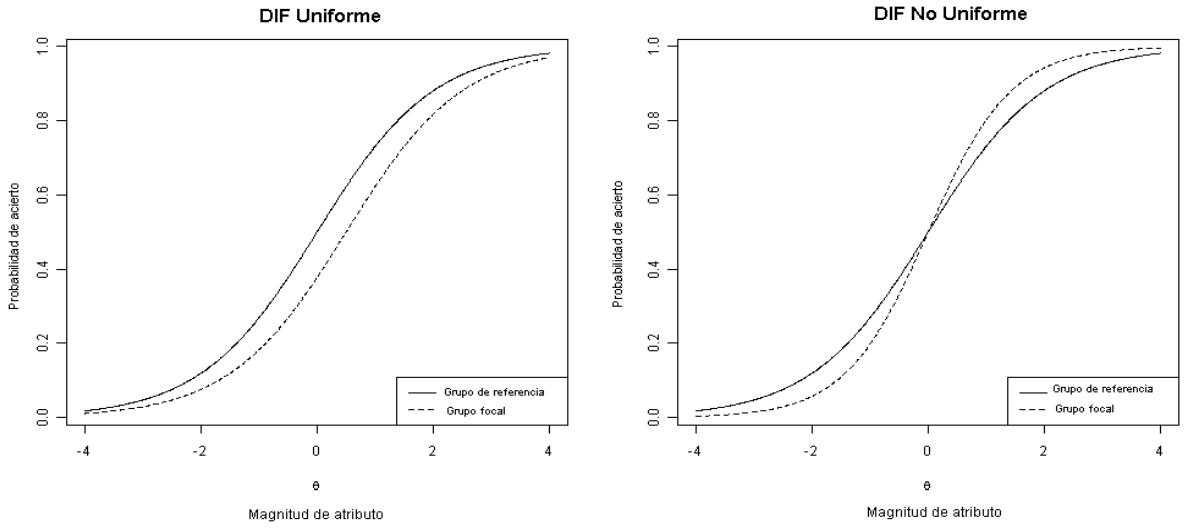
En términos de los grupos de evaluación, un ítem presenta DIF cuando la probabilidad de responder correctamente al ítem es diferente para el grupo focal y el de referencia, y los individuos de ambos grupos pueden equipararse con respecto a su nivel de habilidad o atributo psicológico que se pretende medir. Por lo tanto, los puntajes del ítem estarían reflejando varianza irrelevante para el atributo (factores específicos dados por la pertenencia al grupo: raza, género, idioma), más que la variabilidad del atributo en sí (French & Maller, 2007).

Las diferencias entre los grupos pueden darse en dos sentidos:

1. La probabilidad de responder a un ítem de la manera correcta es mayor para un grupo que para otro, a lo largo del continuo de habilidad, lo que se conoce técnicamente como *DIF uniforme* (Figura 2a). En el DIF uniforme hay una diferencia en el parámetro de dificultad para los dos grupos (Camilli & Shepard, 1994; Hidalgo et al., 2005). En la figura 2a, las curvas características de los ítems (CCI)[3] para los dos grupos son diferentes y no se cruzan a lo largo del continuo de habilidad. Al DIF uniforme en términos de CCI se le conoce también con el nombre de *unidireccional* (Hanson, 1998).

[3] Camilli y Shepard (1994) definen la Curva Característica del Ítem como “La función que relaciona la probabilidad de responder correctamente a un ítem con la habilidad medida por el instrumento que contiene a dicho ítem” (p. 47).

2. La diferencia en las probabilidades de una respuesta correcta para los dos grupos no es la misma en todos los niveles de habilidad (Swaminathan & Rogers, 1990), esto se conoce como *DIF no uniforme* (Mellenbergh, 1982). En el DIF no uniforme se presenta una interacción entre el nivel de habilidad y la pertenencia a un determinado grupo (focal o de referencia) (Zumbo, 1999), de manera que las CCI para los dos grupos son diferentes, pero éstas se cruzan en algún punto de la escala de habilidad (Camilli & Shepard, 1994). En cuanto a los parámetros de los ítems para los dos grupos, en el DIF no uniforme pueden darse diferencias en la discriminación, lo que se conoce como *DIF no uniforme simétrico* (Figura 2b), o bien diferencias tanto en la dificultad como en la discriminación *(DIF no uniforme asimétrico)* (Hidalgo, Gómez & Padilla, 2005). Se observa en la figura 2b que la probabilidad de responder al ítem para los miembros del grupo focal es menor que la del grupo de referencia en niveles bajos de magnitud de atributo (θ<0); por otra parte, cuando el nivel de habilidad es mayor que 0, se aprecia el patrón inverso en la probabilidad de acierto al ítem para ambos grupos.



2a. 2b.

Figura 2. Tipos de DIF. (2a) DIF Uniforme. (2b) DIF no Uniforme.

La investigación en DIF ha buscado desarrollar métodos y estrategias tendientes a aumentar la *potencia* del procedimiento (identificar adecuadamente los ítems con DIF), y en donde se controle al máximo la probabilidad de *Error Tipo I* (identificación errónea de ítems con DIF). Entre las diversas clasificaciones propuestas por múltiples autores

acerca de los métodos de detección de DIF se encuentran la de Camilli y Shepard (1994) y la de Potenza y Dorans (1995).

Camilli & Shepard (1994) realizaron una clasificación de los métodos de detección de DIF en tres categorías:

1. **Los métodos basados en el análisis de varianza y en la Teoría Clásica de los Tests:** En consonancia con los supuestos de la Teoría Clásica de los Tests (TCT), en donde la dificultad de un ítem se mide por la proporción de examinados que aciertan dicho ítem (p-valores), el sesgo de los ítems se equiparó con diferencias en el parámetro de dificultad, lo que dio lugar a la formulación de la técnica del *delta-plot* o índice de dificultad transformada (Angoff, 1972). Este método tenía como objetivo localizar los ítems que maximizan o minimizan las diferencias en la ejecución de los grupos evaluados (Camilli & Shepard, 1994,). Con respecto al ANOVA, Cardall y Coffman desarrollaron en 1964 el primer procedimiento formal de análisis de varianza para la detección de DIF, con el fin de examinar efectos de interacción a doble vía en cuanto a la ejecución del ítem (correcto vs. incorrecto) y la raza (blancos vs. negros) para los datos del SAT (Cromwell, 2006). El ANOVA mantiene la conceptualización de sesgo del ítem en términos de diferencia de dificultad según la TCT. La detección de DIF bajo el enfoque de ANOVA consistía en realizar un ANOVA de medidas repetidas de dos factores, en el que el primer factor era el grupo de pertenencia, y el factor intragrupo estaba conformado por los ítems. De acuerdo con Camilli y Shepard (1994) "las diferencias de medias entre los grupos se reflejaban como efectos principales, mientras que la diferencia en dificultad se reflejaba en los efectos de interacción grupo por ítem" (p. 33).

Los métodos situados en esta primera categoría pronto fueron relegados en la detección de DIF, en la medida en que los procedimientos no conseguían diferenciar sesgo de impacto, además presentaron problemas en cuanto a la equiparación de los grupos (Camilli & Shepard, 1994; Cromwell, 2006; Herrera et al., 2005).

2. **Métodos basados en Teoría de Respuesta al Ítem:** El empleo de modelos de respuesta al ítem en los estudios sobre DIF se inició a comienzos de los años sesenta, con el modelo propuesto por el matemático danés Georg Rasch. El modelo de Rasch establece un sólo parámetro de medición ligado al número de respuestas. Este

parámetro corresponde, de acuerdo con este autor, a la definición de una escala única para evaluar el nivel de atributo de una persona y la calidad de los ítems, calidad abordada en términos de las personas que respondieron correctamente el ítem. (ICI, 1985). Teniendo en cuenta lo anterior, en el modelo de Rasch la probabilidad de acertar un ítem está en función del nivel de habilidad y la *dificultad* del ítem. Posteriormente se formularon los modelos de dos y tres parámetros, en los cuales la probabilidad de acertar el ítem está en función de la dificultad y discriminación, para el caso de dos parámetros; y en función de la dificultad, la discriminación y el pseudoazar, en el caso del modelo de tres parámetros. Los estudios de DIF en la actualidad emplean con mayor frecuencia los modelos de dos y tres parámetros.

La principal función de los métodos de detección de DIF basados en IRT consiste en "determinar si hay una *diferencia en los parámetros* de los ítems entre el grupo focal y el grupo de referencia" (Cromwell, 2006, p. 16). En el marco de los métodos TRI se asume que "un ítem presenta DIF si las funciones de respuesta al ítem no son iguales para diferentes subgrupos, de manera inversa, un ítem no presenta DIF cuando las funciones de respuesta al ítem son iguales para diferentes subgrupos de evaluación" (Hambleton et al., 1991, p. 110). Entre los métodos que se ubican en esta categoría se encuentran la comparación de parámetros, el método Plot, el área entre curvas características del ítem y la evaluación del ajuste en grupos minoritarios a través de estimaciones totales de grupo (Cromwell, 2006). Aun cuando la formulación matemática de los métodos IRT para la detección de DIF es notable, ya que toma en consideración diversos parámetros de análisis, una de las principales limitaciones de estos métodos es la exigencia en cumplimiento de supuestos y tamaño de muestra, lo cual restringe su aplicabilidad en algunas situaciones prácticas (Herrera et al.,2005, p. 32).

3. **Métodos basados en tablas de contingencia:** Bajo esta aproximación hay que señalar que las tablas de contingencia constituyen una estrategia de organización de la información de los individuos que presentaron la prueba y sus respectivos aciertos y errores en los ítems que la conforman, mas no es un método específico de análisis de DIF per se (Camilli & Shepard, 1994). Estos mismos autores afirman además que los métodos basados en tabla de contingencia pueden considerarse *no* paramétricos en la

medida que no invocan un modelo explícito de medida, a diferencia de los modelos TRI empleados en el estudio de DIF.

Los métodos basados en tabla de contingencia, a diferencia de los métodos basados en la TRI son intuitivos, de menor costo computacional y más flexibles en cuanto al cumplimiento de supuestos, lo cual facilita el trabajo en condiciones donde el tamaño de muestra es limitado o donde se requiere calcular índices sencillos, de fácil interpretación y que no demanden grandes costos computacionales (Cromwell, 2006; Herrera et al., 2005).

De acuerdo con Herrera et al. (2005), se distinguen dos enfoques al interior de los métodos basados en el análisis de tablas de contingencia:

(a) Los que se fundamentan en *pruebas de hipótesis sobre igualdad de proporciones mediante el análisis de tablas de contingencia bidimensionales*. En este tipo de métodos se construyen tablas de contingencia en donde se cruza la información de las modalidades de respuesta a los ítems (correcto-incorrecto), con los grupos objeto de análisis (focal y referencia), junto con las proporciones de aciertos y fallos para cada celda de la tabla. Aquí se encuentran algunas aplicaciones de la prueba χ^2, Métodos de Estandarización y la prueba Mantel-Haenszel (MH).

(b) Los que generan *modelos para el análisis de variables categóricas con tablas de más de dos dimensiones*: La característica básica de este tipo de métodos reside en la construcción de modelos con los cuales se pueda predecir la probabilidad de acertar un ítem en función de variables como el nivel de habilidad y el factor de grupo, además permite examinar la bondad de ajuste del modelo a medida que se van introduciendo nuevas variables. En este apartado se incluyen los modelos logit, log-lineales y la regresión logística (RL).

Potenza y Dorans (1995) por su parte, clasifican los métodos DIF de acuerdo con dos criterios: (a) *La manera como se obtiene la variable de equiparación de los grupos* (puntaje total de la prueba vs. estimación del atributo latente que subyace en la ejecución de la prueba); y (b) *si se asumen o no supuestos sobre la forma de la relación entre el puntaje del ítem y la variable de igualación de los grupos*. Cuando se asumen tales supuestos se habla de métodos paramétricos; los métodos no paramétricos no

asumen supuestos a priori sobre la relación entre el puntaje del ítem y la variable de igualación. Ankenmann, Witt y Dunbar (1999) señalan que:

> Bajo este esquema de clasificación, la prueba de Mantel-Haenszel y los métodos de estandarización son considerados métodos no paramétricos basados en el puntaje observado; la regresión logística se considera un método paramétrico basado en el puntaje total de la prueba; el procedimiento SIBTEST se cataloga como método no paramétrico basado en el rasgo latente; y los procedimientos basados en IRT se consideran métodos paramétricos de rasgo latente (p. 278).

Para efectos de este trabajo, se hará una revisión del procedimiento de regresión logística como estrategia para la detección del funcionamiento diferencial del ítem, en el marco de la investigación que el Grupo Métodos e Instrumentos para la Investigación en Salud viene realizando actualmente sobre “Identificación de ítems con sesgo cultural en las pruebas de los Exámenes de Estado en Colombia”.

Capítulo 2:

REGRESIÓN LOGÍSTICA BINARIA Y FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM

La regresión logística binaria o binomial (RL) es un caso particular del modelo de regresión lineal clásico para variables dependientes categóricas dicotómicas (Alderete, 2006). El objetivo general de la regresión logística consiste en determinar el modelo más parsimonioso y mejor ajustado que describa la relación entre un resultado (variable dependiente o variable respuesta), y un conjunto de variables independientes (predictoras o explicativas) (Hosmer & Lemeshow, 1989). De acuerdo con Garson (2006), la regresión logística puede utilizarse para:

1. Predecir la probabilidad de una variable dependiente sobre la base de variables independientes continuas o categóricas, mediante la construcción de un modelo de relación entre las variables.

2. Determinar el porcentaje de varianza en la variable dependiente explicado por las variables independientes.

3. Ordenar por importancia relativa las variables independientes, en función de su contribución a mejorar el modelo.

4. Valorar efectos de interacción entre variables.

5. Analizar el impacto de variables de control que pueden incidir en la probabilidad de ocurrencia del evento.

Las primeras aplicaciones de RL aparecieron en estudios observacionales, de encuesta y experimentales, además del trabajo que se ha adelantado desde la epidemiología (Cornfield, Gordon & Smith, 1961; Silva & Barroso, 2004; Alderete, 2006), y desde la investigación educativa, principalmente en la educación superior. En esta última área, a partir de 1988 se ha observado un incremento en el número de artículos en revistas de educación superior que reportan el empleo de la RL (Peng, So, Stage & St. John, 2002).

La popularidad del modelo de regresión logística para el estudio de diferentes tópicos en ciencias de la salud y ciencias sociales estriba principalmente en dos razones (Kleinbaum, 1994):

1. La función logística que subyace al modelo, que se expresa como $f(y) = \frac{e^z}{1+e^z}$, en donde z simboliza el vector de variables independientes o predictoras, oscila en un rango de 0 a 1, lo cual permite expresar los resultados de la variable dependiente o predicha en términos de probabilidad;

2. La representación de la función logística, que tiene la apariencia de S, permite obtener una aproximación a la naturaleza multivariable del problema y a las relaciones entre las variables predictoras que pueden influir en la probabilidad de ocurrencia del evento de interés.

El empleo de la RL en el análisis de fenómenos de la educación superior se ha extendido principalmente por el reconocimiento de las limitaciones de la regresión por mínimos cuadrados para explicar la complejidad de aspectos de interés en educación como lo son el ingreso, la permanencia y la graduación en las instituciones universitarias. Para estos tópicos usualmente se obtienen medidas categóricas de resultado, y la regresión logística, dada la naturaleza categórica de la variable de respuesta (dicótoma u ordinal), así como las variables continuas y categóricas que se pueden incorporar como variables predictoras de la variable respuesta, constituye una aproximación metodológica de gran utilidad (Peng et al., 2002).

En cuanto a la evaluación sobre funcionamiento diferencial del ítem mediante regresión logística, el uso de la técnica comenzó a reportarse en los trabajos pioneros de Spray y Carlson (1986), Bennet, Rock y Kaplan (1987), y Swaminathan y Rogers (1990), para estudiar diferencias entre grupos en ítems de respuesta dicótoma.

La formulación del modelo de regresión logística en DIF se basa en considerar la probabilidad de acierto a un ítem como función de la *habilidad del sujeto* (generalmente se expresa como el puntaje total obtenido en la prueba), y el *grupo* al cual pertenece el sujeto (focal o referencia). Además de estas dos variables, resulta de interés evaluar *la interacción entre la habilidad y el grupo de pertenencia*, y su influencia en la probabilidad de acierto al ítem.

Teniendo en cuenta lo anterior, la ecuación general de RL (Swaminathan & Rogers, 1990) se expresa como:

$$p\left(y=1\middle|\theta\right)=\frac{e^{z}}{1+e^{z}} \qquad (1)$$

donde $p\left(y=1\middle|\theta\right)$ es la probabilidad de acertar al ítem dado un nivel de atributo θ, y Z es la combinación lineal de las variables predictoras de esa probabilidad de acierto (Hidalgo et al., 2005). Z puede expresarse de la siguiente manera:

$$Z=\beta_0+\beta_1\theta+\beta_2 g+\beta_3\theta g \qquad (2)$$

En esta ecuación θ representa el nivel de habilidad o atributo del individuo (puntuación observada del sujeto en la prueba), g es el grupo al cual pertenece el individuo (referencia o focal), $\theta\, g$ es la interacción entre el nivel de habilidad y el grupo, β_0 representa el intercepto, y β_1, β_2 y β_3 representan los coeficientes para la habilidad, el grupo y la interacción grupo-habilidad, respectivamente. Un ítem se clasifica con DIF uniforme cuando β_2 # *0* y β_3 = 0; un ítem presenta DIF no uniforme si β_3 # *0*, independientemente del valor que asuma β_2 (Swaminathan & Rogers, 1990).

La estimación de los parámetros que acompañan a cada una de las variables en los tres modelos (el intercepto β_0 y los coeficientes β_1, β_2 y β_3), se realiza por el método de máxima verosimilitud, en el que se seleccionan las estimaciones de los parámetros que hagan que los resultados observados sean lo más verosímiles posibles (Kleinbaum, 1994, Alderete, 2006), es decir, que las estimaciones de los parámetros del modelo son los valores que maximizan la función log- verosimilitud (Lemonte & Vanegas, 2005):

$$L(Y,X)=\sum_{k=1}^{n}\left\{Y_k\log\left[\pi(X_k)\right]+(1-Y_k)\log\left[1-\pi(X_k)\right]\right\} \qquad (3)$$

Donde Y_k representa la variable respuesta que asume sólo dos valores, 1 (éxito o acierto al ítem), y 0 (fallo en el ítem), y X_k representa el conjunto de variables que explican o predicen el valor de Y_k. Con estos parámetros, se obtiene el máximo de la función de verosimilitud para cada uno de los modelos, la cual se representa como

$-2LL$ "-2 veces el logaritmo de la verosimilitud", también se le conoce con el nombre de *deviance*. Alderete (2006) afirma que "un buen modelo es aquel que da lugar a una verosimilitud grande, por lo cual el valor de $-2LL$ será pequeño" (p. 57). En el mismo sentido Harrell (2001) señala que un modelo ideal es aquel en el que la verosimilitud es de 1 y el valor del deviance es cero.

Una de las dificultades que se observan en la representación e interpretación de los modelos de variables dicótomas es que la respuesta dada por la probabilidad de un evento no es lineal. Para superar esta dificultad en RL se emplea la transformación logit, que se define como el logaritmo natural de un odds ratio (razón entre la probabilidad de acierto de un suceso (p), y la probabilidad de fracaso del mismo (1-p). La transformación logit se aplica a la variable dependiente (acierto o fallo en el ítem) a fin de expresar una relación lineal entre los resultados de la variable categórica y sus variables explicativas. Por ende, la *regresión logística tiene en cuenta los cambios en el logaritmo natural del odds de la variable dependiente* (Peng et al., 2002). El modelo de regresión logística puede por tanto representarse de la siguiente manera:

$$\log(odds) = LOGIT(P) = \ln\left[\frac{P}{1-P}\right] = Z \qquad (4)$$

En esta formulación, β_0 es el intercepto y Z representa la sumatoria del producto de cada variable predictora X_i (θ , g y $\theta\, g$) con sus respectivos coeficientes β_i (Kleinbaum, 1994).

La regresión logística, a diferencia de otras técnicas empleadas en la detección del DIF, como MH, presenta una serie de ventajas en el análisis del DIF que pueden resumirse así (Swaminathan & Rogers, 1990; Jodoin & Gierl, 2001):

1. *Toma en cuenta la naturaleza continua de la escala de habilidad*, a diferencia del MH, es decir, no requiere categorizar la puntuación total del individuo en la prueba.

2. *Permite modelar DIF uniforme y no uniforme* mediante diferentes estrategias analíticas que implican construcción, ajuste de modelos y prueba de hipótesis sobre la presencia de DIF, bien sea uniforme, no uniforme, o DIF conjunto, según el propósito de la investigación (Swaminathan & Rogers, 1990).

3. Ha demostrado un *poder comparable en la detección de DIF uniforme, comparado con MH y SIBTEST* (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993; Li & Stout, 1996; Zumbo, 1999; Hidalgo & López-Pina, 2004), y un *mayor poder para la detección de DIF no uniforme en comparación con MH* (Swaminathan & Rogers, 1990; Narayanan & Swaminathan, 1996).

4. Puede generalizarse su *uso con ítems de puntuación ordinal* (Zumbo, 1999).

Estrategias de análisis en regresión logística

En la investigación sobre el empleo de la regresión logística para la identificación de DIF, se han propuesto básicamente tres estrategias de análisis, las cuales han sido evaluadas en estudios de simulación bajo diferentes condiciones (Hidalgo et al., 2005).

La primera de ellas se basa en la *comparación de modelos anidados*, y tiene como objetivo evaluar la significación de las variables predictoras que se incorporan sucesivamente al modelo. Para esto, se ajustan tres modelos: El modelo 1 incluye sólo el nivel de habilidad y por tanto representa el comportamiento de un ítem que no presenta DIF; el modelo 2 se compone del nivel de habilidad y el grupo de pertenencia, que se ajustaría a situaciones en las cuales el ítem muestra DIF uniforme; y el modelo 3 o modelo completo, que, además de las variables presentes en el modelo 2, incorpora la interacción entre el grupo y la habilidad; este modelo se ajusta a situaciones en las que el ítem presenta DIF no uniforme.

Con los modelos resultantes, se procede a examinar la significación estadística de las variables que se van incorporando al modelo a través de la comparación de modelos mediante el estadístico de bondad de ajuste para la razón de verosimilitud (Thissen, Steinberg & Gerard, 1986). El procedimiento consiste en la *comparación de dos modelos jerárquicamente anidados*, esto es, un modelo compacto y un modelo aumentado, el cual contiene todos los parámetros del modelo compacto más parámetros adicionales (Ankenmann et al., 1999). Thissen, Steinberg y Wainer (1993) señalan que el objetivo del procedimiento es “probar si los parámetros adicionales en el modelo aumentado son significativamente diferentes de cero” (p. 73). Para ello se emplea el estadístico G^2 que sigue una distribución χ^2 con k grados de libertad,

k resulta de la diferencia de número de parámetros entre los modelos. El G^2 se expresa en la forma (Thissen et al., 1993; Menard, 2000):

$$G^2 = -2\left[\ln\left(L_0\right) - \ln\left(L_M\right)\right] \qquad (5)$$

en donde L_0 representa la función de verosimilitud para el modelo compacto, que contiene sólo el intercepto, y L_M representa la función de verosimilitud para el modelo que contiene todos los predictores.

Para evaluar la presencia de DIF uniforme se evalúa el efecto de la variable grupo comparando la función de verosimilitud del modelo 1 con la del modelo 2, con una distribución $\chi^2 U_{1df}$. Por su parte, el examen de la presencia de DIF no uniforme se realiza analizando el efecto de la variable interacción grupo-habilidad mediante la comparación del modelo 2 con el modelo 3, con una distribución $\chi^2 NU_{1df}$.

La segunda estrategia consiste en realizar *una prueba simultánea de la presencia de DIF uniforme y no uniforme* (Swaminathan & Rogers, 1990). Para ello se parte de la hipótesis $H_0 : \beta_2 = \beta_3 = 0$. En esta estrategia, se procede a ajustar el modelo 1 que incluye sólo el nivel de habilidad, el cual caracteriza a un ítem como no DIF; y el modelo 3, que se compone de la habilidad, el grupo de pertenencia y la interacción entre ambas variables, y se comparan las funciones de verosimilitud de ambos modelos, aquí el G^2 sigue una distribución χ^2_{2df}. Los grados de libertad que acompañan a la prueba de χ^2 son resultado de la diferencia de número de parámetros entre los modelos. A diferencia de la estrategia de modelos anidados, el interés aquí radica en establecer si el ítem presenta o no DIF, independientemente de si éste es uniforme o no uniforme.

La tercera estrategia consiste en *ajustar sólo el modelo 3 (modelo completo) y comprobar la significación de los coeficientes que acompañan a cada una de las variables usando el estadístico de Wald*, que se obtiene dividiendo el coeficiente por su error estándar. De acuerdo con Kleinbaum (1994), este estadístico presenta una distribución aproximadamente normal (0, 1), es decir, una Z en muestras grandes. El cuadrado de este estadístico Z es aproximadamente una χ^2_{1df}. Si sólo el coeficiente β_2 es significativamente diferente de cero, el ítem se clasifica como DIF uniforme;

cuando sólo el coeficiente β_3 es significativamente diferente de cero se estaría hablando de un ítem que presenta DIF no uniforme-simétrico; y cuando ambos coeficientes son significativamente diferentes de cero, se estaría ante un ítem con DIF no uniforme-asimétrico (Hidalgo et al., 2005).

De las anteriores estrategias, la prueba conjunta de DIF uniforme y no uniforme ha mostrado mejores resultados en la detección correcta de ítems con DIF y en el control del error tipo I (Hidalgo et al., 2005), y ha sido ampliamente utilizada tanto en estudios de simulación como en estudios con datos reales.

Análisis de la potencia y error tipo I en regresión logística

La regresión logística, a semejanza de otros estadísticos propuestos para la detección de DIF, como la prueba Mantel-Haenszel, los métodos IRT y el SIBTEST, ha sido evaluada teniendo en cuenta dos criterios esenciales: La potencia y el error tipo I, ambos ligados a la significación estadística de los resultados, es decir, si los efectos observados son consistentes .

En el contexto de DIF, por potencia se define la capacidad de la técnica para identificar correctamente los ítems con DIF en una prueba. La detección de aquellos ítems que presentan funcionamiento diferencial es fundamental para los constructores de pruebas porque permite determinar qué ítems no están reflejando variabilidad del atributo a medir. A partir de procedimientos de purificación de la escala es posible eliminar de la prueba aquellos ítems que presenten DIF, en aras de garantizar la validez de constructo del instrumento y obtener mediciones más *puras* de la habilidad de los individuos.

Por su parte, el error tipo I corresponde a la identificación de ítems no DIF como presentando DIF. El control de este error es un elemento de gran importancia para los constructores de pruebas porque la detección errónea de DIF en ítems que no lo poseen, puede llevar a la supresión de ítems que poseen propiedades psicométricas adecuadas, y por ende, la medición del atributo se ve afectada; además de los sobrecostos que puede significar el ensamblaje de nuevas pruebas una vez se han eliminado ítems que funcionaban apropiadamente o que parecían no funcionar apropiadamente cuando pueden ser buenas medidas.

Durante la década de los 90 y los primeros años de la década actual se han realizado varias investigaciones para la evaluación de la potencia y el control del error tipo I en regresión logística, las cuales han manipulado diversas variables que se ha mostrado inciden en uno o ambos aspectos.

Dentro de estos estudios pueden mencionarse los pioneros de Swaminathan y Rogers, 1990; y Rogers & Swaminathan, 1993; además de algunos que se han ocupado de evaluar el efecto de diferentes variables, como ***tamaño de muestra*** (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993; Hadley, 1995; Narayanan & Swaminathan, 1996; Jodoin & Huff, 2001; Jodoin & Gierl, 2001; Herrera, 2005; Cromwell, 2006; Herrera & Gómez, 2007); ***longitud de la prueba*** (Swaminathan & Rogers; 1990; Rogers & Swaminathan, 1993); ***porcentaje de ítems con DIF*** (Rogers & Swaminathan, 1993; Kennedy, 1994; Narayanan & Swaminathan, 1996; Jodoin & Huff, 2001; French & Maller, 2007); ***magnitud de DIF*** (Narayanan & Swaminathan, 1996; Jodoin & Huff, 2001; Atar, 2007; French & Maller, 2007); **diferencias en habilidad** (Jodoin & Gierl, 2001; Jodoin & Huff, 2001; Finch & French, 2007; French & Maller, 2007); **modelos de simulación** (Rogers & Swaminathan, 1993; Finch & French, 2008); **manipulación de los niveles de dificultad y discriminación** (Rogers & Swaminathan, 1993; Narayanan & Swaminathan, 1996; Herrera, 2005; Herrera & Gómez, 2007); **tipo de estrategia de análisis** (Jodoin & Gierl, 2001; Hidalgo et al., 2005); y **empleo de estrategias de purificación y medidas de tamaño del efecto** (Zumbo, 1999; Jodoin & Gierl, 2001; Hidalgo & Gómez-Benito, 2003; Hidalgo & López-Pina, 2004; Kanjee, 2005).

En general, estos estudios han mostrado las siguientes tendencias:

1. Las condiciones de impacto, es decir, donde existen diferencias en la distribución de la habilidad de los grupos, bien sea en la media, la desviación estándar o en ambas, se asocian con el descenso en la tasa de detecciones correctas de ítems DIF, y con el incremento de las tasas de error tipo I.

2. A mayor longitud de la prueba, aumenta la probabilidad de detección de DIF, en la medida en que a mayor número de ítems se dispondrá de un estimativo más confiable que dé cuenta de la magnitud del atributo del individuo, además el puntaje total obtenido en la prueba se emplea como criterio de igualación entre los grupos focal y de referencia.

3. En relación con el porcentaje de DIF, a medida que se incrementa el porcentaje de ítems con DIF en la prueba, se observan decrementos en la potencia y un aumento de la tasa de falsos positivos.

4. Magnitudes altas de DIF, esto es, la cantidad de DIF que posee un ítem, influyen en el aumento de las tasas de detecciones correctas.

5. En las investigaciones sobre detección de DIF con el empleo de la RL, usualmente se toman modelos de 2 y 3 parámetros.

6. En cuanto a la estrategia de análisis, la mayoría de estudios han empleado la prueba conjunta de DIF propuesta por Swaminathan & Rogers (1990). Las mejores tasas de detección se encuentran en la prueba conjunta de DIF ($\chi^2{}_{2df}$), seguida de la prueba de DIF uniforme ($\chi^2 U_{1df}$), y la prueba de DIF no uniforme ($\chi^2 NU_{1df}$).

7. El estudio del efecto de los parámetros de dificultad y discriminación en la detección de DIF en regresión logística indica que mayores tasas de detecciones correctas se encuentran en ítems con alta discriminación, y dificultad baja o media. Asimismo, esta combinación de parámetros se ha asociado con incrementos en la tasa de falsos positivos.

8. El empleo de la purificación en el análisis de la regresión logística representa un incremento del poder en 13% y una disminución en el error tipo I de 2%, aproximadamente (French & Maller, 2007).

9. El uso exclusivo de las pruebas de significación estadística en la detección de DIF conduce a altas tasas de error tipo I. Cuando se acompaña la prueba estadística de una medida del tamaño del efecto, se controla la tasa de falsos positivos, aunque se observa una reducción en la potencia del procedimiento.

No obstante la amplia literatura sobre regresión logística en DIF que documenta la influencia de estas variables, la razón de tamaños, definida como el número de individuos que hay en el grupo de referencia por cada miembro del grupo focal, y se expresa como $r = nr/nf$,donde *nr* es el número de individuos del grupo de referencia y *nf* el número de individuos del grupo focal, aparece estudiada en pocos trabajos (Herrera, 2005; Cromwell, 2006; Atar, 2007; Herrera & Gómez, 2007) ¿Por qué evaluar el efecto de la razón de tamaños en regresión logística para la detección de DIF? será la pregunta que se responderá a continuación.

Capítulo 3:

RAZÓN DE TAMAÑOS EN REGRESIÓN LOGÍSTICA

Las investigaciones y trabajos que se han adelantado sobre regresión logística en DIF han estudiado el tamaño de muestra (número total de individuos) como una de las variables que influye de modo importante en la detección de DIF. Con la elección de ciertos niveles de tamaño muestral, se procede a efectuar combinaciones para establecer el número de individuos para el grupo focal y de referencia. El análisis de estas combinaciones a su vez permite determinar las condiciones más óptimas de detección correcta y control de falsos positivos.

Los hallazgos de estos estudios pueden resumirse en lo siguiente:

1) *El poder para la detección de DIF aumenta conforme aumenta el tamaño de los grupos, el número de individuos en el grupo focal es mayor y no hay una gran diferencia entre los tamaños de los grupos focal y referencia*, visto esto mediante la evaluación de diferentes distribuciones de la habilidad de los individuos, tipo de DIF y magnitud de DIF (Hadley, 1995; Jodoin & Gierl, 2001; Jodoin & Huff, 2001).

Swaminathan y Rogers (1990) reportan tasas de detección correcta del 75% para DIF uniforme con tamaños de muestra de 250 sujetos en ambos grupos, y del 100% para 500 sujetos. En cuanto al DIF no uniforme, RL es capaz de detectar el 50% de los ítems con DIF para pruebas cortas y tamaño de muestra de 250 sujetos, y el 75% de los ítems en el caso de pruebas largas y con tamaño de muestra de 500 sujetos en ambos grupos. En otro estudio, Rogers y Swaminathan (1993) señalan un incremento del 15% en la detección de DIF uniforme cuando el tamaño de muestra se incrementa de 250 a 500 sujetos. Para DIF no uniforme se observó un incremento del 19% en las tasas de detección a medida que incrementa el tamaño de muestra. French y Maller (2007) apuntan también a un aumento en la tasa de detecciones correctas cuando $nr = 1000$ y $nf = 500$ para DIF no uniforme, en condiciones con alta magnitud de DIF. En relación con distribuciones de habilidad y tamaños de muestra, se observan mayores tasas de detecciones correctas cuando los tamaños de ambos grupos son grandes y no existen diferencias en la distribución de habilidad (Jodoin & Gierl, 2001). Así mismo, en estudios que han tomado un criterio previo para la selección de proporciones

adecuadas de poder (French & Maller, 2007), se ha encontrado que alcanzan este criterio aquellas condiciones con tamaños de muestra grandes.

2) *El error tipo I tiende a ser mayor cuando aumenta el tamaño de los grupos, la magnitud de DIF es grande y hay un alto porcentaje de ítems con DIF* (Kennedy, 1994; Hadley, 1995; Narayanan & Swaminathan, 1996; French & Maller, 2007).

Jodoin y Gierl (2001) reportan incrementos en el error tipo I del 7,3% al 10,3% cuando *nr = 1000,* y *nf = 1000*, con 10% y 20 % de DIF, y sin diferencias en la distribución de habilidad. Cuando los grupos difieren en habilidad el error tipo I para esta condición se incrementa, llegando a un 13,1% y 15.8% para 10% y 20% de DIF, respectivamente. Swaminathan y Rogers (1990), por su parte, en su análisis preliminar del procedimiento de regresión logística, encontraron tasas de error tipo I entre 1 y 6%. En su análisis de diferencias en la distribución de habilidad en RL sobre el error tipo I, Jodoin y Huff (2001) señalan incrementos en la tasa de error tipo I con el aumento del tamaño de muestra y el empleo de la prueba χ^2_{2df} sin considerar el tamaño del efecto.

3). *Los tamaños de muestra estudiados corresponden bien a igual número de individuos en los grupos focal y de referencia* (Swaminathan & Rogers, 1990; Hadley, 1995), *o a combinaciones de niveles muestrales para el grupo focal y de referencia* (Tian, 1999; Aguerri, Galibert, Attorresi & Prieto, 2007; Atar, 2007; Cromwell, 2007). Un ejemplo puede observarse en French y Maller (2007) con n_r = *200, 500* y *1000,* y n_f = *100, 500* y *1000, donde* $n_f \leq n_r$). En estudios con datos simulados, se ha empleado un tamaño de muestra máximo de 3000 individuos (Herrera, 2005).

4). *El número de réplicas de las condiciones experimentales oscila entre 20 y 100 réplicas, sólo de la condición más extrema* (Swaminathan & Rogers, 1990), *o de todas las condiciones experimentales*, característica de estudios posteriores.

Al examinar esos resultados, se aprecia que el tamaño de muestra es una variable crucial que incide en la significación estadística del procedimiento. No obstante, el análisis de regresión logística enfocado sólo en el tamaño de muestra deja de lado dos consideraciones importantes: En primer lugar, existen diversos contextos de evaluación y aplicación de pruebas, especialmente aquellas de aplicación masiva en las que se evalúa un gran número de aspirantes procedentes de distintos grupos étnicos y culturales, grupos cuya distribución no se da de manera proporcionada. Por

ende, la generalización de los resultados obtenidos en estudios de simulación con tamaños iguales para el grupo focal y referencia a estos escenarios de evaluación se ve comprometida. Segundo, las razones de tamaño que pueden extraerse de los estudios simulados en donde se ha examinado el tamaño de muestra son pequeñas: 1:1 (Swaminathan & Rogers, 1990; Hadley, 1995); 1:1 y 2:1 (Tian, 1999; Atar, 2007); 1:1, 2:1 y 4:1 (Jodoin & Gierl, 2001); 1:1, 2:1, 5:1 y 10:1 (French & Maller, 2007), asumiéndose con ello que hay poca o ninguna diferencia en el número de individuos que conforman el grupo focal y el de referencia, situación que no siempre es cierta en la práctica, donde pueden encontrarse diferencias notables en la proporción de individuos del grupo focal en relación con el grupo de referencia.

La hipótesis de un posible efecto de la razón de tamaños sobre el poder y el error tipo I en regresión logística fue sugerida por Narayanan y Swaminathan (1996), a partir del análisis del tamaño de muestra sobre el poder de MH, RL y SIBTEST, combinando dos tamaños de n_r = *500* y *1000*, y dos tamaños de n_f = *200* y *500.* Desde el año 2005 se comienza a observar un incremento de los estudios simulados en DIF que consideran el efecto de razón de tamaño sobre la RL (Herrera, 2005; Cromwell, 2006; Atar, 2007), y sobre otros procedimientos como el Mantel-Haenszel (Herrera; 2005; Herrera & Gómez, 2007; Akelo, 2008; Arias, 2008), la diferencia de dificultad (Berrío, 2008), el χ^2 de Lord (Herrera, 2005); y el SIBTEST (Akelo, 2008). A continuación se mencionarán algunos de los hallazgos más relevantes de las investigaciones enfocadas a examinar el efecto de la razón de tamaños sobre la regresión logística.

En el estudio de Herrera (2005), que tuvo como objetivo evaluar el efecto del tamaño de muestra y la razón de tamaños en la detección de DIF para tres procedimientos (MH, RL y χ^2 de Lord), se tomaron dos tamaños de muestra del grupo de referencia (n_r = *500*, *1500*) y 6 niveles de razón de tamaños de los grupos (1, 2, 2.5, 3, 4 y 5), para obtener un total de 12 condiciones experimentales que fueron replicadas 100 veces. Los resultados obtenidos reflejan un efecto significativo de la razón de tamaños en el error tipo I de la RL, para ítems con baja discriminación y una interacción significativa entre la razón de tamaños y el tamaño de muestra, para ítems con baja dificultad. El porcentaje promedio de falsos positivos de ítems sin DIF no aumentó sistemáticamente con el aumento del tamaño del grupo de referencia o con una

disminución de la razón de tamaños. En relación con la potencia, la razón de tamaños, el tamaño de muestra y la interacción entre estos factores fueron significativos para DIF uniforme, no uniforme y mixto. Con grupos de referencia de 1500 examinados las tasas de detección de DIF no uniforme y mixto fueron altas (mayor de 0.90), independientemente de la razón de tamaños. Por otra parte, se observó una disminución en las tasas de detección correctas con el aumento de la razón de tamaños, es decir, en aquellas condiciones donde el tamaño del grupo focal es muy pequeño.

Entre las conclusiones de este estudio, se señala que los usuarios que empleen regresión logística como técnica de detección de DIF tendrán un adecuado control de error tipo I cuando la prueba estadística se acompaña del empleo del procedimiento de purificación, incluso con grupos de 1500 examinados y grupos focales iguales o hasta 5 veces menores; y obtendrán además un poder satisfactorio de la RL para la detección de DIF con grupos de referencia grandes. Sin embargo, cuando el tamaño del grupo de referencia sea de 500 examinados, se sugiere que el grupo focal sea del mismo tamaño o al menos de 200 examinados para lograr unas adecuadas tasas de detección de DIF no uniforme y mixto, respectivamente.

A partir del estudio de Herrera (2005) y del trabajo de Herrera y Gómez (2007), se han desarrollado otros dos trabajos en torno a explorar el efecto de razones de tamaños extremas para otros procedimientos de detección de DIF [MH (Arias, 2008), y diferencia de dificultad (Berrío, 2008)], como parte de la investigación en curso sobre “Identificación de ítems con sesgo cultural en las pruebas de Estado ICFES en Colombia”.

En el estudio de Arias (2008) con el MH, se fijó un tamaño de muestra de 130000 individuos y se manipularon 5 razones de tamaño (20:1, 100:1, 250:1, 500:1, 1000:1), además se tomaron en consideración otras variables como diferencias en la distribución de la habilidad de los individuos, modelo de simulación y porcentaje de ítems con DIF. Los hallazgos obtenidos en relación con el control de error tipo I muestran que el empleo de ΔMH contribuye a la disminución de falsos positivos, principalmente en las razones de tamaño más grandes (250, 500 y 1000), cuando los datos se simulan con el modelo de un parámetro y hay impacto en la media. Por su parte, la potencia del MH

disminuye cuando aumenta la razón de tamaños. La razón de 250:1 es aquella razón extrema en la cual se observa una adecuada potencia del procedimiento. Condiciones con impacto en la media y 10% de DIF presentan aceptables tasas de detección de DIF.

En la evaluación de la razón de tamaños mediante procedimientos IRT (diferencia de dificultad), bajo las mismas condiciones del anterior estudio, Berrío (2008) encontró que:

"El procedimiento arroja tasas adecuadas de falsos positivos y detección correcta cuando la razón de tamaños oscila entre 1/500 y 1/20, bajo cualquier condición de diferencias entre grupos, con 10% de ítems con DIF, DIF uniforme y modelos ajustados. Cuando la prueba presenta 20% de ítems con DIF y DIF uniforme es preferible que la razón de tamaños se encuentre entre 1/20 y 1/100, manteniendo las demás condiciones constantes" (p. 75).

Con respecto a las tasas de falsos positivos, tasas mayores a 0.60 se presentaron en razones de 20 y 250, cuando hay desajuste y diferencias en media y media y desviación estándar. El análisis de varianza del error tipo I reporta efectos principales e interacciones de la razón de tamaños, modelo de ajuste e impacto, Derivado de lo anterior, se observó una inflación de error tipo I mayor que 0.075 en todas las condiciones, excepto cuando los grupos presentan ajuste y no hay impacto

En relación con la potencia del procedimiento, Berrío (2008) menciona que para las condiciones de 10% y 20% de DIF, a medida que la razón de tamaños va disminuyendo, las tasas de detección correcta aumentan. Cuando aumenta el porcentaje de DIF, la tasa de falsos positivos decrece. Los análisis de varianza para potencia reportaron efectos de la razón de tamaño, razón de tamaño * impacto, e interacción de las tres variables independientes, en la condición de 10%. Para la condición de 20% se reportaron efectos de la razón de tamaño y de la interacción impacto * ajuste.

El estudio de Cromwell (2006) tuvo como uno de sus objetivos principales examinar en qué medida 4 condiciones de tamaño de muestra ($300_R/300_F$ y $1000_R/300_F$; $500_R/100_F$ y $1000_R/100_F$), 3 distribuciones de habilidad (M_F =0, SD_F = 1 / M_R =0, SD_R =1; M_F =1, SD_F =1 / M_R =1, SD_R =2; M_F =0, SD_F =1 / M_R =1, SD_R =2), y 5 formas de la

distribución en las poblaciones (asimetría=0, curtosis=0; asimetría =1, curtosis =0.5; asimetría =0.5, curtosis =0.5; asimetría =0, curtosis =3; asimetría =0, curtosis =-1.0), afectaban la detección de DIF, empleando MH y RL; para un total de 60 condiciones experimentales que fueron replicadas 200 veces.

Entre los hallazgos reportados respecto a la incidencia de las condiciones de tamaño de muestra y la disparidad en cuanto a los grupos de comparación, se reporta que la detección de MH siempre fue mejor que la de RL en todas las combinaciones de razón de tamaño; sin embargo, cabe anotar que el tipo de DIF analizado fue sólo de tipo uniforme, donde el MH presenta mayor solidez. De las cuatro combinaciones de tamaños de muestra considerados, hay una mejor detección de DIF en la condición 500/100, y el peor desempeño fue en la condición de 1000/300, para RL, es decir, *la potencia es mayor cuando existe poca diferencia entre el tamaño de los grupos referencia y focal*. Una de las conclusiones planteadas por Cromwell es que "estos resultados indican que el tamaño total de la muestra es más relevante que el porcentaje de diferencia en el tamaño de muestra entre el grupo de referencia y el grupo focal" (p. 126).

Ante la pregunta de si la interacción entre los niveles de las variables manipuladas influye en la potencia y error tipo I en RL, se encontró que la interacción entre tamaño de muestra y niveles de habilidad fue la única que resultó estadísticamente significativa. Al examinar a través de combinaciones de tamaños de muestra se observó una tendencia lineal, en la medida que si se aumenta el porcentaje de diferencia entre los grupos de examinados, el valor p se incrementa, por consiguiente, la detección de DIF es menor. En relación con las distribuciones de habilidad, se observa que en niveles moderados de la distribución de habilidad (M_F=1, SD_F=1 / M_R =1, SD_R =2), hay una mayor detección de DIF a lo largo de las combinaciones de tamaño de muestra, en comparación con niveles normales y altos de diferencias en distribución de habilidad.

Finalmente, Atar (2007) realizó un estudio de simulación que tuvo como propósito comparar la ejecución de dos aproximaciones diferentes en la detección de DIF (prueba de razón de verosimilitud en TRI y RL) para tres tipos de pruebas (pruebas compuestas por ítems dicótomos, polítomos e ítems mixtos), a partir del análisis de tres

condiciones de tamaño de muestra (n= 600; n= 1200; y n= 2400) que representan tamaño de muestra pequeño, mediano y grande, respectivamente; dos niveles de razón de tamaño (1:1 y 2:1) y tres niveles de magnitud de DIF (0.32, 0.43 y 0.53), que representan magnitud baja, media y alta, respectivamente; para un total de 18 condiciones experimentales que fueron replicadas 200 veces. Entre los resultados obtenidos por el presente estudio cabe anotar que en general las tasas de error tipo I con la prueba de razón de verosimilitud IRT y con regresión logística se encontraron dentro del valor esperado para todas las combinaciones de tamaño de muestra y magnitud de DIF. Hubo una ligera inflación del error tipo I para la combinación de tamaños de muestra medios (n = $600_R/600_F$ o $800_R/400_F$) y magnitud media o alta de DIF (0.43 o 0.53), y para tamaños de muestra grandes (n= $1200_R/1200_F$ o n = $1600_R/800_F$).

Con respecto al efecto de la razón de tamaños sobre el error tipo I, bajo razones de tamaño iguales (1:1) o desiguales (2:1), el error Tipo I de RL y de la razón de verosimilitud IRT se incrementa a medida que aumenta el tamaño de muestra, excepto en condiciones donde la razón de tamaño no es igual y la magnitud de DIF es pequeña. Bajo razones de tamaño iguales (1:1) o desiguales (2:1), el error Tipo I de RL y de la razón de verosimilitud IRT se incrementa a medida que aumenta la magnitud del DIF, excepto en condiciones donde la razón de tamaño no es igual y el tamaño de muestra es pequeño (n= 600).

En cuanto a la potencia, RL y la prueba de razón de verosimilitud IRT obtuvieron índices adecuados de potencia al nivel nominal de 0.05 para todas las condiciones. No obstante, el poder de ambos procedimientos fue inferior a 0.80 en las combinaciones de tamaños pequeños de muestra (n= 600, con igual y distinta razón de tamaño, y magnitud pequeña y mediana de DIF (0.32, 0.43), así como para la combinación de tamaño de muestra medio (n = $800_R/400_F$) y magnitud pequeña de DIF. Por otra parte, la potencia de ambos métodos de detección de DIF (índice mayor a 0.80) se observó en las condiciones de tamaño de muestra grande (n=2400, con razón de tamaño 1:1 y 2:1), alcanzando el valor de 1 cuando los tamaños de muestra grande se toman en conjunción con magnitudes moderadas y altas de DIF. La autora concluye que:

“El tamaño de muestra, la razón de tamaños y la magnitud de DIF fueron factores efectivos en el poder de la prueba de razón de verosimilitud IRT y la regresión logística. Bajo condiciones de igual razón de tamaño (1:1) o diferente razón de tamaño (2:1), el poder de ambos procedimientos se incrementa con el aumento en el tamaño de muestra o la magnitud de DIF. Razones de tamaño iguales proporcionan ligeramente mayor poder que las razones de tamaño diferentes (2:1)” (p.72).

Como puede apreciarse, las investigaciones sobre la razón de tamaños en técnicas de detección de DIF están comenzando a considerar grandes razones de tamaño, aplicadas a amplios tamaños de muestra, de manera que se simulen de la manera más fidedigna situaciones reales de evaluación en donde hay un gran número de individuos que son tomados como referente y un grupo pequeño de individuos que en virtud de ciertas características, tales como grupo étnico, estrato socioeconómico, etc, representa una minoría de la población y puede resultar desfavorecido si los ítems no reflejan la variabilidad inherente al atributo que se desea medir. Además se observa en estos estudios que el error tipo I y la potencia se incrementan cuando aumenta el número de individuos en el grupo focal, además en algunas razones de tamaño el procedimiento presenta una mayor potencia y control de error tipo I, por tanto proporcionan una guía de trabajo al personal encargado del diseño y análisis de ítems para emplear uno u otro procedimiento de detección de DIF teniendo en cuenta ciertas características. La investigación en regresión logística por tanto, no debe ser ajena a estos adelantos y el análisis de la razón de tamaños bajo RL ha de incluir:

1. Definición de un tamaño de muestra apropiado que permita una aproximación a contextos reales de evaluación.

2. Manipulación de varios niveles de razón de tamaño, con el propósito de examinar diversas composiciones de los grupos focal y de referencia.

3. Selección de una estrategia de análisis que posibilite el ajuste de uno o más modelos, de acuerdo con los intereses de la investigación.

4. Elección de una prueba estadística adecuada que con sus correspondientes niveles de significación, conduzca a la identificación correcta de ítems con DIF.

5. Control de error tipo I mediante el empleo de estrategias de purificación, que eliminan del cálculo de la habilidad los ítems identificados con DIF, y por ende, se obtiene un indicador más preciso del nivel de habilidad del sujeto.

6. Examen de otras variables que en conjunción con la razón de tamaños, inciden en la tasa de detecciones correctas y de falsos positivos.

Teniendo en cuenta la revisión anterior, la tarea de proponer métodos para la detección de DIF que tengan en cuenta la razón de tamaños es un objetivo importante, por lo tanto, el presente trabajo tiene como objetivo general *evaluar el efecto de las razones de tamaño extremas en el funcionamiento del procedimiento de regresión logística para la detección del funcionamiento diferencial del ítem, a partir de la realización de un estudio de simulación.*

Como objetivos específicos se plantean los siguientes:

1. Establecer las razones de tamaño que proporcionen una adecuada potencia y control del error tipo I de la RL cuando se emplea para detectar funcionamiento diferencial del ítem.

2. Examinar el efecto de las variables porcentaje de DIF, modelo de simulación e impacto en la potencia y el error tipo I de la RL cuando ésta se emplea para detectar DIF.

3. Examinar las relaciones entre la razón de tamaño y diferentes niveles de distribuciones de los grupos, porcentaje de ítems con DIF en una misma prueba y modelo de simulación, en el marco de la regresión logística.

Capítulo 4:

MÉTODO

Se llevó a cabo un experimento de Monte Carlo con el objeto de analizar los factores que inciden en las tasas de detecciones correctas e incorrectas de la regresión logística, cuando ésta se emplea en el análisis de DIF. Los estudios de Monte Carlo realizados en investigaciones sobre teoría de respuesta al ítem (IRT) proporcionan información sobre la pertinencia de la aplicación de métodos IRT o metodologías (ej. DIF) en conjunción con IRT en datos extraídos de contextos cotidianos de evaluación (Harwell, Stone, Hsu & Kirisci, 1996; Cohen, Kane & Kim, 2001). Estos autores señalan además la estructura que debe poseer un estudio de Monte Carlo en el marco de IRT, la cual es aplicable a estudios sobre DIF con datos simulados (p. 102, 105):

1. **Formulación del problema de investigación**, mediante el establecimiento del problema a indagar, la(s) pregunta(s) de investigación que definen el propósito del estudio, las hipótesis a probar y los efectos a medir.

2. **Diseño del estudio de Monte Carlo**, que incluye la selección de variables dependientes e independientes, el diseño experimental, el número de réplicas y el modelo IRT que subyace a la generación de los datos, a fin de maximizar la replicación y generalización de los resultados.

3. **Estimación de parámetros** a partir de las matrices simuladas de respuestas.

4. **Análisis de los resultados del estudio de Monte Carlo**, con el apoyo de métodos descriptivos e inferenciales. La selección de los análisis inferenciales está guiada por las preguntas de investigación y el diseño experimental.

Los estudios con datos simulados en la investigación sobre DIF permiten controlar algunos factores como el tamaño de muestra y el porcentaje de DIF (French & Maller, 2007). Además, de acuerdo con Nayaranan y Swaminathan (1996), sólo con datos simulados es posible disponer de niveles previos de una variedad de factores, así como determinar adecuadamente las tasas de detección bajo condiciones simuladas por el investigador.

Diseño

El presente estudio se inscribe en la línea de investigación sobre funcionamiento diferencial del ítem, y hace parte del proyecto de investigación en curso titulado "Identificación de ítems con sesgo cultural en las pruebas de Estado ICFES en Colombia", y teniendo en cuenta que se han realizado dos estudios previos de evaluación de técnicas específicas en la detección de DIF (Arias, 2008; Berrío, 2008), se decidió emplear los datos que se simularon para el proyecto de investigación y que fueron utilizados por los dos estudios anteriores. El uso de los mismos datos garantiza la comparabilidad de los resultados con miras a lograr el propósito general del proyecto de investigación, que consiste en identificar los mejores procedimientos de detección de DIF en contextos masivos de aplicación, a fin de proponer al ICFES una rutina de análisis de DIF en pruebas como el Examen de Estado. La elaboración del diseño de investigación para el presente trabajo considera un conjunto de factores fijos, variables independientes y variables de respuesta, que se enuncian a continuación.

Factores fijos

Uno de los propósitos centrales de la investigación sobre sesgo cultural en los exámenes ICFES es la generación de nuevo conocimiento que tenga ulterior impacto en la manera como se analiza e interpreta el DIF en los Exámenes de Estado; es por ello que se controlaron algunas variables en la simulación de los datos como la longitud de la prueba, el tamaño de muestra, magnitud de DIF, y tipo de DIF (uniforme y no uniforme), a fin de obtener datos en condiciones semejantes al contexto real de aplicación de las pruebas de Estado. Para determinar los valores de estas variables se tomó como criterio la información reportada por el ICFES para los Exámenes de Estado del período 2005-2006.

1. **Longitud de la prueba**: Se define como el número total de ítems que conforman la prueba. Rogers y Swaminathan (1993) apuntan que entre mayor sea la longitud de prueba, más confiable será el puntaje total obtenido en la misma, ya que el puntaje total se emplea como criterio de equiparación de los grupos, y en la medida que se tengan más ítems que den cuenta del atributo y se tenga un estimativo total de la habilidad del individuo, serán mejores las estimaciones de los parámetros en regresión

logística. Las longitudes de prueba que han sido manipuladas en diferentes estudios de simulación de DIF oscilan entre 20 y 80 preguntas (French & Maller, 2007). El tamaño de prueba para la investigación sobre sesgo cultural se definió de acuerdo con la longitud promedio de las pruebas que conforman el núcleo común de los Exámenes de Estado ICFES, el cual fue de 30 ítems.

2. **Tamaño de Muestra:** Alude al número total de individuos a los cuales se les ha aplicado la prueba. A diferencia de los dos estudios anteriores (Arias, 2008; Berrío, 2008) que emplearon un tamaño de 130000 individuos, definido a partir de la observación del número de inscritos en las pruebas que conforman el componente flexible interdisciplinar del Examen de Estado, para el estudio de regresión logística se empleó un tamaño de muestra de n= 65000 individuos, lo cual constituye un amplio tamaño de muestra que es manejable desde el punto de vista computacional; ya que la regresión logística, en comparación con otras técnicas de detección de DIF, tiene que ajustar uno o más modelos, lo cual incrementa el tiempo de ejecución del procedimiento y genera una mayor demanda en recursos informáticos. Así mismo, la elección de *n* = 65000 permite simular adecuadamente las condiciones de razón de tamaño que resultaron apropiadas en la identificación de DIF y el control del error tipo I de los estadísticos MH y diferencia de la dificultad.

3. **Magnitud de DIF**: Es la cantidad de DIF que reporta el ítem, la cual se expresa como el valor del área entre las curvas que presentan los ítems en el grupo de referencia y el grupo focal (Rogers & Swaminathan, 1993). Se empleó como medida para cuantificar la magnitud de DIF la fórmula de área no signada entre CCI para modelos de tres parámetros, propuesta por Raju (1988).

$$(1-c)\left|\frac{2(a_r - a_f)}{da_r a_f}\ln\left[1+e^{\frac{da_r a_f (br-bf)}{a_r - a_f}}\right]-(b_r - b_f)\right| \qquad (6)$$

En esta ecuación a_r y a_f representan el parámetro de discriminación para el grupo de referencia y focal; b_r y b_f representan el parámetro de dificultad para los grupos de referencia y focal; *c* simboliza el parámetro de pseudoazar y *d* es una constante igual a 1.7. En términos de área entre las funciones de respuesta al ítem, un ítem presenta DIF si el área entre CCI para los grupos focal y de referencia es diferente de cero.

La magnitud de DIF para la investigación sobre sesgo cultural se fijó en 0.4, que representa una magnitud de DIF moderada (Narayanan & Swaminathan, 1996). La determinación del nivel de magnitud de DIF se da a partir de los resultados obtenidos por Herrera (2005), Herrera y Gómez (2007), Arias (2008) y Berrío (2008), en donde condiciones experimentales bajo esta magnitud reflejan una adecuada identificación de DIF. Cromwell (2006) señala al respecto que "el área entre CCI de ≥ 0.40 se considera lo suficientemente grande para determinar DIF en orden a mantener los datos tan reales como sea posible para los análisis y la interpretación de los resultados" (p. 48).

4. **Tipo de DIF:** Hace alusión a la categoría de DIF que puede presentarse en la prueba (uniforme, no uniforme simétrico o asimétrico). Se simularon ítems con DIF uniforme y DIF no uniforme simétrico (diferencias en discriminación únicamente), siguiendo una proporción de 2 a 1; es decir, por cada dos ítems con DIF uniforme hay un ítem con DIF no uniforme. Narayanan y Swaminathan (1996) afirman que "aunque el DIF uniforme ocurre más que el DIF no uniforme en pruebas estandarizadas, se han identificado ítems con DIF no uniforme en datos reales" (p. 257). La inclusión de DIF no uniforme en la simulación de pruebas en el ámbito educativo constituye un primer paso en el análisis de los tipos de DIF presentes en las pruebas educativas aplicadas en el contexto colombiano; además resulta de gran interés analizar el comportamiento de RL en la detección de este tipo de DIF, ya que precisamente una de las bondades de la técnica de RL es la detección de DIF uniforme y no uniforme, y en la detección de este último, RL se ha mostrado superior con respecto a otros estadísticos como el MH (Swaminathan & Rogers, 1990).

Variables independientes

Las variables que se manipularon fueron razón de tamaños, porcentaje de DIF, Impacto y modelo de simulación. Algunos de los niveles de las variables independientes considerados en los dos estudios anteriores fueron eliminados para este estudio, en virtud de los hallazgos obtenidos por las autoras tanto para MH como en el caso del procedimiento diferencia de la dificultad:

1. **Razón de tamaños**: Se simboliza como $r = nr/nf$, y representa el número de individuos en el grupo de referencia por un miembro del grupo focal. Al analizar los

resultados de investigaciones previas sobre el funcionamiento de diferentes niveles de razones de tamaño para algunos procedimientos de detección de DIF (Herrera, 2005; Arias, 2008; Berrío, 2008), se seleccionaron tres niveles de razón de tamaño a analizar: *20:1*; *100:1* y *250:1*. La elección de estas razones de tamaño en el marco del proyecto de investigación sobre sesgo se dio en virtud de conservar en la simulación de los datos unas condiciones similares a las que se pueden presentar en el territorio nacional, en cuanto a la proporción de individuos de grupos culturales mayoritarios y minoritarios que presentan los Exámenes de Estado.

Teniendo en cuenta el tamaño de muestra (n = 65000), y de acuerdo con las razones de tamaño analizadas, la distribución de los individuos en los grupos focal y de referencia, se dio de la siguiente manera: 20:1 (n_r = 61905, n_f = 3095); 100: 1 (n_r = 64356, n_f = 644) y 250:1 (n_r = 64741, n_f = 259). En relación con el número mínimo de personas que deben conforman el grupo focal para obtener adecuadas estimaciones de los parámetros, González-Romá, Hernández & Gómez-Benito (2006) observaron en su investigación sobre potencia y error Tipo I de un procedimiento basado en el modelo de análisis de estructuras de covarianza para detectar DIF uniforme y no uniforme, que cuando el grupo focal era de n = 100, decrecía el poder para la detección de DIF de magnitud media (0.25), pero cuando el grupo focal era de n = 200 y grupos de referencia de tamaño 200 y 400, había un poder aceptable (≥70%) para detectar DIF uniforme y no uniforme de magnitud media (0.25). Teniendo presente lo anterior, la razón de tamaño más extrema (250:1) posee un n_f de 259, de manera que las estimaciones de los parámetros de los miembros del grupo focal sean apropiadas.

2. **Impacto**: Se define como la diferencia en las distribuciones de habilidad entre los grupos focal y de referencia (Jodoin & Huff, 2001). Para el presente estudio se tomaron dos niveles de esta variable: *Sin impacto*, es decir, una distribución normal de media 0 y desviación estándar 1 tanto para el grupo focal como el de referencia N(0,1); y con *impacto en la media*, es decir N(0, 1) para el grupo de referencia y N(-1, 1) para el grupo focal, en consonancia con los hallazgos obtenidos en los estudios de Arias (2008) y Berrío (2008), en donde la condición de impacto en la media incide en la potencia y el error tipo I de los procedimientos de Mantel-Haenszel y de diferencia de la dificultad en la identificación de DIF.

3. **Porcentaje de DIF**: Expresa el número de ítems que presentan funcionamiento diferencial en relación con el número total de ítems que conforman la prueba. Rogers y Swaminathan (1993) afirman que entre mayor sea el porcentaje de ítems con DIF en una prueba, el puntaje total de la prueba no constituirá una medida adecuada del nivel de atributo, debido a la contaminación del criterio del nivel de habilidad. French y Maller (2007) refieren en cuanto a los porcentajes de DIF que pueden encontrarse en las pruebas, que las proporciones de 5% y 10% de DIF son usuales en DIF relacionado con raza y género, mientras que el 20% de DIF es frecuente cuando el análisis de DIF atañe a oportunidades diferenciales en el aprendizaje de currículos educativos. Se analizaron para el presente estudio tres porcentajes de DIF: 0%, 10% y 20%. Para la condición de 10%, dos ítems presentaron DIF uniforme (3 y 10) y un ítem presentó DIF no uniforme (9); mientras que en la condición de 20% de DIF, cuatro ítems presentaron DIF uniforme (3, 10, 21 y 25) y dos ítems presentaron DIF no uniforme (9 y 24).

4. **Modelo de simulación**: El modelo de simulación de los datos puede generar ajuste o desajuste de éste a los datos, e incide con ello en la distribución del estadístico de prueba de la regresión logística (Rogers & Swaminathan, 1993). Teniendo en cuenta lo anterior, el modelo con el cual se generaron los datos se tomó como una variable de interés para la regresión logística en la detección de DIF. Se analizaron dos modelos: Modelo logístico de 1 parámetro (1PLM), y modelo logístico de 3 parámetros (3PLM).

Variables de respuesta

1. **Error tipo I**: Se definió como la tasa observada de falsos positivos obtenida en las condiciones de 0% de DIF, calculada sobre las 500 réplicas, empleando dos niveles de significación $\alpha = 0.05$ y $\alpha = 0.01$.

2. **Potencia:** Se definió como la tasa observada de detecciones correctas de los ítems con DIF en las condiciones que presentaron DIF (10% y 20%), utilizando valores de significación $\alpha = 0.05$ y $\alpha = 0.01$.

Cruzando completamente las variables independientes a ser manipuladas, se diseñó un estudio experimental con 36 condiciones (tres niveles de razón de tamaño, tres niveles de porcentaje de DIF, dos niveles de impacto, y dos tipos de modelo de simulación), consecuentes con los hallazgos previos de los estudios anteriores realizados en el marco del proyecto de investigación (Tabla 1).

Tabla 1. Diseño de las condiciones experimentales

Condición experimental	Razón de Tamaño	Impacto	Porcentaje de DIF	Modelo de simulación
1	20:1	Sin impacto	0%	1 parámetro
2	20:1	Sin impacto	0%	3 parámetros
3	20:1	Sin impacto	10%	1 parámetro
4	20:1	Sin impacto	10%	3 parámetros
5	20:1	Sin impacto	20%	1 parámetro
6	20:1	Sin impacto	20%	3 parámetros
7	20:1	Diferentes medias	0%	1 parámetro
8	20:1	Diferentes medias	0%	3 parámetros
9	20:1	Diferentes medias	10%	1 parámetro
10	20:1	Diferentes medias	10%	3 parámetros
11	20:1	Diferentes medias	20%	1 parámetro
12	20:1	Diferentes medias	20%	3 parámetros
13	100:1	Sin impacto	0%	1 parámetro
14	100:1	Sin impacto	0%	3 parámetros
15	100:1	Sin impacto	10%	1 parámetro
16	100:1	Sin impacto	10%	3 parámetros
17	100:1	Sin impacto	20%	1 parámetro
18	100:1	Sin impacto	20%	3 parámetros
19	100:1	Diferentes medias	0%	1 parámetro
20	100:1	Diferentes medias	0%	3 parámetros
21	100:1	Diferentes medias	10%	1 parámetro
22	100:1	Diferentes medias	10%	3 parámetros
23	100:1	Diferentes medias	20%	1 parámetro
24	100:1	Diferentes medias	20%	3 parámetros
25	250:1	Sin impacto	0%	1 parámetro
26	250:1	Sin impacto	0%	3 parámetros
27	250:1	Sin impacto	10%	1 parámetro
28	250:1	Sin impacto	10%	3 parámetros
29	250:1	Sin impacto	20%	1 parámetro
30	250:1	Sin impacto	20%	3 parámetros
31	250:1	Diferentes medias	0%	1 parámetro
32	250:1	Diferentes medias	0%	3 parámetros
33	250:1	Diferentes medias	10%	1 parámetro
34	250:1	Diferentes medias	10%	3 parámetros
35	250:1	Diferentes medias	20%	1 parámetro
36	250:1	Diferentes medias	20%	3 parámetros

Procedimiento

El presente estudio se realizó en tres fases: Generación de datos, aplicación del método de detección de DIF (regresión logística) y análisis de datos.

Generación de datos

Se simuló una prueba unidimensional de 30 ítems ajustados a modelos TRI tanto de 1 parámetro como de 3 parámetros con *c* = 0.15, mediante el programa BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002). Los parámetros de los ítems no DIF se simularon de acuerdo con los parámetros de los ítems extraídos del Examen de Estado aplicado en el segundo semestre de 2006. Para ello, se solicitó al ICFES la base de datos de estudiantes que presentaron la prueba en el II-2006, el string de respuestas y las claves de puntuación, para las pruebas del núcleo común. Posteriormente, y a fin de obtener los parámetros de los ítems, se realizó la calificación respectiva de éstos con las claves de puntuación, transformando el string de respuestas original en una cadena de unos y ceros.

La estimación de los parámetros de los ítems se realizó para cada uno de los ítems que componían las pruebas de lenguaje, matemáticas y sociales, con el programa BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002). Luego de la estimación de parámetros, se escogieron 30 ítems de las pruebas anteriores para estructurar la prueba simulada. Los 30 ítems que conforman esta prueba representan diferentes niveles de discriminación y dificultad (Tabla 2). Los parámetros de los ítems fueron los mismos empleados en los estudios de Arias (2008) y Berrío (2008).

Tabla 2. Parámetros de 30 ítems del Examen de Estado ICFES aplicado durante el segundo semestre de 2006

Ítem	*a*	*b*	Ítem	*a*	*b*
1	2,20559	2,07487	**16**	0,75536	-1,0611
2	0,46125	1,46794	**17**	1,16234	2,08233
3	0,39115	0,5167	**18**	1,42727	2,56616
4	0,76307	2,40671	**19**	0,80862	-0,20613
5	0,43264	-0,9214	**20**	0,50056	1,76429
6	0,33913	2,92167	**21**	1,22742	2,08076
7	1,02945	-0,38977	**22**	0,79738	0,03815
8	0,51012	1,60746	**23**	0,48182	2,08688
9	1,29943	-0,95524	**24**	0,63603	0,76112
10	1,40483	-0,99862	**25**	0,36933	2,30862
11	1,08021	0,71909	**26**	0,5267	0,44556
12	0,75375	0,26373	**27**	0,9977	0,32505
13	0,6093	1,98956	**28**	0,58404	0,42478
14	1,55649	3,08462	**29**	1,60704	-1,55029
15	3,01816	1,85731	**30**	0,36103	1,4191

Para simular el DIF uniforme y no uniforme se mantuvo como criterio para la magnitud de DIF un área entre CCI de 0.4, calculada con la fórmula de Raju (1988). De este modo, y teniendo en cuenta los parámetros de los ítems obtenidos de la matriz de respuestas proporcionada por el ICFES, se reemplazaron sus valores en la fórmula de área no signada de Raju y se obtuvieron los valores de los parámetros de los ítems para el grupo focal, de manera que el resultado del área entre las curvas cumpliera el criterio de 0.4.

El DIF se introdujo variando el parámetro de dificultad o discriminación en el grupo focal, según el tipo de DIF. Para los ítems con DIF no uniforme, el parámetro de discriminación se aumentó en el ítem 24, mediante la fórmula $a_F = a_R + [0.40 /(1-c)]$ y se disminuyó en el ítem 9, a partir de $a_F = a_R - [0.40 /(1-c)]$. Para los cuatro ítems con DIF uniforme (3, 10, 21 y 25) el parámetro de dificultad se aumentó en el grupo focal, mediante la expresión $b_F = b_R + [0.40 /(1-c)]$. En la condición de 10% de DIF se modificó el parámetro *a* para el ítem 9, y el parámetro *b* para los ítems 3 y 10. Para la condición de 20%, se modificó el parámetro *a* para los ítems 9 y 24, y el parámetro *b* para los ítems 3, 10, 21 y 25.

En la tabla 3 se pueden observar los parámetros de los ítems, el área entre las curvas y tipo de DIF de los ítems con DIF.

Tabla 3. Parámetros de los ítems con DIF

Ítem	a_r	b_r	a_f	b_f	*c*	Área entre CCI	Tipo DIF
3	0.39115	0.5167	0.39115	**0.999**	0.15	0.410	Uniforme
9	1.29943	-0.95524	**0.74231**	-0.95524	0.15	0.400	No uniforme
10	1.40483	-0.99862	1.40483	**-0.52**	0.15	0.407	Uniforme
21	1.22742	2.08076	1.22742	**2.56054**	0.15	0.408	Uniforme
24	0.63603	0.76112	**1.00591**	0.76112	0.15	0.401	No uniforme
25	0.36933	2.30862	0.36933	**2.79**	0.15	0.409	Uniforme

Convenciones: ar = Parámetro de discriminación para el grupo de referencia. af = Parámetro de discriminación para el grupo focal
br = Parámetro de dificultad para el grupo de referencia bf = Parámetro de dificultad para el grupo focal
c = Parámetro de pseudoazar, común para ambos grupos de evaluación

Luego de simular los parámetros de los ítems, se procedió a simular los parámetros de los individuos, para lo cual se generaron vectores de números aleatorios con distribución normal en los cuales la media cambiaba para el grupo focal según la

condición experimental, a fin de simular las diferencias de habilidad entre los grupos. Las matrices de datos fueron obtenidas a través de rutinas desarrolladas en Lenguaje R (Cervantes, 2007).

Con el objeto de simular las respuestas de los individuos en la mitad de las condiciones experimentales, se obtuvo una matriz de probabilidad de acierto en cada ítem para cada individuo siguiendo un modelo logístico de un parámetro, en el que se asume que el acierto de un ítem depende de la habilidad del sujeto y la dificultad del ítem. Este modelo se expresa como:

$$p_i(\theta) = \frac{e^{d(\theta - b_i)}}{1 + e^{d(\theta - b_i)}} \qquad (7)$$

Donde $p_i(\theta)$ es la probabilidad de que un examinado responda correctamente el ítem *i,* dado un nivel de atributo θ, e es la base de los logaritmos naturales, b_i es el parámetro de dificultad del ítem, y d es una constante de corrección equivalente a 1.7.

Las matrices de probabilidad de acierto para las demás condiciones experimentales se simularon siguiendo un modelo logístico de tres parámetros (Birnbaum, 1968; Herrera, Sánchez & Jiménez, 2001), el cual se expresa de la siguiente manera:

$$p_i(\theta) = c_i + (1 - c_i)\frac{e^{da_i(\theta - b_i)}}{1 + e^{da_i(\theta - b_i)}} \qquad (8)$$

Donde $p_i(\theta)$ es la probabilidad de que un examinado responda correctamente el ítem *i,* dado un nivel de atributo θ, a_i y c_i representan los parámetros de discriminación y pseudoazar, respectivamente. El parámetro c_i se fijó en 0.15. Los demás términos del modelo corresponden con los del modelo de un parámetro.

A partir de la obtención de las matrices de probabilidad de acertar en cada ítem ($P = p_i(\theta)$) para cada individuo con el modelo de uno y tres parámetros, se generaron las respuestas de los 130000 individuos para los 30 ítems que conformaron la prueba. Los valores de 1 (acierto en el ítem) y 0 (fallo en el ítem) se simularon siguiendo una distribución de Bernoulli con parámetro *P* antes definido. Cada una de las 120 condiciones experimentales del proyecto de investigación fue replicada 500 veces,

generando en total 60.000 matrices de respuestas, las cuales fueron obtenidas mediante el programa R versión 2.6 (Ihaka & Gentleman, 2007).

Una vez conocidos los resultados de los estudios de Arias (2008) y Berrío (2008) sobre el efecto de la razón de tamaños en la detección de DIF mediante MH y diferencia de la dificultad, respectivamente, y al analizar el papel de las demás variables manipuladas sobre la potencia y error tipo I de los estadísticos antes mencionados, se procedió a rediseñar las condiciones para el presente estudio de regresión logística, resultando en un total de 36 condiciones experimentales que se muestran en la tabla 1. Los datos se obtuvieron haciendo una selección aleatoria de 65000 individuos en cada una de las 36 condiciones seleccionadas para el estudio. El script de muestreo (Cervantes, 2008a) se diseñó con el programa ActivePerl 5.10.0 (Active State Software Inc., 2008) (Anexo 1). Con respecto al número de réplicas, se conservaron las 500 réplicas para cada una de las 36 condiciones, para un total de 18.000 matrices de respuestas a analizar.

Aplicación del procedimiento de detección de DIF

Para esta fase del estudio se generó un script para regresión logística (Cervantes, 2008b, ver Anexo 2), empleando la función *Logistic Regression Models* (LRM) del paquete Design, que se encuentra en el programa estadístico R v. 2.8.1 (R Development Core Team, 2009). La función *logistic regression model* permite ajustar modelos de regresión logística binaria empleando estimaciones de máxima verosimilitud. Bajo esta aproximación, se asume que los parámetros que acompañan a las variables independientes conservan una relación lineal. El LRM puede expresarse de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k x_k + \varepsilon \qquad (9)$$

Donde Y es la variable respuesta, X_1, X_2 son variables independientes o explicativas, β_i son los parámetros fijos no conocidos para cada una de las X_i, y ε es el error aleatorio.

Considerando lo anterior, la función LRM del paquete R permite ajustar los modelos de regresión logística, a partir de la especificación de los predictores lineales y una descripción de la variable respuesta. La función lrm en R se expresa como **`lrm`**

(resp_var~func_z), donde **resp_var** hace referencia a la variable dependiente (probabilidad de acierto al ítem), y **func_z** alude a la descripción de las variables predictoras que configuran el modelo a ajustar.

En relación con el procedimiento del presente estudio, inicialmente se creó la variable de equiparación, consistente en el puntaje total de la prueba, a fin de comparar los grupos focal y de referencia, y se realizó la regresión logística en dos fases: En la primera fase se ajustaron los tres modelos[4] con la función *lrm*. El modelo 1 contiene sólo el parámetro de habilidad, y se expresa como $P\langle y = 1|\theta\rangle = \beta_0 + \beta_1\theta$ .El modelo 2 contiene los parámetros de habilidad y grupo, y se denota como $P\langle y = 1|\theta\rangle = \beta_0 + \beta_1\theta + \beta_2 g$. El modelo 3 contiene los dos anteriores y la interacción entre ambos, como predictores de la probabilidad de acierto al ítem; este modelo se expresa como $P\langle y = 1|\theta\rangle = \beta_0 + \beta_1\theta + \beta_2 g + \beta_3\theta g$. La estimación de los parámetros en cada uno de los modelos ajustados se realizó por el método de máxima verosimilitud.

Posteriormente, se aplicó la regresión logística sobre los ítems utilizando la *prueba conjunta de DIF*, comparando los modelos 1 (no DIF) y el modelo 3 (modelo completo) mediante el estadístico G^2 que sigue una distribución χ^2_{2df} , poniendo a prueba la hipótesis $H_0 : \beta_2 = \beta_3 = 0$. Con esta primera aplicación se obtuvieron los ítems identificados con DIF con α = 0.01, es decir, los ítems para los cuales el valor *p* es menor a este nivel de significación; luego, se eliminaron los ítems que fueron identificados con DIF del cálculo de la habilidad para cada sujeto, y se efectuó un nuevo cálculo de la habilidad. Este procedimiento se conoce como purificación de la escala, puesto que busca eliminar el efecto de los ítems con DIF en la estimación de habilidad de los individuos.

---

[4] Para estos modelos, θ representa el nivel de habilidad o atributo del individuo, g es el grupo al cual pertenece el individuo, $\theta\, g$ es la interacción entre el nivel de habilidad y el grupo, β_0 representa el intercepto, y β_1, β_2 y β_3 representan los coeficientes para la habilidad, el grupo y la interacción grupo-habilidad, respectivamente.

La segunda fase del procedimiento tuvo lugar posterior a la purificación, y consistió en aplicar nuevamente regresión logística utilizando la prueba conjunta de DIF (comparación de modelo 1 con modelo 3) tomando en cuenta todos los ítems[5], y se reportó el estadístico G^2 que sigue una distribución χ^2_{2df}, en donde se pone a prueba la hipótesis $H_0 : \beta_2 = \beta_3 = 0$, y posteriormente se identificaron los ítems con DIF al nivel de significancia α = 0.05 y α = 0.01. El ítem se declara como DIF cuando el valor *p* es menor al nivel de significancia establecido. Dado que una de las desventajas de la prueba conjunta de DIF es que no permite establecer qué tipo de DIF presenta el ítem (Hidalgo et al., 2005), se empleó para esta fase la comparación de modelos jerárquicamente anidados, es decir, la comparación del modelo 1 con el 2 para determinar si hay DIF uniforme (efecto de la variable grupo); y la comparación del modelo 2 con el 3 para determinar si hay DIF no uniforme (efecto de la interacción grupo * habilidad), a partir del estadístico G^2 que sigue una distribución χ^2 con 1 grado de libertad [$\chi^2 U_{1df}$ y $\chi^2 NU_{1df}$], para DIF uniforme y DIF no uniforme, respectivamente. Por consiguiente, una vez el ítem se identificó con DIF mediante la prueba conjunta y los niveles de significancia establecidos, se procedió a examinar el tipo de DIF.

Análisis de datos

Los análisis de datos se efectuaron en tres etapas: Examen de la potencia y error tipo I, efecto de las variables manipuladas sobre la potencia y el error tipo I de la regresión logística, y análisis adicionales de los parámetros de los ítems y del error tipo I por tipo de DIF.

El análisis de la potencia en regresión logística se realizó mediante el *cálculo de tasas de detección* de los ítems con DIF en las condiciones de 10% y 20% de DIF (24 condiciones), es decir, se obtuvo la proporción de detección correcta de DIF para estas condiciones en sus 500 réplicas, con un nivel de significancia de 0.05 o 0.01. El valor

[5] Sólo se eliminan los ítems identificados con DIF para recalcular la habilidad, no se eliminan del procedimiento de regresión logística como tal

de 0.70 se utilizó como referente para evaluar los resultados de poder, en consonancia con las investigaciones previas realizadas en el marco del proyecto de investigación (Arias, 2008; Berrío, 2008), y en estudios previos de DIF (French & Maller, 2007). Un valor de poder ≥ 0.70 se considera alto, y valores < 0.70 se consideran bajos en la detección de DIF.

El análisis del error tipo I en regresión logística se calculó a través de la *proporción de ítems falsamente identificados con DIF* en las 12 condiciones de 0% de DIF, sobre las 500 réplicas. Se dice que un estadístico es robusto cuando la probabilidad empírica (Π) de cometer error Tipo I es aproximadamente igual al nivel de significación empleado en los análisis (Fidalgo, Mellenbergh & Muñiz, 2000). Bradley (1978) propuso un criterio liberal y otro conservador para determinar en qué medida un test estadístico es robusto. De acuerdo con Fidalgo et al., (2000), "una prueba satisface el criterio liberal de Bradley si $0.5\,\alpha \leq \Pi \leq 1.5\,\alpha$, y el criterio conservador si $0.9\alpha \leq \Pi \leq 1.1\alpha$" (p. 233). Teniendo en cuenta el criterio liberal de Bradley aplicado para el presente estudio, con un α = 0.05, tasas de falsos positivos entre 0.025 y 0.075; y con un α = 0.01, tasas de falsos positivos entre 0.005 y 0.015, no implican una inflación importante de error Tipo I. El criterio conservador por su parte, indica que con un α = 0.05, tasas de falsos positivos entre 0.045 y 0.055; y con un α = 0.01, tasas de falsos positivos entre 0.009 y 0.011, no implican una inflación importante de error Tipo I.

Una vez calculada la tasa de falsos positivos y de detecciones correctas, se procedió a obtener la tasa de detecciones correctas y falsos positivos para las 36 condiciones experimentales, a fin de obtener una panorámica general del comportamiento de las condiciones y un primer indicio sobre la posibilidad de efectos e interacciones entre las variables manipuladas en el estudio.

La evaluación de los efectos que sobre el error tipo I y la potencia de prueba presentaron la razón de tamaños, modelo de simulación, impacto y en el caso de la potencia, el porcentaje de ítems con DIF en la prueba, se hizo mediante descriptivos de las tasas promedio de falsos positivos y detecciones correctas de los ítems, teniendo en cuenta las variables independientes y sus interacciones, así como a través de gráficos que pueden resultar de interés para examinar la interacción de ciertas variables y su influencia en el error tipo I y la potencia de la RL, o bien el comportamiento de algunos

ítems DIF y no DIF. Además se realizaron análisis de varianza de múltiples vías; en donde la variable dependiente fue la tasa de detecciones correctas con un 10% y 20% de ítems DIF en la prueba, ó de falsos positivos en la condición de 0% de DIF, y en las condiciones de 10% y 20% de DIF, seleccionando los ítems no DIF para tales porcentajes de DIF, según se examinara potencia o error tipo I respectivamente.

Finalmente, se realizaron dos análisis adicionales. El primero de ellos consistió en una aproximación exploratoria al comportamiento de los parámetros de dificultad y discriminación de los ítems sobre la potencia y el error tipo I de la RL, mediante correlaciones y gráficos descriptivos; y el segundo consistió en la realización de un análisis de varianza de múltiples vías, a fin de evaluar la influencia de las variables independientes sobre el error tipo I de la RL por tipo de DIF, es decir, la identificación errónea de la clasificación de un ítem simulado como DIF (i.e. se detecta un ítem como uniforme cuando fue simulado como no uniforme, y viceversa) en condiciones de 10% y 20% de DIF.

Las tres etapas se realizaron con los niveles de significación α = 0.05 y α = 0.01, y para las tres pruebas de detección de DIF (conjunta, DIF uniforme y DIF no uniforme) Las tasas de detecciones correctas y falsos positivos, así como los análisis de varianza, se llevaron a cabo en el programa SPSS v15.

Capítulo 5

RESULTADOS

Los resultados se presentarán en tres secciones principales, divididas de la siguiente manera:

1. **Error tipo I para la RL***,* examinado en dos aspectos: (a) *Tasa de detecciones incorrectas de los ítems no DIF en las condiciones de 0% de DIF*; y (b) *Tasa de detecciones incorrectas de los ítems no DIF en las condiciones de 10% y 20% de DIF*, empleando la prueba G^2 en la detección conjunta de DIF (χ^2_{2df}), DIF uniforme ($\chi^2 U_{1df}$) y DIF no uniforme ($\chi^2 NU_{1df}$), con un nivel de significación α = 0.05 y α = 0.01.

2. **Potencia de la RL** *a partir de la tasa de detecciones correctas de los ítems DIF en las condiciones de 10% y 20% de DIF*, empleando la prueba G^2 en la detección conjunta de DIF (χ^2_{2df}), DIF uniforme ($\chi^2 U_{1df}$) y DIF no uniforme ($\chi^2 NU_{1df}$), con un nivel de significación α = 0.05 y α = 0.01.

3. **Hallazgos adicionales**, que incluyen dos aspectos: (a) Exploración preliminar de los parámetros de dificultad y discriminación en la potencia y el error tipo I de la RL; y (b) error tipo I de RL en términos *de la tasa de detecciones incorrectas del tipo de ítem con DIF en las condiciones de 10% y 20% de DIF*, empleando la prueba G^2 en la detección conjunta de DIF (χ^2_{2df}), DIF uniforme ($\chi^2 U_{1df}$) y DIF no uniforme ($\chi^2 NU_{1df}$), con un nivel de significación α = 0.05 y α = 0.01.

Error Tipo I

Tasa global de falsos positivos por condición experimental

Se realizó un análisis descriptivo de la tasa promedio de error tipo I de los ítems no DIF en 0%, 10% y 20% de DIF para cada una de las 36 condiciones experimentales (Tabla 4), con el propósito de delinear en términos generales el comportamiento del estadístico en estas condiciones.

Tabla 4. Tasas promedio de error tipo I en 0%, 10% y 20% de DIF, por condición experimental (α = 0.05)

	Error tipo I 0% DIF			Error tipo I 10% DIF			Error tipo I 20% DIF		
Condición [a]	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}
R 20:1, SI, 1 P	0.049	0.024	0.024						
R 20:1, SI, 3 P	0.049	0.025	0.024						
R 20:1, SI, 1 P				0.048	0.021	0.026			
R 20:1, SI, 3 P				0.045	0.020	0.025			
R 20:1, SI, 1 P							0.049	0.025	0.024
R 20:1, SI, 3 P							0.045	0.019	0.025
R 20:1, DM, 1 P	0.969	0.883	0.931						
R 20:1, DM, 3 P	0.757	0.582	0.510						
R 20:1, DM, 1 P				0.973	0.899	0.934			
R 20:1, DM 3 P				0.771	0.572	0.479			
R 20:1, DM, 1 P							0.969	0.885	0.922
R 20:1, DM, 3 P							0.775	0.585	0.488
R 100:1, SI, 1 P	0.051	0.023	0.026						
R 100:1, SI, 3 P	0.055	0.026	0.027						
R 100:1, SI, 1 P				0.050	0.024	0.025			
R 100:1, SI, 3 P				0.049	0.021	0.026			
R 100:1, SI, 1 P							0.051	0.022	0.029
R 100:1, SI, 3 P							0.046	0.023	0.024
R 100:1, DM, 1 P	0.489	0.284	0.424						
R 100:1, DM, 3 P	0.330	0.241	0.155						
R 100:1, DM, 1 P				0.489	0.263	0.429			
R 100:1, DM, 3 P				0.267	0.164	0.145			
R 100:1, DM, 1 P							0.516	0.283	0.457
R 100:1, DM, 3 P							0.264	0.168	0.141
R 250:1, SI, 1 P	0.052	0.025	0.028						
R 250:1, SI, 3 P	0.052	0.026	0.026						
R 250:1, SI, 1 P				0.052	0.023	0.027			
R 250:1, SI, 3 P				0.050	0.025	0.024			
R 250:1, SI, 1 P							0.054	0.025	0.028
R 250:1, SI, 3 P							0.051	0.026	0.025
R 250:1, DM, 1 P	0.227	0.099	0.180						
R 250:1, DM, 3 P	0.200	0.144	0.079						
R 250:1, DM, 1 P				0.242	0.103	0.191			
R 250:1, DM, 3 P				0.156	0.099	0.073			
R 250:1, DM, 1 P							0.253	0.104	0.205
R 250:1, DM, 3 P							0.161	0.103	0.075

[a] CONVENCIONES: R 20:1 = Razón de Tamaño 20:1; R 100:1 = Razón de Tamaño 100:1; R 250:1= Razón de Tamaño 250:1; SI = Sin impacto; DM = Diferencia en la media; 1P = 1 parámetro; 3 P = 3 parámetros. Las tasas promedio resaltadas en negrita no cumplen el criterio liberal de Bradley.

En términos del **error tipo I** para las **condiciones de 0%** de DIF, cuando no hay impacto en las tres razones de tamaño y se contemplan los dos modelos de simulación (1P y 3P), las tasas de error se mantienen en el criterio liberal propuesto por Bradley (1978). La tasa de error tipo I más alta se presenta cuando se posee *razón de tamaño 20:1, diferencias en la media, y modelo de 1 parámetro* (tasas entre 0.88 y 0.96); y *modelo de 3 parámetros* (tasas de falsos positivos entre 0.51 y 0.75), para las tres pruebas de detección de DIF.

Para las razones 100:1 y 250:1, con diferencias en la media y tomando los dos modelos de simulación de los datos, se observan tasas de error tipo I > 0.075. Cuando se presentan diferencias en la media y modelo de 1 parámetro, independientemente de la razón de tamaño, la mayor tasa de error tipo I se encuentra en la prueba conjunta de DIF seguida por la prueba de DIF no uniforme; no obstante cabe anotar que a medida que aumenta la razón de tamaños, la tasa de error tipo I disminuye. En las restantes condiciones experimentales, la prueba de DIF uniforme ocupa el segundo lugar, luego de la prueba conjunta de DIF.

El comportamiento del **error tipo I** en los **niveles de 10% y 20% de DIF** para las 36 condiciones experimentales**,** refleja una tendencia similar a la observada en el análisis del error tipo I para 0% DIF. En situaciones donde no hay impacto y se contempla tanto el modelo de 1 parámetro como el de 3 parámetros, para las tres razones de tamaño consideradas, hay un adecuado control de la tasa de falsos positivos. La *razón de tamaño 20:1, con diferencias en la media, y modelo de 1 parámetro* (tasas entre 0.89 y 0.97 para 10% de DIF; y 0.88 y 0.96 para 20% de DIF); *y modelo de 3 parámetros* (0.48 y 0.77 para 10% de DIF; y 0.49 y 0.78 para 20% de DIF), presenta la mayor tasa de error tipo I en las condiciones de 10% y 20% de DIF, para las tres pruebas de detección de DIF. Cuando se tiene una razón 20:1, diferencia en la media y modelo de 1 parámetro, en cada uno de los niveles de DIF, se aprecia una mayor tasa de falsos positivos en la condición de 10% de DIF; mientras que la presencia de una razón 20:1, diferencia en la media y modelo de 3 parámetros reporta una mayor tasa de error tipo I en la condición de 20% de DIF.

En las razones 100:1 y 250:1, con diferencias en la media y tomando los dos modelos de simulación, se observan así mismo tasas de falsos positivos > 0.075. Al

tomar sólo el modelo de 1 parámetro para estas dos razones de tamaño y presencia de diferencias en la media, la mayor tasa de falsos positivos se aprecia en la condición de 20% de DIF; mientras que si se considera el modelo de 3 parámetros, conservando las mismas razones de tamaño y las diferencias en las medias, la condición de 0% es aquella que presenta mayor cantidad de falsos positivos, seguida de la condición de 10% en la razón de 100:1. En la razón 250:1, la condición de 20% de DIF sucede a la condición de 0% de DIF.

Cuando se presentan diferencias en la media y modelo de 1 parámetro, independientemente de la razón de tamaño, la mayor tasa de error tipo I se encuentra en la prueba conjunta de DIF seguida por la prueba de DIF no uniforme; y a medida que aumenta la razón de tamaños, la tasa de error tipo I disminuye. En las restantes condiciones experimentales, la prueba de DIF uniforme ocupa el segundo lugar, luego de la prueba conjunta de DIF.

Al examinar la tasa promedio de error tipo I de ítems no DIF en los tres niveles de DIF para cada una de las 36 condiciones experimentales con un nivel de significancia α = 0.01 (Anexo 3), se observa en términos generales la misma tendencia que con α = 0.05 en cuanto al error tipo I.

Teniendo presente los anteriores resultados, uno de los factores determinantes en la presentación de altas tasas de error tipo I es el impacto, en donde las diferencias en la distribución de habilidad de los individuos son sensibles a ser identificadas como presencia de funcionamiento diferencial. Para lograr un adecuado control del error tipo I en la RL, de acuerdo con los resultados observados, se requeriría que los individuos de los grupos evaluados fuesen iguales en la distribución del nivel de habilidad.

En relación con la razón de tamaños, razones pequeñas, es decir, aquellas en las que hay un mayor número de individuos en el grupo focal, presentan mayores tasas de error tipo I, y a medida que disminuye la cantidad de individuos en el grupo focal en relación con el de referencia, las tasas de falsos positivos disminuyen. No obstante, la tasa de falsos positivos que se reporta en las razones extremas de 100:1 y 250:1 excede el criterio liberal de Bradley.

El tipo de modelo de simulación de los datos, en donde el modelo de 1 parámetro señala la habilidad del individuo y la dificultad del ítem como predictores de la

probabilidad de acierto a un ítem, incrementa la ocurrencia de falsos positivos en la RL, ya que el modelo no toma en consideración el parámetro de discriminación, que puede presentar valores altos en ciertos ítems y por ende ser detectados como DIF no uniforme, esto se aprecia particularmente en aquellas situaciones donde el modelo de 1 parámetro estuvo acompañado por diferencias en la media. En las condiciones donde se simularon los datos con el modelo de 3 parámetros, se aprecia una disminución de las tasas de falsos positivos con respecto a las condiciones de 1 parámetro, aunque la presencia de diferencias en la media en conjunción con un modelo de 3 parámetros muestra una tasa de falsos positivos que excede el criterio de Bradley.

En relación con el porcentaje de DIF que contiene la prueba, éste parece ser un factor importante que interactúa junto con otros niveles de las variables manipuladas e incide en la presentación de altas tasas de error tipo I. Tal es el caso de las razones 100:1 y 250:1, cuando se presentan diferencias en la media y modelo de 1 parámetro, reportan una alta tasa de falsos positivos en la condición de 20%. Esto sugiere que en situaciones donde los grupos difieren en la distribución de habilidad, la composición numérica de los grupos es marcadamente desigual, con amplia desventaja para el grupo focal; se emplee un modelo de 1 parámetro y una prueba con alto porcentaje de ítems con DIF, la probabilidad que los ítems no DIF sean detectados como DIF es mayor.

Una posible explicación a este comportamiento puede darse en dos vías: Por una parte, los parámetros de dificultad y de discriminación de los ítems no DIF, cuyos valores superiores o inferiores conducen a que sean detectados por la RL como ítems DIF; y una segunda posibilidad es que a medida que aumenta el porcentaje de ítems con DIF en una prueba, aumenta la tasa de falsos positivos. También habría que establecer hasta qué punto el aumento del porcentaje de DIF en la prueba y la consecuente inclusión de más ítems con DIF no uniforme hace que la tasa de falsos positivos aumente, particularmente en las situaciones de modelo 1 parámetro, donde el parámetro de discriminación no se considera en este modelo.

Por otra parte, la condición de modelo de 3 parámetros, en las razones 100:1 y 250:1 y diferencias en la media, presentó mayores inflaciones del error tipo I en la condición de 0% de DIF; esto sugiere que cuando se simula con un modelo de 3

parámetros, donde la discriminación y el factor de pseudoazar forman parte del modelo, en situaciones donde hay un menor número de individuos en el grupo focal, y los grupos evaluados difieren en la distribución de la habilidad, resulta un incremento de falsos positivos en una prueba que no posee ítems DIF. Nuevamente, es preciso examinar en qué medida los valores de los parámetros de dificultad y de discriminación contribuyen a la presentación de estas tasas de falsos positivos.

Factores que afectan el error Tipo I de la RL (condición 0% de DIF)

Para establecer cuáles factores influyen en el error tipo I de la RL en las condiciones de 0% de DIF, se realizó un ANOVA de tres vías, en donde se tomó como variable dependiente la tasa de falsos positivos en los ítems no DIF en la condición de 0% de DIF en las tres pruebas de detección de DIF ($\chi^2{}_{2df}$, $\chi^2 U_{1df}$ y $\chi^2 NU_{1df}$), y como variables independientes la razón de tamaños, el impacto y el modelo de simulación.

De acuerdo con el análisis de varianza de la tabla 5, se observa un efecto significativo de las variables impacto, razón y modelo en el error tipo I para la prueba conjunta de DIF, DIF uniforme y DIF no uniforme, con α = 0.05. Teniendo en cuenta los efectos principales, la variable más importante en las tres clasificaciones de DIF es el *impacto*, seguida de la razón de tamaños y el modelo. Si se examinan comparativamente las pruebas de detección de DIF en términos de la importancia que cada una de las variables manipuladas ejerce en el error tipo I para 0% DIF, se puede apreciar que el impacto presenta una mayor incidencia en la prueba conjunta de DIF; y la razón y el modelo en la prueba de DIF no uniforme.

Con respecto a las interacciones de dos vías, aquella que ejerce una mayor influencia sobre el error tipo I en las condiciones de 0% de DIF para los tres tipos de DIF es razón * impacto, seguida de impacto * modelo. En la prueba de DIF uniforme y la prueba de DIF no uniforme se observa un efecto importante de la interacción razón* modelo de simulación. Esta interacción razón * modelo ocupa el segundo lugar de importancia en la prueba de DIF uniforme. La interacción impacto * modelo se ubica en el segundo lugar de importancia después de razón * impacto en la prueba conjunta de DIF y en la prueba de DIF no uniforme. Similares tendencias reportadas con respecto a

los efectos principales y las interacciones aplican para el análisis de varianza al nivel de significancia α = 0.01 (Anexo 4).

Tabla 5. Valor F y significación de los efectos sobre el error tipo I de la RL en la condición de 0% de DIF (α = 0.05)

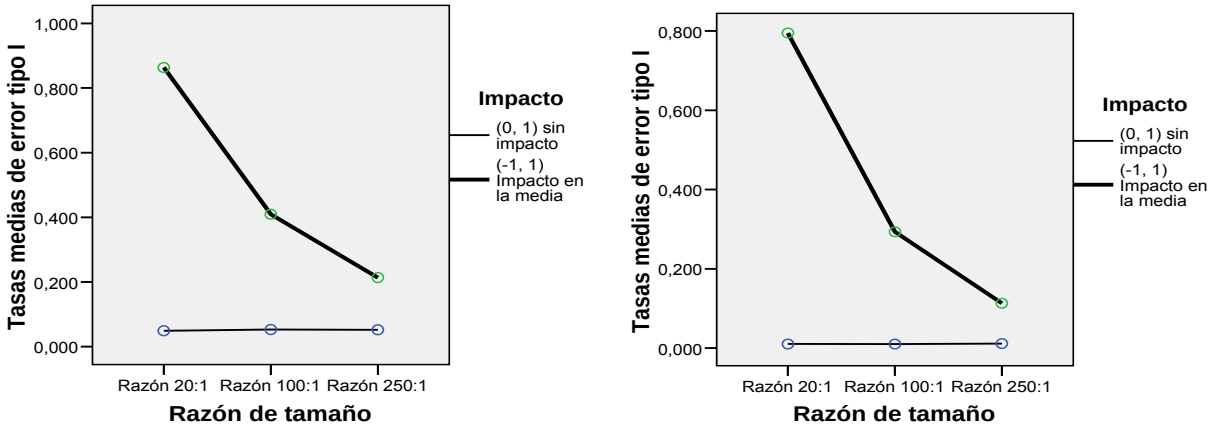
	χ^2_{2df}		χ^2U_{1df}		χ^2NU_{1df}	
Factor	F	Significación	F	Significación	F	Significación
Razón	96.901	.000	89.781	.000	101.480	.000
Impacto	521.967	.000	318.431	.000	410.911	.000
Modelo	11.439	.001	6.273	.013	57.283	.000
Razón * Impacto	98.962	.000	90.328	.000	103.261	.000
Razón * Modelo	1.942	.145	7.163	.001	6.896	.001
Impacto * Modelo	11.837	.001	6.680	.010	56.840	.000

En relación con los efectos de las interacciones en la prueba conjunta de DIF con un α = 0.05 (Figura 3a), cabe anotar que con respecto a la razón de tamaño * impacto, la tasa de falsos positivos es mayor en la razón de 20:1 y diferencias en la media (86.3%). La condición de 100:1 refleja una disminución importante de la tasa de error tipo I (41%), y cuando se tiene una razón extrema de 250:1 y diferencias en la media, la tasa de falsos positivos se ubica en un 21.4%. En la prueba de DIF no uniforme, donde esta interacción resultó altamente significativa, en la razón 20:1 con diferencias en la media, la tasa de falsos positivos alcanza un 79%, y al llegar a la razón 100:1, la tasa de falsos positivos se ubica en 28.9%. En la razón 250:1 y diferencias en la media, la tasa de error tipo I es del 12.1%.

Si se examina con un α = 0.01 la interacción entre razón e impacto para la prueba conjunta de DIF (Figura 3b) se observa una mayor reducción en la tasa de error tipo I al pasar de una razón 20:1 a una razón 100:1, cuando existen diferencias en la media (79.5 a 29.3%). Las tasas de error tipo I en general son menores a las obtenidas con α = 0.01, no obstante, exceden el criterio liberal de Bradley.

Teniendo en cuenta lo anterior, razones de tamaño pequeñas, en conjunción con diferencias en la distribución de la media de habilidad, son condiciones que conducen a altas tasas del error tipo I, y a medida que aumenta la proporción de individuos en el grupo focal en relación con los individuos del grupo de referencia, se observa una disminución en la tasa de falsos positivos de la RL, aunque no lo suficientemente

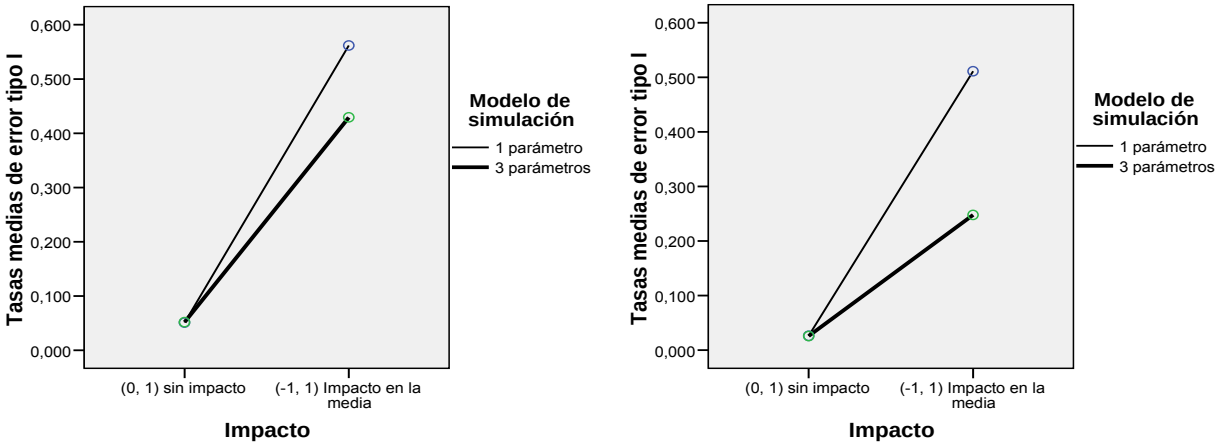
grande como para lograr un control adecuado del error tipo I, ya que excede los criterios liberales de Bradley para α = 0.05; esto es, la tasa de falsos positivos es mayor a 0.075.



3a. 3b.

Figura 3. Tasas medias de error tipo I cuando hay interacción entre razón de tamaño e impacto, para la prueba conjunta de DIF. (3a) α = 0.05. (3b) α = 0.01.

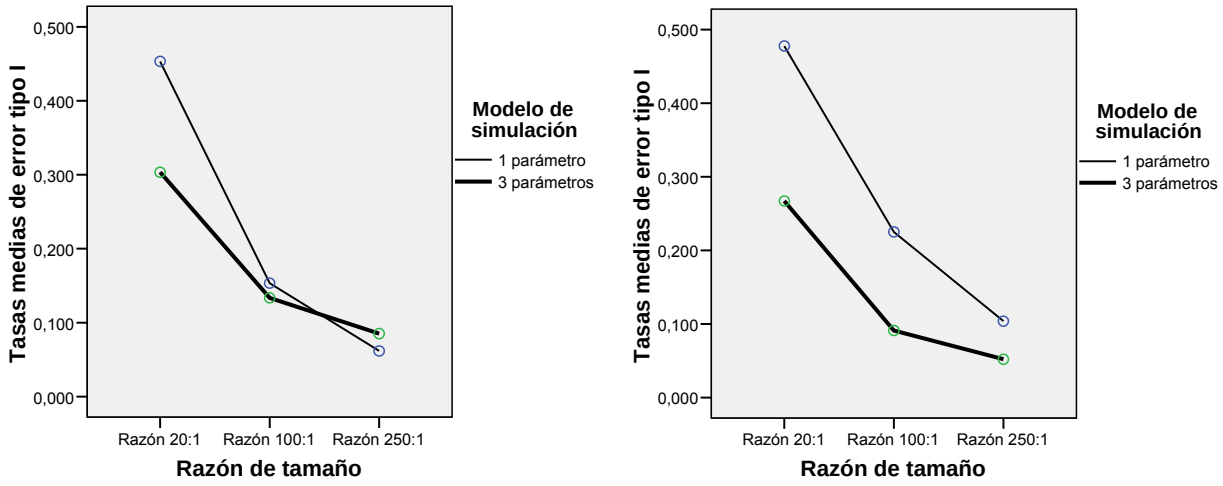
En la interacción de los niveles de impacto con el modelo de simulación, se puede apreciar que en condiciones donde hay diferencias en la media y modelo de1 parámetro, se presenta una tasa mayor de falsos positivos (56.2%) para la prueba conjunta de DIF (Figura 4a). No obstante, para esta prueba de detección de DIF, el modelo de 3 parámetros reporta una alta tasa de error tipo I. En la prueba de DIF no uniforme, donde la interacción impacto*modelo es significativa e importante, el modelo de 1 parámetro junto con diferencias en la media reporta mayores tasas de error tipo I, al compararlo con el modelo de 3 parámetros (Figura 4b).



4a. 4b.

Figura 4. Tasas medias de error tipo I con α = 0.05, cuando hay interacción entre impacto y modelo de simulación. (4a) prueba conjunta de DIF. (4b) DIF no uniforme.

La interacción razón * modelo resultó significativa para las pruebas de DIF uniforme y no uniforme con α = 0.05, aunque el comportamiento de esta relación difiere según la prueba de detección de DIF. En la prueba de DIF uniforme (Figura 5a), la razón 20:1 y modelo de 1 parámetro presentó la mayor tasa de error tipo I (45.3%). Al aumentar la razón de tamaño a 100:1, se reduce la tasa de error tipo I en más del 50% para ambos modelos, sin embargo, cuando se tiene la razón de tamaño más extrema y modelo de 3 parámetros, la tasa de falsos positivos se incrementa en comparación con el modelo de 1 parámetro; el modelo de 1 parámetro presenta un adecuado control de error tipo I al considerar la razón más extrema. Para la prueba de DIF no uniforme (Figura 5b), el modelo de 1 parámetro presenta una mayor tasa de falsos positivos que el modelo de 3 parámetros, en todos los niveles de razón de tamaño. A medida que aumenta la razón de tamaños disminuye el error tipo I en ambos modelos, siendo la razón de 100:1 aquella en la que se observan los mayores decrementos en la tasa de falsos positivos. La razón de 250:1 con modelo de 3 parámetros presenta un adecuado control de la tasa de error tipo I.



5a. 5b.

Figura 5. Tasas medias de error tipo I con α = 0.05, cuando hay interacción entre razón y modelo de simulación. (5a) DIF uniforme. (5b) DIF no uniforme.

El análisis de la tasa de detecciones incorrectas para la condición de 0% DIF empleando las tres pruebas de detección de DIF (Tabla 6), y teniendo en cuenta las variables manipuladas, señala un incremento considerable en las tasas de error tipo I

para la mayoría de condiciones, al examinar estas tasas en función de los criterios propuestos por Bradley (1978) de intervalos de confianza para el error tipo I. Entre los hallazgos más relevantes se destacan:

1. Las condiciones de no impacto mantuvieron las tasas de falsos positivos dentro de los límites propuestos por Bradley (1978) (2 - 5% para α = 0.05; y 0 -1% para α = 0.01). No se observan diferencias entre estas condiciones para las tres pruebas de detección de DIF ni en el caso del aumento de la razón de tamaño.

2. La condición de diferencias en la media presenta altas tasas de falsos positivos, particularmente en la razón de 20:1, donde hay una mayor cantidad de individuos en el grupo focal (entre 72 y 86% para α = 0.05, y entre 64 y 79% α = 0.01). Estos porcentajes corresponden a la prueba conjunta de DIF. A partir de la razón 100:1 se comienzan a observar descensos del 50 y 60% en la tasa de falsos positivos, con un leve incremento en el error tipo I en la prueba de DIF no uniforme en comparación con la prueba de DIF uniforme para las razones de tamaño 100:1 y 250:1. No obstante, todas las condiciones de diferencias en la media en conjunción con la razón de tamaño a un nivel α = 0.05 y α = 0.01 presentan tasas de falsos positivos que exceden el criterio liberal de Bradley para estos niveles de significancia.

3. En relación con el modelo de simulación, las tasas de falsos positivos más altas se presentan en el modelo de 1 parámetro para la razón de 20: 1 (tasas entre 45 y 51% con α = 0.05; y entre 40 y 47%, con α = 0.01). Las tasas más altas de error tipo I en el modelo de 3 parámetros también se encuentran en la razón 20:1 (27- 40% con α = 0.05, y 20 - 33% con α = 0.01). El modelo de 1 parámetro reporta una tasa mayor de falsos positivos que el modelo de 3 parámetros, a medida que incrementa la razón de tamaños; sin embargo, el aumento de la razón de tamaño se vincula con una disminución del error tipo I en ambos modelos. Cabe resaltar que en el modelo de 1 parámetro, los mayores porcentajes de error tipo I se encuentran para la prueba conjunta de DIF y la prueba de DIF no uniforme; mientras que en el modelo de 3 parámetros, la prueba conjunta de DIF es la que reporta mayor error tipo I, seguida de la prueba de DIF uniforme y finalmente de la prueba de DIF no uniforme; esta tendencia se observa en los tres niveles de razón de tamaño y en los dos niveles de significación.

Tabla 6. Tasa promedio de falsos positivos de la RL en condición de 0% de DIF, con interacción entre las variables (α = 0.05 y α = 0.01)

	Razón de tamaño									
	Razón 20:1			Razón 100:1			Razón 250:1			
α = 0.05	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	TOTAL GENERAL
(0, 1) sin impacto	0.049	0.024	0.024	0.053	0.025	0.027	0.052	0.026	0.027	0.034
(-1, 1) Impacto en la media	0.863[a]	0.732[a]	0.720[a]	0.410[a]	0.262[a]	0.289[a]	0.214[a]	0.121[a]	0.129[a]	0.416[a]
1 parámetro	0.509[a]	0.453[a]	0.477[a]	0.270[a]	0.153[a]	0.225[a]	0.140[a]	0.062	0.104[a]	0.266[a]
3 parámetros	0.403[a]	0.303[a]	0.267[a]	0.192[a]	0.134[a]	0.091[a]	0.126[a]	0.085[a]	0.052	0.184[a]
TOTAL GENERAL	0.456[a]	0.378[a]	0.372[a]	0.231[a]	0.144[a]	0.158[a]	0.133[a]	0.074[a]	0.078[a]	0.225[a]
α = 0.01										
(0, 1) sin impacto	0.010	0.004	0.005	0.010	0.004	0.005	0.011	0.004	0.005	0.006
(-1, 1) Impacto en la media	0.795[b]	0.641[b]	0.639[b]	0.293[b]	0.178[b]	0.180[b]	0.113[b]	0.063[b]	0.051[b]	0.328[b]
1 parámetro	0.474[b]	0.402[b]	0.440[b]	0.187[b]	0.102[b]	0.141[b]	0.062[b]	0.026[b]	0.037[b]	0.208[b]
3 parámetros	0.331[b]	0.243[b]	0.204[b]	0.116[b]	0.079[b]	0.045[b]	0.06 [2b]	0.042[b]	0.019[b]	0.135[b]
TOTAL GENERAL	0.403[b]	0.323[b]	0.322[b]	0.152[b]	0.091[b]	0.093[b]	0.062[b]	0.034[b]	0.028[b]	0.167[b]

[a]. La tasa de error tipo I estimada no cumple el criterio liberal de Bradley, con α = 0.05 (FP > 0.075).

[b]. La tasa de error tipo I estimada no cumple el criterio liberal de Bradley, con α = 0.01 (FP > 0.015).

Factores que afectan el error Tipo I de la RL (condiciones 10% y 20% de DIF)

Para establecer cuáles factores influyen en el error tipo I de la RL en las condiciones de 10% y 20% de DIF, se realizó un ANOVA de cuatro vías, en donde se tomó como variable dependiente la tasa de falsos positivos en los ítems no DIF en las condiciones de 10% y 20% de DIF para las tres pruebas de detección de DIF ($\chi^2{}_{2df}$, $\chi^2 U_{1df}$ y $\chi^2 NU_{1df}$), y como variables independientes la razón de tamaños, el impacto, el modelo de simulación, y el porcentaje de DIF.

Tabla 7. Valor F y significación de los efectos sobre el error tipo I de la RL en la condición de 10% y 20% de DIF (α = 0.05)

	$\chi^2{}_{2df}$		$\chi^2 U_{1df}$		$\chi^2 NU_{1df}$	
Factor	F	Significación	F	Significación	F	Significación
Razón	197.515	.000	194.197	.000	160.765	.000
Impacto	959.112	.000	562.359	.000	695.007	.000
Porcentaje de DIF	.063	.802	.036	.849	.053	.818
Modelo	39.724	.000	26.186	.000	118.657	.000
Razón * Impacto	202.899	.000	197.602	.000	161.992	.000
Razón * Porcentaje de DIF	.016	.984	.014	.986	.031	.969
Impacto * Porcentaje de DIF	.050	.823	.015	.901	.057	.812
Razón * Modelo	2.484	.084	11.130	.000	11.775	.000
Impacto * Modelo	36.947	.000	25.361	.000	115.685	.000
Porcentaje de DIF * Modelo	.042	.838	.005	.943	.032	.859

De acuerdo con el análisis de varianza de la tabla 7, se observa un efecto significativo de las variables impacto, razón y modelo en el error tipo I para la prueba conjunta de DIF, DIF uniforme y DIF no uniforme, con α = 0.05. Al examinar los efectos principales, la variable más importante en las tres clasificaciones de DIF es el *impacto*, seguida de la razón de tamaños y el modelo de simulación. *El porcentaje de DIF no resultó significativo para ninguna de las tres pruebas de detección de DIF*. Si se examinan comparativamente las pruebas de detección de DIF en términos de la importancia que cada una de las variables manipuladas ejerce en el error tipo I para las condiciones de 10% y 20% de DIF, se puede apreciar que el impacto y la razón presentan una mayor incidencia en la prueba conjunta de DIF; y el modelo de simulación en la prueba de DIF no uniforme.

Con respecto a las interacciones de dos vías, aquella que ejerce una mayor influencia sobre el error tipo I en las condiciones de 10% y 20% de DIF para los tres tipos de DIF es razón * impacto, seguida de impacto * modelo, en las tres pruebas de

detección de DIF; es decir, razones pequeñas acompañadas de diferencias en la media incrementan la tasa de falsos positivos, y el incremento de la razón de tamaño se acompaña de reducciones en la tasa de error tipo I. En relación con el impacto y el modelo, la presencia de diferencias en la media y modelo de 1 parámetro genera mayores tasas de falsos positivos en las condiciones de 10% y 20% de DIF.

La interacción razón* modelo es significativa sólo para la prueba de DIF uniforme y la prueba de DIF no uniforme; y la interacción impacto * modelo es más importante para la prueba de DIF no uniforme en comparación con las otras pruebas de detección de DIF.

Como se observa, el análisis de varianza de error tipo I para las condiciones de 10% y 20% de DIF, reporta resultados similares al ANOVA de error tipo I para la condición de 0% de DIF, en cuanto a efectos principales e interacciones se refiere. El ANOVA para condiciones de 10% y 20% de DIF con α = 0.01 refleja la misma tendencia (Anexo 5). Sólo cabe anotar que para el ANOVA con una significación α = 0.01, la interacción razón * modelo es significativa para las tres pruebas de detección de DIF, a diferencia de los análisis anteriores de error tipo I, en donde esta interacción resultó significativa únicamente para la prueba de DIF uniforme y la prueba de DIF no uniforme.

El análisis de la tasa de detecciones incorrectas para la condición de 10% y 20% de DIF empleando las tres pruebas de detección de DIF (Tabla 8), y teniendo en cuenta las condiciones experimentales, señala un incremento considerable en las tasas de error tipo I para la mayoría de condiciones, al examinar estas tasas en función de los criterios propuestos por Bradley (1978) de intervalos de confianza para el error tipo I. Las tasas de error tipo I y las tendencias para los niveles de impacto, razones de tamaño y modelo de simulación muestran comportamientos semejantes a las tasas observadas en las condiciones de 0% de DIF, es decir, las tasas de error tipo I son más altas en la razón 20:1, con diferencias en la media y modelo de 1 parámetro. El incremento en la razón de tamaños se acompaña de una disminución en la tasa de falsos positivos, aunque las tasas de error tipo I en las razones más extremas exceden el criterio liberal de Bradley. En relación con el porcentaje de DIF, no se presentan diferencias entre la condición de 10% y 20% de DIF al comparar por razones de tamaño y pruebas de detección de DIF, lo que es concordante con lo observado en el análisis de varianza, donde el porcentaje de DIF no resultó significativo en α = 0.05 y α = 0.01.

Tabla 8. Tasa promedio de falsos positivos de la RL en condición de 10% y 20% de DIF, con interacción entre las variables (α = 0.05 y α = 0.01)

	Razón de tamaño									
	Razón 20:1			Razón 100:1			Razón 250:1			
α = 0.05	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	TOTAL GENERAL
(0, 1) sin impacto	0.047	0.021	0.025	0.049	0.023	0.026	0.052	0.025	0.026	0.033
(-1, 1) Impacto en la media	0.872[a]	0.735[a]	0.706[a]	0.384[a]	0.219[a]	0.293[a]	0.203[a]	0.102[a]	0.136[a]	0.406[a]
1 parámetro	0.510[a]	0.458[a]	0.477[a]	0.276[a]	0.148[a]	0.234[a]	0.150[a]	0.064	0.113[a]	0.270[a]
3 parámetros	0.409[a]	0.299[a]	0.254[a]	0.157[a]	0.094[a]	0.084[a]	0.104[a]	0.063	0.049	0.168[a]
10 % DIF	0.459[a]	0.378[a]	0.366[a]	0.214[a]	0.118[a]	0.156[a]	0.125[a]	0.063	0.079[a]	0.218[a]
20 % DIF	0.459[a]	0.378[a]	0.365[a]	0.219[a]	0.124[a]	0.163[a]	0.130[a]	0.064	0.083[a]	0.221[a]
TOTAL GENERAL	0.459[a]	0.378[a]	0.366[a]	0.217[a]	0.121[a]	0.159[a]	0.127[a]	0.064	0.081[a]	0.219[a]
α = 0.01										
(0, 1) sin impacto	0.009	0.003	0.005	0.010	0.003	0.005	0.010	0.004	0.005	0.006
(-1, 1) Impacto en la media	0.804[b]	0.647[b]	0.622[b]	0.266[b]	0.140[b]	0.189[b]	0.099[b]	0.046[b]	0.055[b]	0.319[b]
1 parámetro	0.476[b]	0.408[b]	0.439[b]	0.194[b]	0.099[b]	0.151[b]	0.068[b]	0.027[b]	0.042[b]	0.212[b]
3 parámetros	0.337[b]	0.243[b]	0.188[b]	0.081[b]	0.044[b]	0.042[b]	0.042[b]	0.023[b]	0.018[b]	0.113[b]
10 % DIF	0.407[b]	0.323[b]	0.317[b]	0.135[b]	0.070[b]	0.095[b]	0.054[b]	0.025[b]	0.030[b]	0.162[b]
20 % DIF	0.406[b]	0.328[b]	0.310[b]	0.141[b]	0.074[b]	0.099[b]	0.056[b]	0.025[b]	0.031[b]	0.163[b]
TOTAL GENERAL	0.407[b]	0.325[b]	0.314[b]	0.138[b]	0.072[b]	0.097[b]	0.055[b]	0.025[b]	0.030[b]	0.162[b]

[a]. La tasa de error tipo I estimada no cumple el criterio liberal de Bradley, con α = 0.05 (FP > 0.075).

[b]. La tasa de error tipo I estimada no cumple el criterio conservador de Bradley, con α = 0.01 (FP > 0.015).

Potencia

Tasa global de detecciones correctas por condición experimental

Se realizó un análisis descriptivo de la tasa de detecciones correctas de los ítems DIF en 10% y 20% de DIF para cada una de las 36 condiciones experimentales (Tabla 9), con el propósito de delinear en términos generales el comportamiento del estadístico en estas condiciones.

Al examinar la **potencia de las condiciones con 10% de DIF** se aprecia que cuando se da una razón de tamaño 20:1, sin impacto y modelo de 3 parámetros, la tasa de detecciones correctas es del 100% en las tres pruebas de detección de DIF. Las razones 100:1 y 250:1, sin impacto y modelo de 3 parámetros presentan detecciones mayores a 0.70. La presencia de diferencias en la media y modelo de 1 parámetro, con independencia de las razones de tamaño, refleja una tasa de detecciones correctas mayor que 0.90. Cuando aumenta la razón de tamaño a 100:1 y se tiene modelo de 1 y 3 parámetros, se aprecia una leve disminución de la potencia de la RL, y los valores resultantes conservan una adecuada potencia. La tasa más baja de potencia para 10% de DIF (<0.70) se encuentra en la razón 250:1, con diferencia en la media y modelo de 3 parámetros. Así mismo, se presentan tasas bajas de potencia cuando los grupos a evaluar poseen la misma distribución de habilidad y modelo de 1 parámetro, para la prueba conjunta de DIF en las tres razones de tamaño.

En la prueba conjunta de DIF, se observa una potencia de 99.7% en las razones 20:1 y 100:1, con diferencia en la media y modelo de 1 parámetro. Asimismo se observan tasas de detecciones mayores a 0.70 en la razón de 20:1 y 250:1, con diferencia en la media y modelo de 3 parámetros; y en la razón de 100:1 con y sin impacto, y modelo de 3 parámetros.

Para la prueba de DIF uniforme, las tasas de detecciones correctas más altas se presentan cuando no hay impacto y el modelo es de 1 parámetro, con independencia de la razón de tamaño (rango de 0.99 a 1). Cabe anotar que se observa una potencia mayor que 0.90 cuando hay diferencias en la media y modelo de 1 parámetro, siendo la razón de 100:1 aquella que presenta un 100% de detecciones correctas.

En la prueba de DIF no uniforme, se observan detecciones correctas del 100% cuando la razón es 20:1, el modelo es de 3 parámetros, y hay o no impacto. En la razón de 100:1, donde hay o no impacto, y el modelo es de 3 parámetros, se observan asimismo detecciones entre 0.968 y 0.978. La menor potencia para DIF

no uniforme se presenta cuando la razón es 250:1, no hay impacto y el modelo es de 1 parámetro.

Tabla 9. Tasas promedio de detecciones correctas en 10% y 20% de DIF, por condición experimental (α = 0.05).

	Potencia DIF 10%			Potencia DIF 20%		
Condición [a]	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}
R 20:1, SI, 1 P						
R 20:1, SI, 3 P						
R 20:1, SI, 1 P	0.687	1	0.046			
R 20:1, SI, 3 P	1	1	1			
R 20:1, SI, 1 P				0.686	1	0.033
R 20:1, SI, 3 P				0.992	0.968	0.999
R 20:1, DM, 1 P						
R 20:1, DM, 3 P						
R 20:1, DM, 1 P	0.997	0.994	0.98			
R 20:1, DM 3 P	0.955	0.93	1			
R 20:1, DM, 1 P				1	1	0.974
R 20:1, DM, 3 P				0.903	0.680	0.634
R 100:1, SI, 1 P						
R 100:1, SI, 3 P						
R 100:1, SI, 1 P	0.690	1	0.054			
R 100:1, SI, 3 P	0.921	0.873	0.978			
R 100:1, SI, 1 P				0.652	0.954	0.022
R 100:1, SI, 3 P				0.697	0.552	0.736
R 100:1, DM, 1 P						
R 100:1, DM, 3 P						
R 100:1, DM, 1 P	0.997	1	0.9			
R 100:1, DM, 3 P	0.807	0.693	0.968			
R 100:1, DM, 1 P				0.811	0.778	0.759
R 100:1, DM, 3 P				0.613	0.366	0.495
R 250:1, SI, 1 P						
R 250:1, SI, 3 P						
R 250:1, SI, 1 P	0.685	0.991	0.042			
R 250:1, SI, 3 P	0.739	0.68	0.664			
R 250:1, SI, 1 P				0.534	0.761	0.029
R 250:1, SI, 3 P				0.469	0.375	0.443
R 250:1, DM, 1 P						
R 250:1, DM, 3 P						
R 250:1, DM, 1 P	0.900	0.974	0.636			
R 250:1, DM, 3 P	0.677	0.594	0.596			
R 250:1, DM, 1 P				0.562	0.579	0.38
R 250:1, DM, 3 P				0.453	0.326	0.341

[a] CONVENCIONES: R 20:1 = Razón de Tamaño 20:1; R 100:1 = Razón de Tamaño 100:1; R 250:1= Razón de Tamaño 250:1; SI = Sin impacto; DM = Diferencia en la media; 1P = 1 parámetro; 3 P = 3 parámetros. Las tasas promedio resaltadas en negrita corresponden a potencia > 0.70.

Con respecto a la **potencia de las condiciones con 20% de DIF** se aprecia que en general las tasas de detecciones correctas son inferiores a las condiciones que poseen 10% de DIF. Las potencias más altas se observan en la razón 20:1, con diferencias en la media y modelo de 1 parámetro (100% para la prueba conjunta de DIF y la prueba de DIF uniforme; y de 97.4% en la prueba de DIF no uniforme). Cuando se posee una razón 20:1, no hay impacto y el modelo es de 3 parámetros; y en la razón 100:1, con diferencia en la media y modelo de 1 parámetro, se observan tasas de detección mayores a 0.70 en las tres pruebas de detección de DIF.

Para la prueba de DIF uniforme, tasas de detecciones correctas mayores que 0.90 se observan en la razón 20:1, cuando hay o no impacto, y el modelo es de 1 parámetro; y en la razón 100:1 cuando no hay impacto y el modelo es de 1 parámetro. En la razón 100:1, con diferencia en la media, y en la razón 250:1 cuando no hay impacto, y en ambas razones de tamaño el modelo es de 1 parámetro, las tasas de detección son ligeramente mayores a 0.70. Las tasas de detección más bajas (rango entre 33% y 45%) tanto para la prueba de DIF uniforme como para la prueba conjunta de DIF se observan en la razón 250:1, con diferencias en la media y modelo de 3 parámetros.

En la prueba de DIF no uniforme, tasas mayores a 0.90 se observan en la razón de 20:1 cuando no hay impacto y el modelo es de 3 parámetros, seguida por la razón 20:1 con diferencias en la media y modelo de 1 parámetro. En la razón 100:1, sin impacto y modelo de 3 parámetros; y en la razón 100:1 con diferencias en la media y modelo de 1 parámetro, las tasas de detección oscilan entre el 73% y el 76%. Cabe anotar que para las razones 20:1 y 100:1, sin impacto y modelo de 3 parámetros, las tasas de detección para este tipo de DIF fueron más altas en comparación con la prueba conjunta de DIF y la prueba de DIF uniforme. La peor potencia para este tipo de DIF (tasa de 2.2%) se presentó cuando se emplea una razón 100:1, no hay impacto y el modelo es de 1 parámetro.

El análisis de la tasa de detecciones correctas en 10% y 20% de DIF con α = 0.01 (Anexo 6) revela un patrón similar al obtenido con α = 0.05. Como elementos divergentes con α = 0.05 se puede mencionar que en la condición de

10% de DIF cuando se tiene una razón de 100:1, los grupos difieren en habilidad y el modelo es de 1 parámetro, sólo la prueba de DIF no uniforme reporta una tasa de detecciones correctas menores a 0.70. En la razón de 250:1, con modelo de 3 parámetros y sin impacto, la tasa de falsos positivos para la prueba conjunta de DIF es menor a 0.70. En la condición de 20% de DIF, a partir de la razón 100:1, con diferencias en la media y modelo de 1 parámetro, la potencia de la RL es menor al 70%.

Una revisión de los resultados de potencia para 10% y 20% de DIF sugiere que las detecciones correctas de ítems DIF mediante el empleo de RL se obtendrían en situaciones donde la diferencia en proporción numérica de los grupos focal y de referencia sea pequeña, las distribuciones de habilidad de los grupos sean iguales y se consideren la habilidad, la dificultad y la discriminación modelo de 3 parámetros reporta altas tasas de detección correctas, la conjugación de diferencias en las distribuciones de habilidad de los grupos y modelo de 1 parámetro reporta una mayor potencia de la RL en la detección de DIF.

Si se observan las tasas de detecciones correctas por razón de tamaño, se aprecia una adecuada potencia en la medida que las razones de tamaño sean pequeñas; al aumentar la razón de tamaño, la potencia de la RL disminuye. Esta tendencia es similar para las tres pruebas de detección de DIF.

Es preciso señalar los efectos diferenciales de los factores que contribuyen a la potencia para cada una de las pruebas de detección de DIF. Para la prueba conjunta, las diferencias en la media y el modelo de 1 parámetro se asocian con una potencia cercana al 100% en la detección de DIF. Los resultados extraídos en la prueba de DIF uniforme sugieren que la detección de un ítem como uniforme se encuentra dada por el modelo de 1 parámetro y las diferencias en la media; mientras que en la prueba de DIF no uniforme, la potencia estaría más influenciada por el modelo de 3 parámetros y las diferencias en la media.

En relación con el porcentaje de ítems DIF en una prueba en la potencia de la RL, los incrementos en éste se acompañan de una disminución en la potencia, ya que en la medida que haya mayor número de ítems DIF, el puntaje total no constituye un criterio confiable como medida de magnitud de atributo.

Factores que afectan la potencia de la RL (condición 10% y 20% de DIF)

Para establecer cuáles factores influyen en la potencia de la RL en las condiciones de 10% y 20% de DIF, se realizó un ANOVA de cuatro vías, en donde se tomó como variable dependiente la tasa de detecciones correctas de los ítems DIF en las condiciones de 10% y 20% de DIF en las tres pruebas de detección de DIF (χ^2_{2df} , $\chi^2 U_{1df}$ y $\chi^2 NU_{1df}$), y como variables independientes la razón de tamaños, el impacto, el modelo de simulación y el porcentaje de DIF.

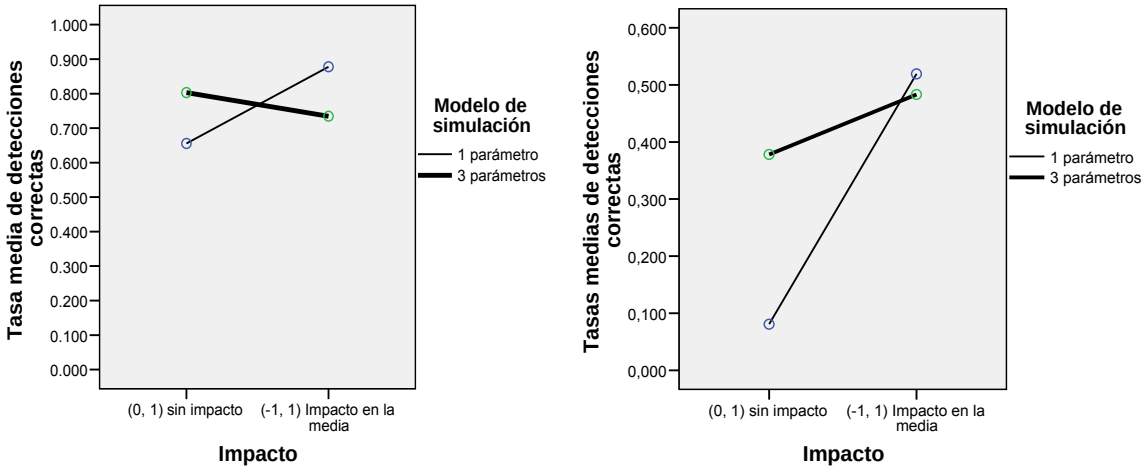
De acuerdo con el análisis de varianza de la tabla 10, se observa un efecto principal significativo de la razón de tamaño en la potencia de RL, para la prueba conjunta de DIF, DIF uniforme y DIF no uniforme, con α = 0.05. El porcentaje de DIF es significativo para la prueba de DIF uniforme y la prueba conjunta de DIF; mientras que el impacto y el modelo de simulación sólo fueron significativos para la prueba de DIF no uniforme.

En términos de la incidencia de las variables según prueba de detección de DIF, la razón y el porcentaje de DIF en su orden son los factores más importantes para la prueba conjunta de DIF y la prueba de DIF uniforme; para esta última prueba, el valor F en ambas variables es ligeramente mayor al obtenido en la prueba conjunta de DIF. Para la prueba de DIF no uniforme el impacto, la razón de tamaños y el modelo de simulación, son, en su orden, las variables consideradas fundamentales en la detección de ítems DIF

Tabla 10. Valor F y significación de los efectos sobre la potencia de la RL en la condición de 10% y 20% de DIF (α = 0.05)

Factor	χ^2_{2df}		$\chi^2 U_{1df}$		$\chi^2 NU_{1df}$	
	F	Significación	F	Significación	F	Significación
Razón	5.528	.006	6.498	.002	11.647	.000
Impacto	1.296	.258	.652	.422	20.576	.000
Porcentaje de DIF	4.304	.041	5.419	.022	2.141	.147
Modelo	.001	.975	.067	.797	4.756	.032
Razón * Impacto	.126	.882	.094	.910	2.318	.105
Razón * Porcentaje de DIF	.995	.374	.756	.473	.013	.987
Impacto * Porcentaje de DIF	.136	.713	.312	.578	.836	.363
Razón * Modelo	.823	.443	.622	.539	.074	.929
Impacto * Modelo	4.631	.034	2.576	.112	7.738	.007
Porcentaje de DIF * Modelo	.104	.748	.359	.551	.004	.951

Con respecto a las interacciones de dos vías, sólo resultó significativa la interacción impacto * modelo para la prueba de DIF no uniforme y la prueba conjunta de DIF, respectivamente.



6a. 6b.

Figura 6. Tasas medias de detecciones correctas con α = 0.05, cuando hay interacción entre impacto y modelo de simulación. (6a) Prueba conjunta de DIF. (6b) Prueba de DIF no uniforme.

En la prueba conjunta de DIF (Figura 6a), las diferencias en la media en conjunción con el modelo de 1 parámetro generan la mayor tasa de detecciones correctas de ítems DIF (87.8%). Cabe resaltar que cuando no hay impacto y el modelo es de 3 parámetros, la RL presenta una potencia cercana al 80%. En la prueba de DIF no uniforme, por su parte, cuando se asume que no hay diferencias entre los grupos y se trabaja con un modelo de 1 o 3 parámetros, la potencia de la RL oscila entre 8 y 38%, respectivamente. Cuando los grupos difieren en la media, la tasa de detecciones correctas aumenta en los dos modelos de simulación, con un ligero incremento al tener un modelo de 1 parámetro. Sin embargo, las tasas que se obtienen cuando hay diferencias en la media no son lo suficientemente altas como para indicar que la regresión logística tenga una adecuada potencia en la detección de este tipo de DIF.

Al realizar el análisis de varianza a cuatro vías, con α = 0.01 (Anexo 7) se encontró, a semejanza del análisis anterior, efectos principales significativos para

razón de tamaños en las tres pruebas de detección de DIF y efectos del impacto, porcentaje de DIF y modelo de simulación, para una o más pruebas de detección de DIF, siguiendo el mismo patrón de α = 0.05. A diferencia del ANOVA con α = 0.05, la interacción impacto * modelo sólo resultó significativa para la prueba de DIF no uniforme (F= 4.065, p= 0.047).

Teniendo en cuenta que la razón de tamaños fue el único factor que resultó significativo para la detección correcta de ítems DIF en las tres pruebas de detección de DIF, un análisis de la tasa global de detecciones correctas por razón de tamaño permite extraer hallazgos relevantes (Tabla 11):

Tabla 11. Tasa promedio de detecciones correctas de la RL en las condiciones de 10% y 20% de DIF, según razón de tamaño (α = 0.05 y α = 0.01)

	Razón de Tamaño									
	Razón 20:1			Razón 100:1			Razón 250:1			
α	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	Total
0.05	**0.900** [a]	**0.935** [a]	0.692	**0.747** [a]	**0.739** [a]	0.577	0.586	0.610	0.360	0.683
0.01	**0.888** [a]	**0.910** [a]	0.657	0.690	0.660	0.472	0.450	0.533	0.215	0.608
Total	**0.894** [a]	**0.923** [a]	0.675	**0.719** [a]	**0.700** [a]	0.525	0.518	0.572	0.288	0.646

a Detecciones resaltadas en negrita representan potencia > 0.70

1. La prueba de DIF uniforme es aquella que reporta mayor potencia, seguida de la prueba conjunta de DIF y la prueba de DIF no uniforme, en la razón 20:1. La razón 100:1 reporta una mayor potencia para la prueba conjunta de DIF, seguida de la prueba de DIF uniforme y la prueba de DIF no uniforme, respectivamente. Al comparar las tasas de detecciones correctas para χ^2_{2df} y χ^2U_{1df} se observa que no existen marcadas diferencias entre estas dos tasas, en los niveles de razón de tamaño. No obstante, la diferencia entre las tasas de χ^2_{2df} y χ^2NU_{1df} es mayor, mostrando que las detecciones correctas con la prueba de DIF no uniforme son inferiores al 70% en todas las condiciones de razón de tamaño.

2. La prueba conjunta de DIF y la prueba de DIF uniforme alcanzan tasas iguales o mayores a 70% en las razones de 20:1, para α = 0.05 y α = 0.01. En la

razón 100:1, la prueba conjunta de DIF y la prueba de DIF uniforme presentan un comportamiento similar, sólo en α = 0.05.

3. Las mayores tasas de detecciones de tasas correctas se observan en la razón 20:1 (tasas entre 0.69 y 0.94 para α = 0.05; y entre 0.66 y 0.91 para α = 0.01). A medida que aumenta la razón de tamaños decrece la tasa de detecciones correctas.

4. Al examinar la potencia en términos de los niveles de significación, se observa que no existen mayores diferencias entre $\chi^2{}_{2df}$ y $\chi^2 U_{1df}$, mientras que al interior de la prueba $\chi^2 NU_{1df}$, la diferencia entre α = 0.05 y α = 0.01 es mayor, principalmente cuando el tamaño del grupo focal es cada vez menor (razones 100:1 y 250:1).

Un análisis general de la tasa media de detecciones correctas de DIF teniendo en cuenta las cuatro variables consideradas en el ANOVA permite detallar algunas tendencias de interés (Tabla 12):

1. La condición de diferencia en la media presenta mayores tasas de detección de ítems DIF que la condición de no impacto, en los tres niveles de razón de tamaño. En la razón 20:1 se observan tasas mayores al 90% en las tres pruebas de detección de DIF; sin embargo, la potencia disminuye a medida que la razón de tamaño aumenta; para la razón 100:1 aún se observan tasas mayores a 0.70 para la prueba conjunta de DIF y la prueba de DIF no uniforme. La razón 250:1 presenta tasas de detección iguales o menores que 0.60 en los niveles de impacto.

Un examen de la incidencia del impacto sobre la detección de ítems DIF tomando en cuenta la prueba de DIF, muestra para la prueba conjunta una mayor detección en razón de 20:1 y diferencias en la media. En el caso de prueba de DIF uniforme se observa una potencia cercana al 100% en la razón de tamaño más pequeña y asumiendo que los grupos son iguales en la distribución de habilidad. La diferencia más notoria entre los niveles de impacto se observa para la prueba de DIF no uniforme, donde tasas mayores que 0.70 se aprecian sólo cuando los grupos difieren en la media y se tienen razones de tamaño 20:1 y 100:1.

2. En relación con el porcentaje de DIF, la tasa de detecciones correctas es mayor en la condición de 10% de DIF, en los tres niveles de razón de tamaño. En la condición de 10% de DIF, con un α = 0.05, una potencia mayor a 0.70 se

observa para las tres pruebas de detección de DIF en las razones 20:1 y 100:1; cuando la razón es de 250:1, sólo alcanzan este criterio la prueba conjunta de DIF y la prueba de DIF uniforme. Con un 20% de ítems DIF en la prueba sólo se aprecian tasas mayores a 0.80 en la razón 20:1, para la prueba conjunta de DIF y la prueba de DIF uniforme. Para la prueba de DIF no uniforme, sólo se alcanza una potencia > 0.70 en la razón 20:1, con un 10% de ítems DIF en la prueba.

3. Respecto al modelo de simulación, el modelo de 3 parámetros reporta mayores tasas de detecciones correctas que el modelo de 1 parámetro en todos los niveles de razón de tamaño. Las tasas de detecciones correctas más altas se presentan en el modelo de 3 parámetros para la razón de 20: 1 (tasas entre 88% y 96% con α = 0.05; y entre 85% y 94%, con α = 0.01), para las tres pruebas de detección de DIF. En la razón de 20:1 y 100:1, el modelo de 1 parámetro reporta tasas mayores a 0.70 para la prueba conjunta de DIF y la prueba de DIF uniforme; mientras que con una razón 100:1 y modelo de 3 parámetros, se observa una potencia mayor que 0.70 para la prueba conjunta de DIF y la prueba de DIF no uniforme. Los resultados obtenidos con el modelo de simulación, por tanto, sugieren un efecto diferencial del modelo de 1 parámetro para la detección de DIF uniforme, y del modelo de 3 parámetros en la detección del DIF no uniforme. Para la razón 250:1 sólo la prueba de DIF uniforme y modelo de 1 parámetro alcanza una potencia mayor que 0.70.

Tabla 12. Tasa promedio de detecciones correctas de la RL, con interacción entre las variables (α = 0.05 y α = 0.01)

	Razón de tamaño									
	Razón 20:1			Razón 100:1			Razón 250:1			
α = 0.05	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	TOTAL GENERAL
(0, 1) sin impacto	0.840	0.989	0.518	0.718	0.814	0.425	0.572	0.657	0.275	0.645
(-1, 1) Impacto en la media	0.960	0.881	0.866	0.775	0.663	0.729	0.601	0.563	0.446	0.720
10% DIF	0.910	0.981	0.757	0.854	0.892	0.725	0.750	0.810	0.485	0.796
20% DIF	0.895	0.912	0.660	0.693	0.662	0.503	0.505	0.510	0.298	0.626
1 parámetro	0.843	0.999	0.507	0.769	0.911	0.419	0.629	0.774	0.249	0.678
3 parámetros	0.957	0.871	0.878	0.725	0.567	0.735	0.543	0.446	0.471	0.688
TOTAL GENERAL	0.901	0.939	0.698	0.756	0.752	0.589	0.600	0.627	0.371	0.692
α = 0.01										
(0, 1) sin impacto	0.828	0.971	0.498	0.695	0.736	0.352	0.410	0.565	0.153	0.579
(-1, 1) Impacto en la media	0.948	0.948	0.816	0.685	0.584	0.592	0.489	0.501	0.276	0.649
10% DIF	0.905	0.980	0.736	0.810	0.841	0.637	0.661	0.756	0.306	0.737
20% DIF	0.880	0.875	0.618	0.631	0.569	0.390	0.344	0.422	0.169	0.544
1 parámetro	0.835	0.999	0.467	0.760	0.847	0.288	0.472	0.690	0.109	0.607
3 parámetros	0.941	0.821	0.847	0.620	0.473	0.656	0.427	0.376	0.320	0.609
TOTAL GENERAL	0.890	0932	0.664	0.700	0.675	0.486	0.467	0.552	0.222	0.621

a Detecciones resaltadas en negrita representan potencia > 0.70.

Hallazgos adicionales

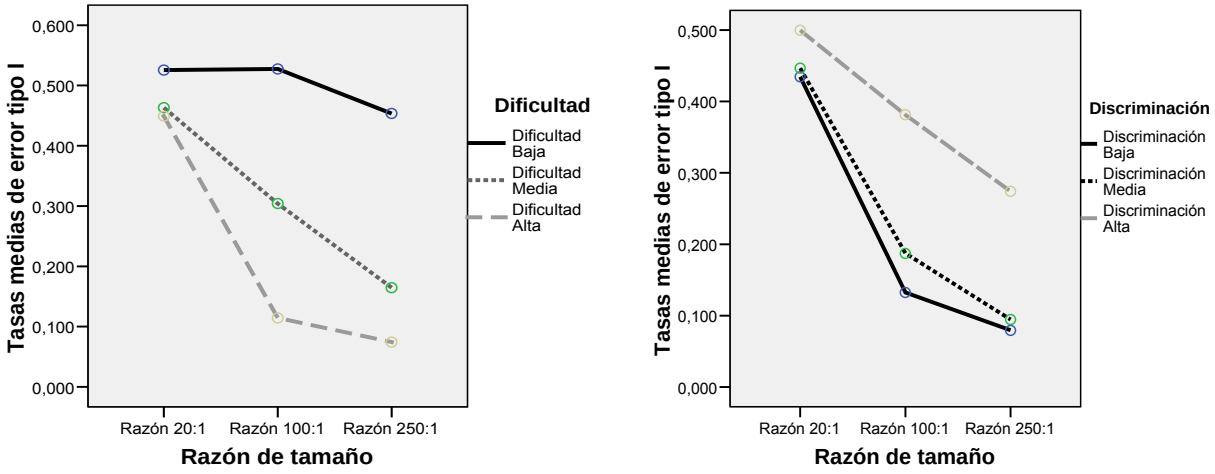
Análisis de la dificultad y la discriminación

Una exploración preliminar del comportamiento de los parámetros de dificultad y discriminación en relación con las variables manipuladas en el estudio, permitirá obtener una primera aproximación a los efectos que pueden tener los parámetros de los ítems sobre la potencia y el error tipo I de la RL. Un posible efecto de los parámetros en la RL ya ha sido reportado en estudios preliminares (Narayanan & Swaminathan, 1996; Herrera, 2005). Para el presente análisis, los niveles de dificultad y discriminación para los ítems se clasificaron teniendo presente los niveles propuestos en Herrera (2005). La dificultad se clasifica como baja si $b \leq -1.5$, media si $-1.5 \leq b \leq 1.5$, y alta si $b \geq 1.5$. La discriminación se clasifica como baja si $a < 0.5$, media si $0.5 \leq a \leq 1$, y alta si $a > 1$.

Error tipo I (0% DIF)

El examen de la tasa de falsos positivos de la prueba conjunta de DIF en relación con la razón de tamaños y los parámetros de dificultad y discriminación revela que *la razón de 20:1, baja dificultad y alta discriminación* reporta altas tasas de error tipo I, cercanas al 50% (Figura 7a y 7b). A medida que aumenta la razón de tamaño a 100:1 aumenta levemente la tasa de error tipo I en ítems con baja dificultad. La razón 100:1 en niveles de dificultad media y alta, y en los tres niveles de discriminación (Figura 7b) comienza a reportar una disminución importante del error tipo I. Esta tendencia se observa también en la prueba de DIF uniforme y no uniforme.

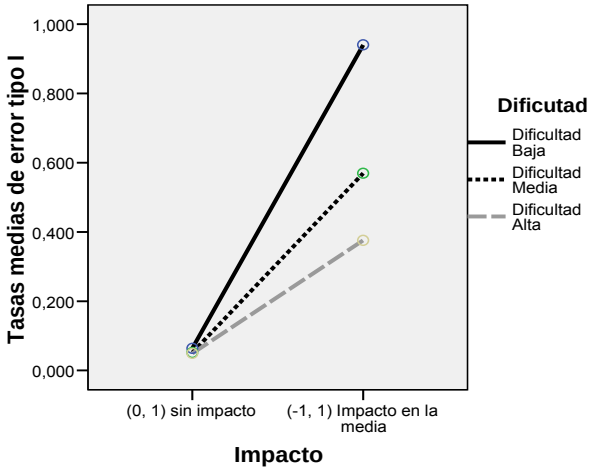
La correlación entre la tasa de falsos positivos y la dificultad (r = -0.162, p = 0.002) y falsos positivos y discriminación (r = 0.122, p = 0.021) confirma esta tendencia, es decir, ítems con baja dificultad y alta discriminación se asocian con altas tasas de falsos positivos en la prueba conjunta de DIF. En la prueba de DIF uniforme sólo se da una correlación de la tasa de falsos positivos con el parámetro de dificultad (r = -0.218, p = 0.000), y las tasas de error tipo I para la prueba de DIF no uniforme correlacionan con el parámetro de discriminación (r = 0.118, p = 0.025) y el de dificultad (r = -0.124, p=0.019).



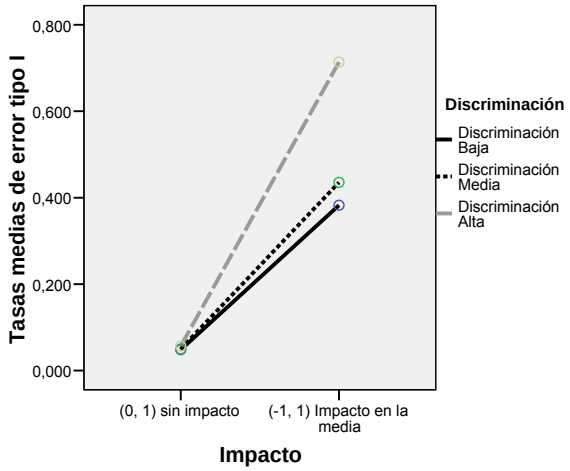
7a. 7b.

Figura 7. Tasas medias de falsos positivos en la prueba conjunta de DIF, con α = 0.05, (7a) Razón de tamaño * dificultad. (7b) Razón de tamaño * discriminación.

La exploración del impacto y los parámetros de dificultad y discriminación sugiere que en la interacción impacto * dificultad (Figura 8a), la mayor tasa de falsos positivos se observó en la condición de diferencia en la media y baja dificultad (tasas mayores a 80%), con el consecuente decremento en las tasas de detecciones incorrectas a medida que aumenta el parámetro de dificultad. En la interacción impacto * discriminación (Figura 8b), se aprecia una alta tasa de falsos positivos en la condición de diferencias en la media y alta discriminación. Al disminuir el nivel de discriminación, desciende la tasa de falsos positivos en la condición de diferencias en la media. En ambas interacciones, la condición de no impacto reportó tasas similares de falsos positivos para los tres niveles de dificultad y discriminación. Esta tendencia se observa también en la prueba de DIF uniforme y no uniforme.



8a.



8b.

Figura 8. Tasas medias de falsos positivos en la prueba conjunta de DIF, con α = 0.05, (8a) Impacto * dificultad. (8b) Impacto * discriminación.

Teniendo en cuenta los presentes hallazgos, en donde la tasa de falsos positivos se encuentra fuertemente influenciada por los parámetros de dificultad y de discriminación, se procedió a examinar los ítems no DIF y su proporción de detecciones incorrectas, con el objetivo de establecer cuáles ítems presentan mayor tasa de falsos positivos y sus parámetros asociados (Tabla 13).

De acuerdo con esta tabla, se observa que la mayoría de los ítems presentan tasas de falsos positivos superiores al 10%. Los ítems que presentan una mayor tasa de error tipo I en la condición de 0% de DIF son los ítems 1, 5, 7, 9, 10, 11, 12, 15, 16, 17, 19, 21, 22, 27 y 29. En la tasa de falsos positivos por ítem pueden observarse dos tendencias principales: (a) Altas tasas de error tipo I en las tres pruebas de detección de DIF (ítems 5,7, 9-12, 16, 19, 22, 27 y 29); y (b) Altas tasas de error tipo I en la prueba conjunta de DIF y en la prueba de DIF no uniforme (ítems 1, 15, 17, 21). Los niveles de dificultad y discriminación para los ítems que conforman estos dos grupos, se observan en la tabla 14.

Tabla 13. Media de falsos positivos para la prueba conjunta de DIF, DIF uniforme y no uniforme con α = 0.05 y α = 0.01, en la condición de 0% de DIF[a]

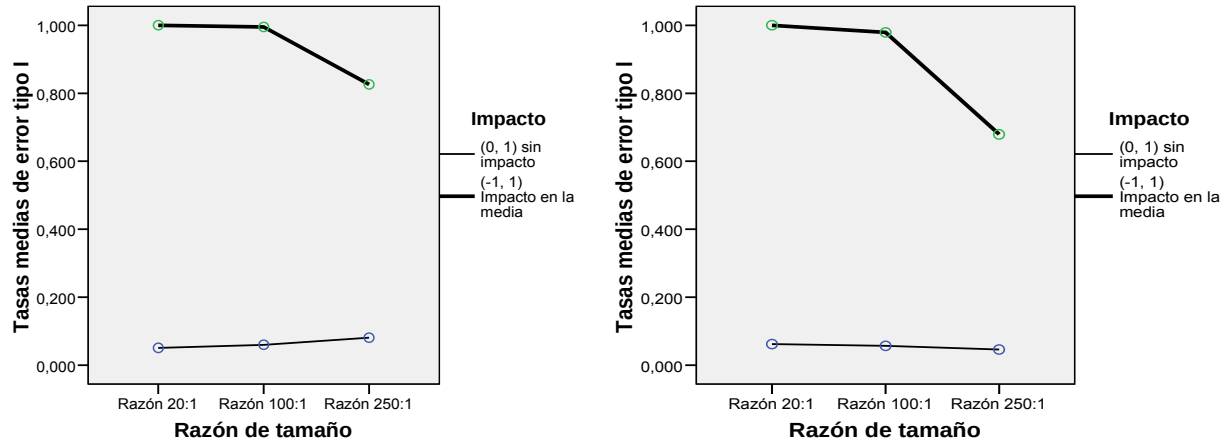
	α = 0.05			α = 0.01		
Ítem	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}
1	0.272	0.115	0.235	0.197	0.060	0.177
2	0.183	0.144	0.133	0.126	0.105	0.099
3	0.234	0.151	0.187	0.156	0.111	0.129
4	0.223	0.126	0.175	0.151	0.059	0.118
5	0.323	0.227	0.269	0.245	0.177	0.186
6	0.173	0.128	0.072	0.089	0.060	0.025
7	0.395	0.329	0.282	0.317	0.268	0.218
8	0.191	0.152	0.138	0.136	0.111	0.103
9	0.464	0.410	0.318	0.399	0.361	0.230
10	0.471	0.418	0.323	0.411	0.367	0.237
11	0.316	0.232	0.285	0.250	0.196	0.231
12	0.279	0.205	0.210	0.203	0.160	0.155
13	0.205	0.159	0.163	0.160	0.109	0.127
14	0.123	0.070	0.061	0.045	0.022	0.020
15	0.315	0.129	0.276	0.236	0.074	0.221
16	0.393	0.354	0.260	0.337	0.294	0.178
17	0.276	0.130	0.237	0.203	0.066	0.184
18	0.193	0.086	0.141	0.104	0.035	0.076
19	0.342	0.271	0.230	0.272	0.220	0.174
20	0.191	0.152	0.132	0.133	0.107	0.099
21	0.272	0.124	0.233	0.203	0.068	0.184
22	0.321	0.247	0.221	0.242	0.193	0.170
23	0.203	0.156	0.136	0.142	0.102	0.095
24	0.199	0.143	0.167	0.139	0.110	0.125
25	0.205	0.157	0.117	0.141	0.094	0.073
26	0.203	0.136	0.170	0.143	0.105	0.128
27	0.310	0.237	0.267	0.242	0.199	0.212
28	0.220	0.153	0.174	0.150	0.110	0.133
29	0.502	0.451	0.331	0.446	0.413	0.219
30	0.210	0.164	0.142	0.148	0.120	0.104

a. Los ítems resaltados en negrita presentan FP> 15%.

Tabla 14. Parámetros de dificultad y discriminación para los ítems de los dos grupos de detecciones incorrectas

Altas tasas de error Tipo I en las tres pruebas de DIF			Altas tasas de error tipo I en la prueba conjunta de DIF y en la prueba de DIF no uniforme		
Ítem	*a*	*b*	Ítem	*a*	*b*
5	baja	media	1	alta	alta
7	alta	media	15	alta	alta
9	alta	media	17	alta	alta
10	alta	media	21	alta	alta
11	alta	media			
12	media	media			
16	media	media			
22	media	media			
27	media	media			
29	alta	baja			

Como se aprecia en la tabla 14, las mayores detecciones de falsos positivos para las tres pruebas de DIF se encuentran particularmente en los ítems con dificultad baja y media y discriminaciones altas. Dentro de este grupo, los ítems con más tasa de error tipo 1 son el ítem 29 y el 10, ambos se caracterizan por valores altos de discriminación (1.607 y 1.404), respectivamente; sin embargo, a diferencia del ítem 10, el ítem 29 posee *baja dificultad* (*b* = -1.550), razón por la cual es detectado por las tres pruebas de DIF; recordemos que la baja dificultad del ítem se asocia con altas tasas de error tipo I y éste es el único ítem de la prueba que posee dificultad baja. En relación con la interacción razón de tamaño * impacto, el ítem 29 (Figura 9a) y el 10 (Figura 9b) presentan tasas de falsos positivos cercanas al 100% en la prueba conjunta de DIF, razón 20:1 y diferencia en la media, con α = 0.05. La disminución más marcada en la tasa de error tipo I se observa en la condición de 250:1, siendo el ítem 10 el que más baja tasa de error tipo I reporta (67.9%).



9a. 9b.

Figura 9. Tasas de falsos positivos en la condición de diferencias entre grupos (impacto) y razón de tamaños (χ^2_{2df}, α = 0.05). (9a) Ítem 29. (9b) ítem 10.

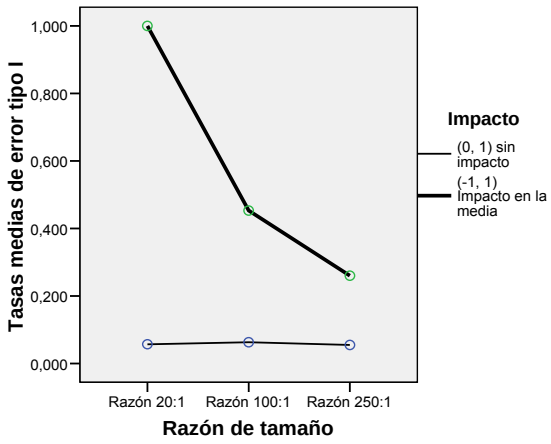


Figura 10. Tasas de falsos positivos para el ítem 15 en la condición de diferencias entre grupos (impacto) y razón de tamaños ($\chi^2 NU_{1df}, \alpha$ = 0.05)

En los 4 ítems que conforman el segundo grupo de detección, todos se caracterizan por valores altos en dificultad y discriminación. El ítem con valores más altos dentro de este grupo es el ítem 15 (*a*: 3.018, *b*: 1.857). Para este ítem (Figura 10), se puede observar que las diferencias de media y la razón 20:1 para la prueba de DIF no uniforme con α = 0.05 presentan la mayor tasa de error tipo I (99.3%). A partir de la razón 100:1, la disminución en la tasa de falsos positivos es cercana al 60%, y en 250:1, alcanza una tasa de 18%.

Potencia

El examen de la tasa de detecciones correctas de la prueba conjunta de DIF en relación con la razón de tamaños y los parámetros de dificultad muestra que *la razón de 20:1,* y *dificultad media y alta* reporta una tasa de detecciones correctas mayor a 80% (Figura 11). Las detecciones correctas de ítems DIF correlacionan con la dificultad (r = -0.234, p = 0.015), y con el porcentaje de ítems DIF en la prueba (r = -0.193, p = 0.046). Asimismo, existe asociación entre la dificultad y el porcentaje de DIF (r= 0.378, p=0.000), lo que indica que altas tasas de detecciones correctas se presentan cuando la dificultad presenta niveles medios y el porcentaje de ítems DIF en la prueba es pequeño, y en segundo lugar, el aumento de la dificultad del ítem se asocia con alta tasa de detecciones correctas cuando la prueba tiene 20% de ítems con DIF.

En relación con los niveles de dificultad óptimos para determinar que un ítem con DIF sea detectado correctamente por la RL, cabe anotar que los ítems DIF simulados en el presente estudio sólo poseen dificultad media y alta, no se dispone de ítems con dificultad baja que permitan contrastar en mayor profundidad el significado de la correlación obtenida entre potencia y dificultad.

La correlación entre la tasa de detecciones correctas y la razón de tamaños (r = -0.369 p=0.000), sugiere que a medida que aumenta la razón de tamaño a 100:1 disminuye la potencia en los ítems que presentan dificultad media y alta. Con un 20% de DIF y razón 100:1, el descenso en la potencia para los ítems con dificultad alta es cercano al 50%, y con una razón 250:1 la potencia es menor a 0.30. Esta tendencia se observa también en la prueba de DIF uniforme y no uniforme.

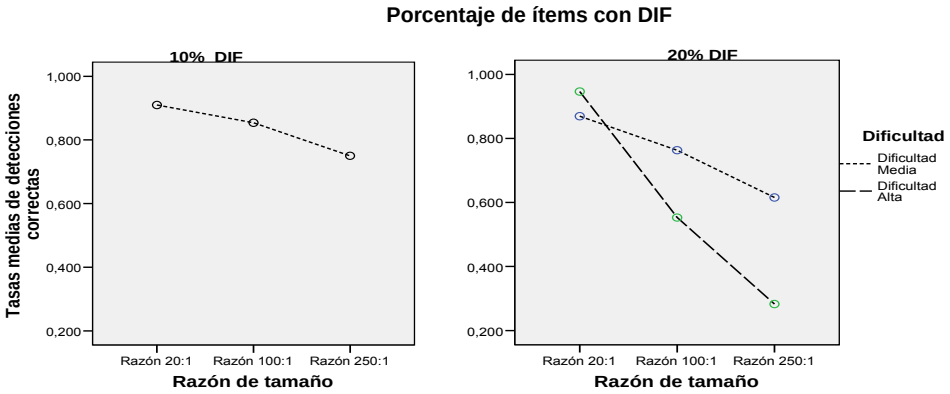


Figura 11. Tasas medias de detecciones correctas en la prueba conjunta de DIF, tomando razón * dificultad, con α = 0.05.

Con relación a la discriminación y la razón de tamaños (Figura 12), se observa que los ítems con discriminación alta son detectados en un 100% en una prueba con 10% de ítems DIF, y no hay disminución de la potencia cuando incrementa la razón de tamaño. A partir de la razón 100:1 se aprecia una disminución de la potencia en los ítems con discriminación media y baja. La peor potencia para la condición de 10% de DIF se presenta en la razón 250:1 en ítems con discriminación media. Cuando la prueba posee un 20% de ítems DIF, hay una disminución de la potencia en ítems con discriminación alta y baja (valores cercanos a 90%). A partir de la razón 100:1 hay una disminución de la potencia en los tres niveles de discriminación, siendo la discriminación baja en conjunción con la razón 250:1, la que peor potencia presenta.

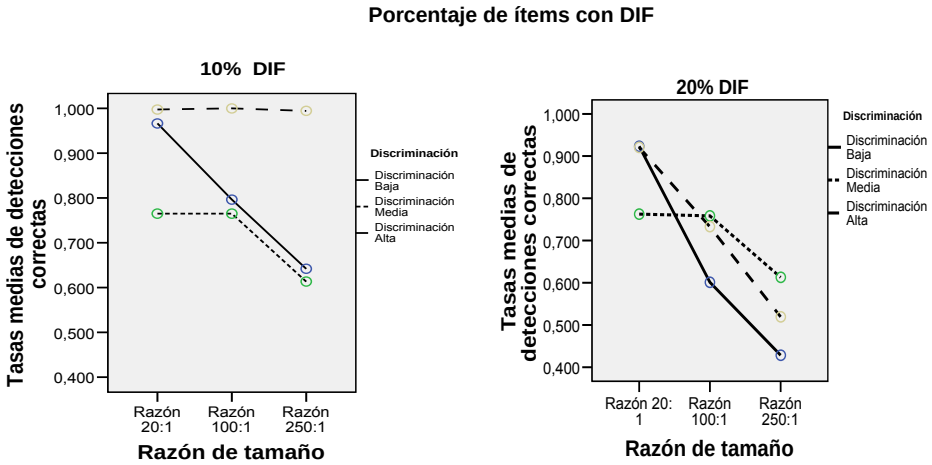


Figura 12. Tasas medias de detecciones correctas en la prueba conjunta de DIF, tomando razón * discriminación, con α = 0.05.

Para la prueba de DIF uniforme se encontró una correlación entre la tasa de detecciones correctas con la razón de tamaños (r = -0.379, p=0.001), el parámetro de dificultad (r = -0.535, p=0.000), y con el porcentaje de ítems con DIF (r = -0.278, p=0.018); y de la dificultad con el porcentaje de ítems DIF (r = 0.50, p=0.000), conservando la misma tendencia que la prueba conjunta de DIF.

Una correlación que resultó significativa para la prueba de DIF uniforme fue la asociación entre la tasa de detecciones correctas y el modelo de simulación (r = -0.395, p=0.001). Si se examina el modelo de simulación y el parámetro de dificultad, la condición de modelo de 1 parámetro en ambos porcentajes de ítems DIF y considerando ítems de dificultad media, presenta una potencia cercana al 100%. En la condición de 3 parámetros la potencia disminuye en 10 y 20% de DIF para dificultad media y alta; siendo los ítems con dificultad alta en condición de 3 parámetros y 20% de ítems DIF, aquellos que presentan una menor potencia.

En el caso de la discriminación y el modelo de simulación para un 10% de ítems DIF, los ítems que presentan discriminación alta son detectados en un 100%, independientemente del modelo de simulación. Si se tiene un 20% de ítems DIF, la potencia decrece para los ítems con discriminación alta y baja, siendo más notoria esta disminución en los ítems con discriminación baja y modelo de 3 parámetros (potencia menor a 0.50).

La prueba de DIF no uniforme revela correlaciones entre la tasa de detecciones correctas con la razón de tamaños (r = -0.344, p=0.040), la discriminación (r = -0.376, p=0.024), el impacto (r = 0.343, p=0.040), y el modelo de simulación (r = 0.379, p=0.023); y de la discriminación con el porcentaje de ítems DIF (r = 0.500, p=0.002).

El impacto en relación con la dificultad indica que las diferencias en la media generan una mayor tasa de detecciones correctas en la prueba de 10% de DIF que en la de 20%. Sin embargo, los dos ítems simulados como DIF no uniforme son de dificultad media, lo que impide realizar una mayor discusión en torno al comportamiento de ítems con dificultad alta o baja respecto al impacto.

La discriminación y el impacto muestran que, en niveles medios de discriminación y diferencias en la media. las tasas de detección son las mismas, independientemente del porcentaje de DIF. Cuando se consideran ítems con discriminación alta y sin impacto, en una prueba con 20% de ítems DIF, la potencia de

la RL es menor. Similar tendencia aplica para la interacción entre el modelo de simulación y los parámetros de dificultad y discriminación.

El examen de las detecciones correctas de los ítems DIF en condiciones de 10 y 20% de DIF con α = 0.05 y α = 0.01 muestra tendencias de interés en relación con los parámetros de dificultad y discriminación (Tabla 15):

1. En la condición de 10% de DIF, en la que se incluyeron dos ítems con DIF uniforme (3 y 10) y un ítem con DIF no uniforme (9), se observa una mayor tasa de detecciones correctas para los ítems 10 y 3 en la prueba conjunta de DIF y en la prueba de DIF uniforme. Con α = 0.05, la detección del ítem 10 es cercana al 100% y para el ítem 3 las tasas son del 80%. Para el ítem 9 la detección de DIF en la prueba conjunta fue mayor al 70% con α = 0.05, sin embargo, la prueba de DIF no uniforme sólo fue mayor al 60% con α = 0.05. Al examinar los parámetros de los ítems 10 y 3 se encuentra que ambos poseen dificultad media; la diferencia radica en la discriminación del ítem, ya que el ítem 10 posee discriminación alta, mientras que el ítem 3 posee baja discriminación.

2. En la condición de 20% de DIF, en la que se incluyeron cuatro ítems con DIF uniforme (3, 10, 21 y 25) y dos ítems con DIF no uniforme (9 y 24), los ítems con DIF uniforme presentan tasas más altas de detección correcta que los ítems con DIF no uniforme. El ítem 10 es el que posee mayor tasa de detecciones correctas, seguido del 3, 21 y 24. El ítem 3 muestra una reducción en la tasa de detección en comparación con la condición de 10%, y los ítems 21 y 25 poseen tasas menores al 70% para la prueba conjunta de DIF y la prueba de DIF uniforme.

Al revisar los ítems con DIF no uniforme, el ítem 9 presenta una tasa mayor al 70% sólo para la prueba conjunta de DIF con α = 0.05, sin embargo la prueba de DIF no uniforme arroja un valor menor a 0.70 en ambos niveles de significación. El ítem 24, por su parte, muestra tasas menores a 0.70, para ambos niveles de significación. En relación con los parámetros de los ítems 9 y 24 se encuentra que ambos poseen dificultad media (*b*: -0.955 y *b*= 0.761, respectivamente); la diferencia radica en la discriminación del ítem, ya que el ítem 9 posee discriminación media (*a*= 0.74), mientras que el ítem 24 posee alta discriminación (*a*= 1.005).

Tabla 15. Media de detecciones correctas para la prueba conjunta de DIF, DIF uniforme y no uniforme, por porcentaje de DIF (α = 0.05 y α = 0.01) [a]

% DIF	Ítem	α = 0.05			α = 0.01		
		χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}
10% DIF	3	**0.802**	**0.791**	0.190	**0.729**	**0.724**	0.108
	9	**0.715**	0.630	0.655	0.651	0.561	0.559
	10	**0.997**	**0.997**	0.383	**0.995**	**0.994**	0.254
20% DIF	3	**0.779**	**0.769**	0.182	**0.713**	**0.708**	0.104
	9	**0.711**	0.621	0.644	0.647	0.547	0.543
	10	**0.998**	**0.998**	0.362	**0.915**	**0.995**	0.226
	21	0.665	0.512	0.318	0.552	0.397	0.228
	24	0.509	0.413	0.331	0.481	0.326	0.241
	25	0.523	0.501	0.093	0.402	0.388	0.037

a. Detecciones resaltadas en negrita representan potencia > 0.70.

Un aspecto que resulta de gran interés en la observación de la presente tabla es la alta detección incorrecta de ítems DIF según su clasificación (i.e., un ítem simulado como DIF uniforme es detectado como DIF no uniforme, y viceversa), lo que conduciría a un error tipo I de clasificación del ítem DIF. Algunos ejemplos de ello pueden verse en las tasas de detección del ítem 9 como DIF uniforme (63% en α = 0.05), y el ítem 10 detectado como DIF no uniforme (38.3%), con α = 0.05.

Factores que afectan la clasificación incorrecta de ítems DIF en RL (condición 10% y 20% de DIF)

Con el objetivo de indagar un poco más a profundidad qué factores inciden en la clasificación incorrecta de ítems DIF en las condiciones de 10% y 20% de DIF, según lo observado en la tabla 15, se realizaron dos análisis de varianza de cuatro vías, en donde se tomó como variable dependiente la tasa de clasificaciones incorrectas de los ítems DIF, para cada tipo de DIF, en las condiciones de 10% y 20% de DIF en las tres pruebas de detección de DIF (χ^2_{2df} , χ^2U_{1df} y χ^2NU_{1df}), y como variables independientes la razón de tamaños, el impacto, el modelo de simulación y el porcentaje de DIF.

Tabla 16. Valor F y significación de los efectos sobre la clasificación incorrecta de ítems DIF en la condición de 10% y 20% de DIF (a). DIF Uniforme. (b). DIF no uniforme

16a.

Anova Error Tipo I, DIF Uniforme, Detección No Uniforme	α = 0.05		α = 0.01	
Factor	F	Significación	F	Significación
Razón	17.242	.000	14.431	.000
Impacto	23.303	.000	17.741	.000
Porcentaje de DIF	.792	.378	.394	.533
Modelo	.305	.584	.333	.566
Razón * Impacto	4.463	.017	6.340	.004
Razón * Porcentaje de DIF	.103	.903	.119	.888
Impacto*Porcentaje de DIF	.609	.439	.721	.400
Razón * Modelo	.074	.929	.037	.963
Impacto*Modelo	.403	.529	.083	.775
Porcentaje de DIF*Modelo	.925	.341	.938	.338

16b.

Anova Error Tipo I, DIF No Uniforme, Detección Uniforme	α = 0.05		α = 0.01	
Factor	F	Significación	F	Significación
Razón	9.081	.004	13.438	.001
Impacto	24.829	.000	20.794	.001
Porcentaje de DIF	2.161	.167	2.250	.159
Modelo	32.917	.000	22.118	.001
Razón * Impacto	.857	.449	1.225	.328
Razón * Porcentaje de DIF	.575	.577	.671	.529
Impacto*Porcentaje de DIF	.216	.650	.223	.645
Razón * Modelo	.257	.778	.653	.538
Impacto*Modelo	12.554	.004	7.131	.020
Porcentaje de DIF*Modelo	.141	.713	.308	.589

Puede verse en la tabla 16a que la razón de tamaños y el impacto influyen en la presentación de un error tipo I muy especial, esto es, un ítem con DIF uniforme es detectado anormalmente como si fuera un ítem con DIF no uniforme. Este tipo de error tiende a presentarse en condiciones de diferencia de medias e ítems con discriminación alta, y disminuye conforme aumenta la razón de tamaños; la conjugación de las

condiciones de diferencia en la media y una razón de tamaños pequeña puede también determinar la ocurrencia de este error de detección.

Así mismo, puede verse en la tabla 16b la presentación de un error tipo I converso del anterior, esto es, un ítem con DIF no uniforme que es detectado anormalmente como presentando DIF uniforme. En este caso, las variables determinantes son la razón de tamaños (disminuye en las razones altas), el impacto (aumenta en las condiciones sin impacto), y el modelo (aumenta en el modelo de 1 parámetro). La interacción impacto * modelo señala que las tasas más altas de error tipo I se presentan cuando hay diferencia de medias y modelo de 1 parámetro.

Capítulo 6
DISCUSIÓN Y CONCLUSIONES

La presente investigación tenía como objetivo general evaluar el efecto de la razón de tamaños en el funcionamiento de la regresión logística para la detección del funcionamiento diferencial del ítem, a partir de la realización de un estudio de simulación. Para lograr este objetivo en primer lugar se consideraron tres niveles de razón de tamaño (20:1, 100:1 y 250:1) en un n = 65000, a fin de conservar la proporción de personas entre los grupos focal y de referencia que se pueden presentar a pruebas masivas de selección en ámbitos educativos. Además de la razón de tamaños como variable central que puede incidir en la potencia y error tipo I de la regresión logística, se tomaron en cuenta otras variables como porcentaje de DIF, diferencias en las distribuciones de habilidad de los grupos y modelos de simulación, las cuales han sido reportadas en la literatura sobre la evaluación de procedimientos DIF.

Responder a estas dos preguntas de investigación implicaba tener presente a su vez dos elementos esenciales: Cuáles son los efectos principales de cada una de las variables manipuladas tomadas en forma separada y qué efectos de interacciones o combinaciones de variables pueden aportar en la explicación de la potencia y el error tipo I del estadístico en cuestión. El análisis de los efectos principales e interacciones de las variables manipuladas, por ende, proporciona indicaciones sobre las condiciones de razón de tamaño que permitirían una adecuada potencia y control de error tipo I de la RL en contextos masivos de aplicación de pruebas como el aquí simulado, en donde las diferencias en cuanto a número de individuos del grupo focal por número de individuos del grupo de referencia son considerables. Así mismo, la exploración de cuáles condiciones en donde dos o más variables pueden presentarse en un instrumento de medición da herramientas a los constructores de pruebas para identificar aquellas características que conjugadas pueden hacer que la probabilidad de detectar ítems como DIF sea mayor, o qué situaciones permiten un adecuado control de la tasa de ítems falsamente identificados con DIF.

Una primera aproximación al error tipo I consistió en examinar el efecto de las variables manipuladas sobre la tasa de falsos positivos de los ítems no DIF en la condición de 0% de DIF. En relación con los efectos principales, *el impacto y la razón de tamaños fueron los más significativos para las tres pruebas de detección de DIF*. La tasa de error tipo I es mayor cuando hay diferencias en las distribuciones de habilidad (en este caso, diferencias en la media), hallazgo que es concordante con lo reportado en los estudios de Jodoin & Gierl (2001) y Jodoin & Huff (2001). Las diferencias en la distribución de habilidad de los grupos, que de acuerdo con Jodoin & Huff (2001) pueden expresarse también en términos de área de solapamiento entre las CCI para los grupos focal y de referencia, por ende, juegan un papel decisivo en la tasa de falsos positivos de la regresión logística, de manera que al decrecer el área de solapamiento entre las CCI, aumenta la proporción de falsos positivos. French & Maller (2007), quienes emplearon los mismos niveles de impacto que los del presente trabajo, encontraron que las mayores tasas de error tipo I se concentran en las condiciones de impacto, tamaño del grupo focal mayor a 500, y con el empleo de la prueba conjunta de DIF, sin considerar una medida del tamaño del efecto. Estos hallazgos confirman la sugerencia que Raju formulara en 1988, la cual afirmaba que son las diferencias en habilidad entre el grupo focal y el grupo de referencia lo que afecta al error tipo I más que el porcentaje de DIF (Akelo, 2008; Wang & Yeh, 2003).

El comportamiento de la razón de tamaños en el error tipo I de la RL señala que razones de tamaño pequeñas (20:1, con N_f= 3095), se acompaña de incrementos en la tasa de error tipo I (49.6%). A medida que aumenta la razón de tamaños, es decir, disminuye el número de personas en el grupo focal respecto del grupo de referencia, el error tipo I disminuye. A partir de la condición 100:1 la tasa de falsos positivos disminuye casi un 50% si se compara con la razón de tamaño anterior, no obstante, las tasas de falsos positivos para los tres niveles de razón de tamaño exceden el criterio liberal de Bradley.

Si se observa el efecto de las interacciones, aquella que resultó más significativa fue la de razón * impacto, en donde la tasa de falsos positivos es mayor en la razón de 20:1 y con diferencias en la media, alcanzando un valor

cercano al 90% de error tipo I. Al aumentar la razón de tamaño a 100:1, se modula el efecto del impacto en la media sobre la tasa del error tipo I, logrando una disminución considerable del mismo. Bolt y Gierl (2004) señalan que con amplios tamaños de muestra y diferencias entre los tamaños de muestra de los grupos focal y de referencia, se observa una importante inflación del error tipo I para los procedimientos de DIF. Arias (2008) y Berrío (2008) reportaron similar tendencia, aunque las tasas de falsos positivos obtenidas en estas dos investigaciones fueron menores a las del presente estudio. Akelo (2008) en su estudio sobre la detección de DIF en grupos con diferentes tamaños de muestra, a partir de la comparación del SIBTEST con corrección de la regresión y con el MH muestra como uno de sus hallazgos más importantes que:

> *No sólo las diferencias en habilidad pueden afectar el error tipo I de un procedimiento, pero también la razón de tamaños entre el grupo focal y de referencia también tiene efectos significativos sobre el poder estadístico de los procedimientos de detección de DIF* (p. 75).

La literatura sobre evaluación de técnicas de DIF y efecto de la razón y el impacto en el error tipo I muestra tasas de error entre el 5 y el 20%, aproximadamente. Sin embargo, para este estudio se observan inflaciones del error tipo I mayores al 80% para dos condiciones experimentales, que precisamente son las *de razón 20:1 y diferencias en la media*, además se observa alto error tipo I por ítem (tasas de FP mayores a 10%), y tomando la interacción entre las variables. Entre las explicaciones a ello pueden mencionarse:

1. **Las dificultades de la regresión logística para diferenciar impacto de DIF.** Uno de los propósitos en regresión consiste en formular un modelo ideal que cubra la mayor cantidad de casos dentro de éste. En razones de tamaño pequeñas, cuando las diferencias entre los tamaños en el grupo focal y de referencia son menores y el grupo focal tiene una cantidad considerable de individuos, el modelo tiende a ajustarse a diferencias intrafocales o intrarreferenciales, de ahí que la RL sea sensible a diferencias considerables en

los ítems o en los individuos, y por ende, sean identificadas erróneamente como DIF. Tal es el caso del impacto en la media, en donde las diferencias en la distribución de la habilidad con respecto a la media para los grupos focal y de referencia fueron identificadas por la RL como DIF. En consecuencia, razones pequeñas acompañadas de diferencias en la media presentan incrementos en la tasa de error tipo I. French y Maller (2007) mostraron en sus estudios que grupos focales mayores a 500 individuos reportaban mayor tasa de error tipo I. La falta de diferenciación de la RL entre DIF e impacto conlleva implicaciones prácticas importantes, dado que las diferencias reales en la magnitud de atributo en los individuos evaluados son tomadas como diferencias artificiales generadas por el instrumento de medida, con lo cual se ignorarían las diferencias reales que pueden existir en la distribución del atributo en los individuos

Cuando la razón de tamaños se incrementa, es decir, hay un menor número de individuos en el grupo focal y las diferencias numéricas con relación al grupo de referencia son mayores, el modelo de regresión tiende a responder mejor a las diferencias intrarreferenciales que a las intrafocales, dada la poca cantidad de individuos en el grupo focal. Esto conlleva a que aumente la probabilidad de que la diferencia de un individuo en el grupo focal no se encuentre dentro del modelo de regresión, generando con ello una disminución del error tipo I a medida que la razón de tamaños aumenta. No obstante, las tasas de error tipo I para las razones de tamaño más extremas, cuando hay diferencias en la media, exceden los criterios liberales de Bradley, incidiendo con ello en la robustez de la RL.

2. **El amplio tamaño de muestra empleado para la investigación** (**n=65000).** El hecho de contar con un tamaño de muestra grande evidencia que entre mayor sea el tamaño de muestra, es más probable que se presente una alta tasa de falsos positivos (Cohen, 1990, 1994; Traub, 1983), ya que pequeñas diferencias pueden ser reportadas como significativas (Akelo, 2008).

3. **Los niveles de razón de tamaño considerados en la investigación.** El estudio de Cromwell (2006) señaló que el tamaño de muestra total que se emplea en las investigaciones sobre DIF era más relevante que el porcentaje de diferencia entre el tamaño del grupo focal y de referencia. No obstante, empleó 5

combinaciones de tamaño de muestra ($1000_r/100_f$, $500_r/100_f$, $300_r/300_f$, $1000_r/300_f$ y $1000_r/1000_f$), que representan razones de tamaño de 0.10:1, 0.05:1, 1:1, 0.30:1 y 1:1, respectivamente, por tanto representan diferencias leves o nulas en cuanto a la distribución del tamaño de los grupos, que no reflejan grandes discrepancias en cuanto a error tipo I se refiere. Los tres niveles de tamaño escogidos para la presente investigación buscaban evidenciar las marcadas diferencias respecto a la proporción de individuos de los grupos definidos como focal y de referencia que presentan pruebas educativas a gran escala, en el contexto colombiano. Se esperaría por consiguiente que estas altas razones de tamaño generen incrementos importantes en el error tipo I de los procedimientos de detección de DIF.

Una variante del error tipo I que se buscó indagar en la presente investigación fue examinar los factores que afectaban el error tipo I en los ítems no DIF para las condiciones de 10% y 20% de DIF. A diferencia del anterior análisis, se tuvo en cuenta el porcentaje de DIF como variable independiente. Los resultados obtenidos para las condiciones de 10% y 20% de DIF conservan las mismas tendencias en cuanto a efectos principales e interacciones según la prueba de detección de DIF. En relación con el porcentaje de DIF en la prueba, éste no reporta efectos significativos. Aún cuando el análisis de la media de detecciones incorrectas de DIF de los ítems no DIF indica que el error tipo I se incrementa cuando aumenta el porcentaje de DIF, las diferencias entre 10% y 20% de DIF son pequeñas para ser consideradas no significativas. Akelo (2008); Narayanan & Swaminathan (1996) y Wang & Yeh (2003) han argumentado a favor de esta idea, señalando que aunque la tendencia general del porcentaje de DIF consiste en incrementos del error tipo I cuando aumenta el porcentaje de DIF, no han resultado ser significativos estadísticamente.

En relación con la potencia de la regresión logística, la *razón de tamaños* es la variable determinante en la detección de ítems con DIF para las tres pruebas de detección de DIF. Las razones de tamaño pequeñas, en donde la diferencia de tamaños de los grupos focal y de referencia no es muy grande, favorecen la detección de DIF, y a medida que aumenta la razón de tamaños, la tasa de

detecciones correctas disminuye. Los estudios iniciales sobre RL habían documentado inicialmente el efecto del tamaño de muestra total en la potencia del procedimiento (Swaminathan & Rogers, 1993; Rogers & Swaminathan, 1993; Narayanan & Swaminathan, 1996), mostrando que los incrementos en la potencia de los procedimientos DIF se relacionan con el aumento en el tamaño de muestra. Al respecto Thompson (1996) indica que si se posee un tamaño de muestra bastante amplio, siempre se rechazará la hipótesis nula, y todos los métodos de DIF emplean la prueba de significancia de la hipótesis nula. Teniendo en cuenta el tamaño de muestra empleado en el presente estudio, se esperaría encontrar una alta tasa de detecciones correctas.

Sin embargo, el tamaño de muestra no es el único elemento determinante en la potencia, sino que las razones de tamaño influyen de manera decisiva. A partir del estudio de Narayanan y Swaminathan (1996), en el que se planteó un posible efecto de la razón de tamaños, se han conducido estudios posteriores que han trabajado con combinaciones de tamaños de muestra para el grupo focal y de referencia; los hallazgos señalan que cuando las diferencias en tamaños de muestra en los grupos a evaluar son menores, la detección correcta de ítems DIF es mayor (Gierl, Bisanz, Boughton & Khaliq. 2001; Jodoin & Gierl, 2001; Herrera, 2005). Es por ello que varios autores afirman que amplios tamaños de muestra y pequeñas diferencias entre el grupo focal y de referencia en cuanto al tamaño de muestra por subgrupo se constituyen en la condición ideal para el estudio del DIF cuando el DIF está presente (Jodoin & Gierl, 2001).

Examinando las razones de tamaño empleadas en el estudio, se observa que la mayor tasa de detecciones correctas se observa en la razón 20:1, con una tasa cercana al 90%, y desciende de manera importante al aumentar la razón de tamaño. Cabe señalar que los resultados de potencia obtenidos para la razón 20:1 deben ser tomados con cautela, ya que esta misma razón de tamaño presentó la mayor tasa de error tipo I, por ende, la significación estadística de la potencia se ve comprometida por las altas tasas de error tipo I. En la condición 100:1, la potencia de la RL para la prueba conjunta de DIF con α = 0.05, es mayor que 0.70, criterio que ha sido considerado por varios autores como un nivel

aceptable de tasa de detección de DIF (French & Maller, 2007; González-Romá et al., 2006; Kaplan & George; 1995).

A diferencia del análisis de error tipo I, en donde el impacto, la razón de tamaños y el ajuste incidieron en las tres pruebas de detección de DIF, para la potencia de la RL se encontraron efectos principales diferenciales del porcentaje de ítems DIF en la prueba conjunta de DIF y en la prueba de DIF uniforme; y del impacto y el modelo de simulación en la prueba de DIF no uniforme.

Con relación al impacto, se observó que las diferencias en la media se asocian con tasas de detecciones correctas mayores a 0.70 en las tres pruebas de detección de DIF para las razones más pequeñas, hallazgo es concordante con lo obtenido por otros autores (Jodoin & Gierl, 2001; French & Maller, 2007; Arias, 2008; Berrío, 2008). French y Maller (2007) manipularon el impacto de la misma manera que el presente estudio, y en cuanto a potencia encontraron tasas de detecciones correctas > 0.70 para DIF no uniforme en diferente habilidad, con 500_r y 500_f. Para 1000_r y 500_f se alcanza la potencia con el uso de la prueba estadística para DIF uniforme, y con prueba estadística y prueba estadística con purificación iterativa para DIF no uniforme, en diferencias de habilidad. Para 1000_r, 1000_f, se alcanza buena potencia con prueba estadística y prueba estadística con purificación iterativa para DIF. Estos mismos autores indican además que *aún cuando el poder parezca adecuado en muchos casos con impacto, es más probable que se presenten inflaciones importantes del error tipo I.* Esto se observa para el presente estudio en el análisis de las tasas de detecciones correctas de los ítems DIF, las cuales son mayores al 70% en la razón 20:1 y 100:1, en la prueba conjunta de DIF y en la prueba de DIF uniforme. Aunque el poder parece ser adecuado en muchos casos con impacto, es más probable que sea producto de una inflación de error tipo I (French & Maller, 2007).

A semejanza de lo encontrado en el análisis de error tipo I, razones de tamaño pequeñas y diferencias en la media se asocian con alta tasa de detecciones correctas de ítems DIF. Nuevamente, podría argumentarse la hipótesis de una dificultad de la regresión logística para diferenciar impacto de DIF como explicación de los altos valores de potencia registrados en RL, en la medida

en que las diferencias en la distribución de habilidad son susceptibles a ser identificadas como DIF, particularmente cuando hay un mayor número de individuos en el grupo focal, es decir, en razones de tamaño pequeñas.

En relación con el modelo de simulación, el modelo de 3 parámetros como aquel que reporta la mayor tasa de detecciones correctas no se sigue de lo encontrado en investigaciones anteriores. Rogers y Swaminathan (1993) encontraron que el empleo del modelo de 3 parámetros se asociaba con disminución en el poder de la RL (73% para DIF uniforme y 67% para DIF no uniforme), ya que había una interacción entre el modelo y el impacto, en donde al emplear un modelo de 3 parámetros en presencia de diferencias de habilidad, el poder de la RL disminuía. Sin embargo, cabe anotar que el modelo de ajuste empleado por los autores no fue un modelo de un parámetro sino de 2PL, en donde se busca explicar la probabilidad de acierto al ítem en función de la habilidad del individuo, y los parámetros de dificultad y discriminación. Para la condición de 2PL, no se observa interacción entre el modelo y el impacto, y las tasas de detecciones correctas son mayores (73% para DIF uniforme y 67% para DIF no uniforme).

Finch & French (2007) en su estudio sobre la comparación de 4 métodos de detección de DIF no uniforme (LR, SIBTEST, IRTLR y CFA) reportan que la RL bajo el modelo de 2 parámetros no se deja afectar por las diferencias en las distribuciones de habilidad de los grupos (tasa de 52.2%), mientras que la condición de modelo de tres parámetros y diferencias en la distribución de habilidad, da lugar a menor tasa de identificación de ítems DIF (40.2%). Tanto el estudio de Rogers como el de Swaminathan afirman que la razón por la cual la potencia es menor para el modelo de 3PL es que el modelo de la RL no se ajusta a la presencia del parámetro de pseudoazar. Aún cuando la mayoría de estudios DIF trabajan con *c* fija, con valores que oscilan entre 0.15 y 0.20, la inclusión de este parámetro hace diferencias respecto a la tasa de detecciones correctas para regresión logística.

Teniendo en cuenta lo anterior, se podría pensar que el modelo de 1 parámetro, tal como se concibió en la investigación, no es adecuado para la

regresión logística, en la medida que sólo toma en cuenta la dificultad y la habilidad, asumiendo que la discriminación es igual. Cromwell (2006) afirma al respecto que "El modelo de Rasch, sin conocimiento del parámetro *a* y *c*, conduce a decisiones incorrectas respecto a sesgo o carencia de sesgo. Los modelos de dos y tres parámetros son los más usados en los estudios de DIF" (p. 16).

Si se comparan las tasas de detección por tipo de DIF, se encuentra que la detección conjunta de DIF y la prueba de DIF uniforme son más altas que la de DIF no uniforme, hallazgo consistente con lo reportado en la literatura sobre DIF (Jodoin & Gierl, 2001). En la prueba de DIF no uniforme, sin embargo, las tasas de detecciones correctas no alcanzan el 70% para ninguno de los dos porcentajes de DIF: el ítem 9 obtuvo 65.5% para 10% de DIF y 64.4% para 20% de DIF; y el ítem 24 en condición de 20% de DIF obtuvo sólo el 33.1%, ambos ítems evaluados con α = 0.05 (Tabla 15).

Como se mencionaba en el capítulo de Resultados, un aspecto que resultó de interés en la observación de la tasa de detecciones correctas de ítems DIF es la alta detección incorrecta de ítems DIF según su clasificación (i.e., un ítem simulado como DIF uniforme es detectado como DIF no uniforme, y viceversa), lo que conduciría a un error tipo I de clasificación del ítem DIF. El análisis de este error tipo I por tipo de DIF conduce a hallazgos preliminares que podrían explorarse a profundidad en próximas investigaciones.

En el caso del error tipo I para los ítems con DIF uniforme, se encontró que, la razón de tamaños y el impacto influyen en la presentación de falsos positivos para DIF uniforme. Este tipo de error tiende a presentarse en ítems con discriminación alta e impacto en la media, y disminuye conforme aumenta la razón de tamaños; la conjugación de las condiciones de impacto en la media y una razón de tamaños pequeña puede también determinar la ocurrencia de este error de detección. En relación con este error tipo I, los ítems 10 y 21 aparecen mayormente identificados como ítems con DIF no uniforme; para el ítem 10 se observa un promedio de 38.3% y 36.3% para las condiciones de 10% y 20% de DIF, con α = 0.05, respectivamente; y para el ítem 21 se registra un promedio de 31.8% para la condición de 20% de DIF. Como características de estos ítems, se

encuentra que poseen una alta discriminación (*a*= 1.40 para el ítem 10 y *a*= 1.22 para el ítem 21). Con estos valores altos de discriminación, es factible que aun cuando ambos ítems fueron simulados como DIF uniforme, haya cruce entre las CCI de los grupos focal y de referencia en los niveles extremos de magnitud de atributo (Anexo 8), lo que es identificado por la RL como interacción y por tanto conduce a la identificación de estos ítems como DIF no uniforme.

Finch y French (2008) reportan altas tasas de este error tipo I para RL cuando el ítem posee dificultad media (32.8%), alta discriminación (23.2%), magnitud de DIF de 1 (25.2%), modelo de tres parámetros (28.1%), tamaño de muestra $1000_f/1000_r$ (25%) y diferencias en la media de -0.5 para el grupo focal (20.7%). Aun cuando el estudio de estos autores sólo reportó efectos significativos del ajuste del modelo, se observa que para nuestro estudio el efecto de la razón de tamaños e impacto sobre este error tipo I es más notorio, e incluso los ítems identificados en esta categoría presentan tasas superiores a las reportadas por Finch y French.

Cuando se presenta el error tipo I en los ítems con DIF no uniforme, las variables determinantes son la razón de tamaños (disminuye en las razones altas), el impacto (aumenta en las condiciones sin impacto) y el modelo (aumenta en el modelo de 1 parámetro). Curiosamente, cuando se conjugan simultáneamente una razón de tamaños pequeña y un ítem sin impacto, estos dos factores explican la detección errónea independientemente del modelo. En relación con este error tipo I, los ítems 9 y 24 aparecen identificados como ítems con DIF uniforme; para el ítem 9 se observa un promedio de 63% y 62.1% para las condiciones de 10% y 20% de DIF, con α = 0.05, respectivamente; y para el ítem 24 se registra un promedio de 41.3% para la condición de 20% de DIF. Como características de estos ítems, se encuentra que poseen una dificultad media (*b*= *-0.955* para el ítem 9 y *b*= *0.761* para el ítem 24).

Finch y French (2008) reportan altas tasas de este error tipo I para RL cuando el ítem posee dificultad alta (83.5%), discriminación media (85.5%), magnitud de DIF de 0.8 (65.6%), modelo de dos parámetros (64%), tamaño de

muestra 1000f/1000r (70%) y diferencias en la media de -0.5 para el grupo focal (65.1%).

Una posible explicación que proporcionan Finch y French respecto a la identificación de ítems con DIF no uniforme como uniforme es que

> *El punto de cruce entre las CCI no se encuentra localizado centralmente en la distribución de habilidad, creando así una situación en la cual uno de los grupos tiene una mayor probabilidad condicional de una respuesta correcta que el otro, a lo largo de la distribución de habilidad* (p. 755)

La observación de las CCI para los ítems 9 y 24 (Anexo 8) parecería confirmar esta tendencia, ya que en el ítem 9 el punto de cruce entre las CCI de los grupos focal y de referencia se encuentra en θ = -1.0. En niveles de atributo mayores a θ = -1.0, la CCI sugiere que la diferencia en la probabilidad de acierto para el ítem es mayor para el grupo de referencia que para el focal en gran parte del continuo de habilidad. Para el ítem 24, el cruce de las CCI se da en θ=-0.6, mostrando que para niveles bajos de magnitud de atributo la diferencia en la probabilidad de acierto para el ítem es mayor para el grupo de referencia que para el focal en buena parte del continuo de habilidad.

Por otra parte, aún cuando la dificultad y la discriminación no se manipularon sistemáticamente dentro de la presente investigación; la revisión de estudios previos (Akelo; 2008; Arias, 2008; Herrera, 2005) y la observación de las tasas de detecciones correctas y falsos positivos de los datos del presente trabajo, sugirió la posibilidad de un efecto de la dificultad y la discriminación, tanto para la potencia como para el error tipo I en la RL. Es por ello que la dificultad y la discriminación se incluyeron como variables a considerar en la potencia y el error tipo I para la regresión logística.

Para comprender la influencia de los parámetros de los ítems en la potencia y error tipo I de la RL, cabe recordar que la regresión logística es un procedimiento que tiende a buscar estimaciones de tendencias con el propósito de construir un modelo que explique la probabilidad de acierto al ítem en función de ciertas variables, que se acompañan por parámetros de los ítems (dificultad-

discriminación). Cuando existen variaciones extremas superiores o inferiores de estos parámetros, la regresión es sensible frente a estas diferencias, asignando dicho dato como elemento por fuera del modelo ideal trazado por las estimaciones de máxima verosimilitud. En consecuencia, al utilizar la regresión logística se incurre en el riesgo de aceptar detecciones incorrectas de DIF con mayor probabilidad, esto es, es más factible la ocurrencia de error tipo I. Tal es el caso de ítems con *discriminación alta y dificultad baja*, como el ítem 29, que fue detectado como DIF en una tasa promedio del 50%. Además se observaron interacciones importantes entre impacto * dificultad, e impacto*discriminación para la prueba de DIF no uniforme.

El efecto de los parámetros de los ítems sobre las tasas de error tipo I en regresión logística ha sido reportado en algunos estudios previos: Narayanan y Swaminathan (1996) enunciaron que la tasa de error tipo I mayor se encuentra en ítems con discriminación alta y dificultad media (8.8%) o dificultad baja (9.8%)[6]. Herrera (2005) sugiere la existencia de un efecto de la razón de tamaños sobre el error tipo I de la prueba conjunta de DIF para ítems con baja dificultad y discriminación entre 0 y 1. Aguerri et al. (2007) mostraron que cuando el grupo focal es de n= 350, los ítems son de baja dificultad y alta discriminación (*b*=-1, *a*=1.2), así como aquellos ítems cuya dificultad toma valores extremos y la discriminación posee valores muy altos (*b*= -2 o 2, *a*=1.6), poseen tasas de FP > 0.10. Arias (2008) señaló en su estudio sobre el MH que las tasas de error tipo I para los ítems 10 y 29 fueron las más altas (tasas de detección de 0.50 y 0.59). Como característica principal de estos ítems es que poseen una discriminación mayor a 1.40 y dificultad que oscila entre -1 y -1.55 (ítems fáciles). Finch y French (2007) y Akelo (2008) encontraron que los ítems con alta discriminación y baja dificultad tenían más probabilidad de ser clasificados incorrectamente como DIF. Akelo afirmó además que el SIBTEST era más sensible a las diferencias entre

[6] En este estudio, la clasificación de los parámetros tomó estos valores: Dificultad baja *b*=-1.5; media *b*=0; y alta *b*= 1.5. Discriminación baja (0.40 ≤*a* ≤ 0.50 en Gr, y 0.72 ≤ *a* ≤1.03 para Gf), y Discriminación alta (0.47 ≤ *a* ≤ 0.90 en Gr, y 1.68 ≤ *a* ≤ 2.01).

dificultad y discriminación. En relación con la estimación de los parámetros de los ítems, Atar (2007) indicó que "la razón de tamaños afecta la precisión de la estimación de los parámetros de dificultad y de discriminación en las pruebas dicótomas" (p. 10). Esta última hipótesis que vincula razón de tamaños y parámetros de los ítems, merece ser analizada a profundidad en posteriores investigaciones sobre DIF.

La influencia de los parámetros de los ítems puede evidenciarse también en las tasas de error tipo I vía el modelo de simulación. Cuando los datos se simulan y estiman con modelo de un parámetro, la tasa de error tipo I es mayor en comparación al modelo de tres parámetros. Esta variable es significativa particularmente para la prueba de DIF no uniforme, ya que el modelo de 3 parámetros, al tener en cuenta los parámetros de dificultad y discriminación, permite que se consideren las estimaciones de los parámetros a y b, de manera que ambas tendrían peso dentro del modelo de regresión, disminuyendo con ello la tasa de error tipo I. Partiendo de lo anterior, tasas altas de error tipo I se encuentran en ítems con discriminación media y modelo de 1 parámetro, ya que sólo se consideraría la habilidad y la dificultad del ítem como elementos determinantes, a expensas del grado de discriminación, además, si se observa la tabla 8, los ítems con tasas de falsos positivos más altas y que presentan discriminación media se acompañan de dificultad media; dificultad que podría enmascarar de cierta forma el efecto que pueda ejercer la discriminación.

Akelo (2008) reportó una tendencia similar en el empleo del SIBTEST al indicar que: "El procedimiento SIBTEST también identifica erróneamente ítems como DIF aquellos que poseen parámetros de dificultad y discriminación que son semejantes entre sí para DIF, por ello, se observa una elevada tasa de error tipo con α = 0.05" (p. 64). Cuando el modelo es de 3 parámetros y se conjuga con discriminación media o baja, las tasas de falsos positivos disminuyen considerablemente.

En relación con la potencia, *la razón de 20:1,* y *discriminación media y alta* reporta una tasa de detecciones correctas mayor a 80%. Con respecto a esto, los resultados de estudios previos muestran que hay una mayor potencia de la RL en

ítems con dificultad media y discriminación alta (82% para DIF uniforme y 77% para DIF no uniforme) (Rogers & Swaminathan, 1993) [7] . Narayanan & Swaminathan (1996) en su estudio reportan que en cuanto al tipo de ítem, la mayor potencia se encuentra en ítems con *baja b alta a* (90%) y *media b alta a* (70%). En relación con los resultados obtenidos en la presente investigación se encuentra que las mayores tasas de detección de ítems con DIF uniforme se dan efectivamente en aquellos ítems que poseen dificultad media y discriminación alta, mientras que en el caso del DIF no uniforme las mayores tasas de detección de ítems se encuentra en el ítem 9, que posee dificultad media y discriminación media.

La discusión de los hallazgos obtenidos en el presente estudio en cuanto a la potencia y error tipo I de la regresión logística en la detección de DIF permite dilucidar algunas tendencias generales sobre la razón de tamaños y el efecto de las demás variables independientes sobre la potencia y error tipo I.

1. El aumento de la razón de tamaños se asocia con decrementos en la potencia y el error tipo I. Aún cuando la razón de 100:1 presenta una tasa de detecciones correctas aceptable (74%), su error tipo I es superior al 20%, presentando una alta inflación del mismo. Autores como Zumbo (1999), Jodoin & Gierl (2001) y French y Maller (2007), entre otros, han señalado que la regresión logística reporta altas tasas de error tipo I, particularmente cuando se realiza la detección de DIF con la prueba conjunta de DIF, sin tener en cuenta una medida del tamaño del efecto.

2. La condición de diferencias en la distribución de las medias entre los grupos se asoció con incrementos en el error tipo I y la potencia del estadístico. Así mismo las interacciones con razón de tamaño revisten gran interés teórico y práctico, ya que si los grupos a ser evaluados presentan una gran diferencia en proporción numérica y además se distribuyen diferencialmente en cuanto a la

[7] En este estudio, la clasificación de los parámetros tomó estos valores: Dificultad baja *b*=-1.5; media *b*=0; y alta *b*= 1.5. Discriminación baja= 0.6, Discriminación Media=1 y Discriminación alta = 1.6

media de magnitud de atributo, la probabilidad que los ítems sean detectados como DIF es mayor, aun cuando las diferencias en magnitud de atributo de los individuos son reales.

3. En cuanto a la tasa global de detección de DIF uniforme y no uniforme mediante RL, se observa que la detección de DIF no uniforme no alcanza valores mayores a 0.70, ahí pueden entrar en consideración variables como el tamaño de muestra total, las razones de tamaño, el número de ítems simulados como no uniforme, la definición de los modelos de simulación, variables que pueden interactuar para la obtención de tales tasas de detección de ítems DIF.

Ante este panorama, puede concluirse que la regresión logística no es el mejor método de detección de DIF en razones de tamaño extremas, ya que presenta altas tasas de error tipo I y potencia inferior a los valores mínimos aceptables de identificación de ítems DIF, visto esto particularmente en las situaciones donde se simulan contextos de aplicación masiva de pruebas en un ámbito educativo (i.e., Examen ICFES), escenarios en donde el tamaño total de los individuos que se presentan semestralmente es amplio, $n \geq 500000$ personas, y las diferencias entre grupos mayoritarios y minoritarios son grandes; ya que, como hemos visto, el aumento del tamaño de muestra comporta un aumento de la potencia del estadístico, pero así mismo un incremento en la tasa de error tipo I; de la misma manera, razones de tamaño pequeñas, donde haya un mayor número de individuos en el grupo focal conducen a aumentos en la potencia y en el error tipo I. Si se revisan las pruebas de detección de DIF, se encuentra además que las tasas de detección son mayores para la prueba conjunta de DIF y la prueba de DIF uniforme con respecto a la prueba de DIF no uniforme, aunque los errores tipo I son también altos para estas condiciones.

Algunas limitaciones del presente estudio se encontraron en la no manipulación sistemática de los niveles de dificultad y discriminación como variable independiente que pueda aportar en la comprensión de la potencia y el error tipo I. Aun cuando se encontraron tendencias interesantes en cuanto a potencia y error tipo I para los niveles de dificultad y discriminación, que son concordantes con hallazgos previos, es preciso una investigación más exhaustiva

sobre el efecto de estos niveles de los parámetros en conjunción con razones de tamaños grandes y amplios tamaños de muestra. Otro aspecto que no se definió adecuadamente fue la condición de modelo de 1 parámetro, ya que ésta, al considerar sólo la dificultad, excluía el papel de la discriminación; además, en los estudios sobre regresión logística el ajuste del modelo se ha definido tradicionalmente en términos del modelo de dos parámetros.

La observación de los resultados obtenidos y de la literatura concerniente al empleo de la regresión logística en la detección de DIF abre la posibilidad de sugerir propuestas de investigación tendientes a comprender el funcionamiento de la regresión logística en la identificación de DIF, especialmente en escenarios de aplicación masiva de pruebas, donde se cuenta con un gran número de participantes y se presentan diferencias marcadas en cuanto a la composición de los grupos. Entre los objetos materia de posteriores investigaciones cabe señalar los siguientes:

1. Complementar las pruebas estadísticas de detección de DIF con el empleo de medidas de tamaño del efecto, a fin de controlar las tasas de error tipo I. Algunos autores como Zumbo (1999) y Jodoin & Gierl (2001), han propuesto como medidas de tamaño del efecto el $R^2\Delta$, bajo tres versiones: $R^2\Delta$, que a semejanza de la prueba conjunta de DIF, se basa en la comparación del modelo que posee sólo la habilidad, y el modelo completo, donde además de la habilidad se incorporan el grupo y la interacción grupo * habilidad.; $R^2\Delta U$ y $R^2\Delta NU$ para la extracción de una medida del tamaño del efecto en la prueba de DIF uniforme y la prueba de DIF no uniforme, respectivamente. Estos autores han propuesto diferentes criterios de clasificación de DIF; Zumbo clasificó el DIF como leve= $R^2\Delta$ ≤ 0.13, moderado = 0.13 ≤ $R^2\Delta$ ≤ 0.26, amplio = $R^2\Delta$ ≥ 0.26; mientras que Jodoin y Gierl clasificaron el DIF en leve= $R^2\Delta$ ≤ 0.035, moderado = 0.035 ≤ $R^2\Delta$ ≤ 0.070, amplio = $R^2\Delta$ ≥ 0.070. Los resultados que se han obtenido con el empleo de estas medidas han sido contradictorios en cuanto a la efectividad de estas medidas (Hidalgo & López-Pina, 2004; Kanjee, 2007), y sugieren la necesidad de redefinir

los puntos de corte a fin de controlar la tasa de error tipo I y lograr niveles adecuados de potencia (Gómez-Benito, Hidalgo & Padilla, 2009). Es por ello que una de las primeras tareas a realizar por parte de los investigadores que aborden el estudio de la regresión logística y el empleo de medidas de tamaño de efecto en DIF ha de ser el establecimiento de una métrica adecuada de clasificación de ítems con DIF. Otros estadísticos como el MH y la diferencia de dificultad poseen una medida de tamaño del efecto asociada, que al ser utilizada en conjunción con la prueba estadística, proporciona adecuadas tasas de potencia y un mayor control de error tipo I (Arias, 2008; Berrío, 2008).

2. Examinar el efecto de la razón de tamaños extremas en la detección de DIF mediante el empleo de la regresión logística, a partir del uso de datos provenientes de otras pruebas aplicadas en el contexto educativo, como es el caso de las pruebas SABER o ECAES, a fin de colectar mayor evidencia que proporcione información sobre el efecto de la razón de tamaños en la regresión logística, en pruebas simuladas con ciertas características no sólo en términos de composición de los grupos focal y de referencia, sino teniendo en cuenta otras variables que se han probado pueden influir en la potencia y error tipo I de la RL como la inclusión de modelo de 2 parámetros como modelo ajustado y manipulación de los niveles de dificultad y discriminación del ítem.

3. Examinar más a profundidad el efecto que puede tener la razón de tamaños en la clasificación incorrecta de ítems DIF; lo aquí expuesto es una aproximación preliminar a la influencia de la razón de tamaños y de los parámetros de los ítems en la presentación de este tipo de error. Sería de gran interés conocer de una manera más sistemática qué razones de tamaño generan inflación o control de este error tipo I, y asimismo, determinar qué características de los ítems son decisivas en la clasificación errónea de ítems DIF.

4. Analizar otras posibles fuentes de funcionamiento diferencial de los ítems en los Exámenes Nacionales: Género, Establecimiento Educativo.

5. Capacitar a los equipos de las evaluaciones nacionales en el uso e interpretación de los resultados de las pruebas, así como las implicaciones técnicas y prácticas de las mismas.

REFERENCIAS

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

Active State Software Inc., (2008). *ActivePerl 5.10.0 standard distribution* [Computer software]. Vancouver, Canada. Retrieved from http://www.activestate.com/Products/activeperl/

Aguerri, M.T., Galibert, M. S., Attorresi, H. F., & Prieto, P. (2007). Erroneous detection of nonuniform DIF using the Breslow-Day test in a short test. *Quality and Quantity* (version online).

Akelo, R. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-based Monte Carlo study with SIBTEST and Mantel-Haenszel procedures.* Tesis de Doctorado en Filosofía en Investigación Educativa y Evaluación. Virginia Polytechnic Institute and State University.

Alderete, A. M. (2006). Fundamentos del análisis de regresión logística en la investigación psicológica. *Evaluar*, *6,* 52-67.

Angoff, W.H. (1972, September). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 069686)

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.

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

Arias, E. (2008). *Efecto de la razón de tamaño y el ajuste del modelo sobre el estadístico Mantel_Haenszel y su métrica delta en la detección de DIF.* Tesis de Maestría para optar por el título de Magíster de Psicología. Departamento de Psicología, Universidad Nacional de Colombia, Bogotá Colombia.

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression and GLLAMM procedures.* Tesis de Doctorado en Filosofía. College of Education. The Florida State University.

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

Berrío, A. I. (2008). *La razón de tamaños y desajustes al modelo en la detección de ítems con funcionamiento diferencial mediante el procedimiento diferencia de la dificultad*. Tesis de Maestría para optar por el título de Magíster en Psicología, Departamento de Psicología, Universidad Nacional de Colombia, Bogotá, Colombia.

Birnbaum, A. (1968). In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, Mass: Adisson-Wesley.

Bolt, D. M., & Gierl, M. J. (2004, April). *Application of a regression correction to three nonparametric test of DIF: Implications for global and local DIF detection*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

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

Cervantes (2007). *Simulación de matrices de probabilidad y simulación de bases de datos* [Script para Perl]. Proyecto de identificación de sesgo cultural en el Examen de Estado ICFES. Grupo de Métodos e Instrumentos de Investigación en Salud. Universidad Nacional de Colombia.

Cervantes (2008a). *Muestreo de matrices de respuesta estudio Regresión Logística* [Script para Perl]. Proyecto de identificación de sesgo cultural en el Examen de Estado ICFES. Grupo de Métodos e Instrumentos de Investigación en Salud. Universidad Nacional de Colombia.

Cervantes (2008b). *Detección de DIF-Dos etapas. Purificación de medida de equiparación y reporte de ítems que presentan DIF. Procedimiento de R-L. Uso de Pseudos R^2 como medidas de tamaño del efecto.* [Script para R]. Proyecto de identificación de sesgo cultural en el Examen de Estado ICFES. Grupo de Métodos e Instrumentos de Investigación en Salud. Universidad Nacional de Colombia.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Cohen, J. (1994). The earth is round (p< .05). *American Psychologist*, *49*, 997-1003.

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

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Thayer (Eds.), *Differential item functioning* (pp. 25-30), Hillsdale, NJ: Lawrence Erlbaum Associates.

Cornfield, J., Gordon, T. & Smith, W. N. (1961). Quantal response curves for experimentally uncontrolled variables. *Bulletin of The International Statistical Institute, 38*, 97-115.

Cromwell, S. (2006). *Improving the prediction of differential item functioning: A comparison of the use of an effect size for logistic regression DIF and Mantel-Haenszel DIF methods.* Tesis de Doctorado en Filosofía. Texas A & M University. Estados Unidos.

Departamento Administrativo Nacional de Estadística. (2007). *Colombia: Una nación multicultural. Su diversidad étnica*. Bogotá: Departamento Administrativo Nacional de Estadística.

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

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

Fidalgo, A. M., Mellenbergh, G. J. & Muñiz, J. (2000). Effects of amount of DIF, test length and purification typo on rubustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*(3), 43-53.

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

Finch, W. H., & French, B. F. (2008). Anomalous Type I Error Rates for Identifying One Type of Differential Item Functioning in the Presence of the Other. *Educational and Psychological Measurement, 68*(5), 742-759.

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

Garson, G. D. (1998). Logistic regression. Obtenido el 22 de noviembre de 2007 del sitio http://www2.chass.ncsu.edu/garson/PA765/logistic.htm.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, 20, 26-36.

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

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

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research, 41*(1), 29-53.

Hadley, P. (1995). The performance of the Mantel-Haenszel and logistic regression dif detection procedures across sample size and effect size. Tesis Doctoral. University of Ottawa.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *MMSS fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*(3), 244-253.

Harrell, F. (2001). *Regression modelling strategies with applications to linear models, logistic regression, and survival analysis.* United States: Springer New York.

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

Herrera, A. N., Sánchez, N. R. & Jiménez, H. (2001). De la teoría clásica de los tests a la teoría de respuesta al ítem. *Aula Psicológica, 3,* 293-332.

Herrera, A. N. (2005). *Efecto del tamaño de muestra y la razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems*. Tesis de Doctorado en Evaluación y Tecnología Informática en Ciencias del Comportamiento. Universidad de Barcelona, España.

Herrera, A. N. & Gómez, J. (2007). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality & Quantity, 42*(6), 739-755.

Herrera, A. N., Gómez, J. & Hidalgo, M. D. (2005). Detección de sesgo en los ítems mediante el análisis de tablas de contingencia. *Avances en Medición, 3*(1), 29-52.

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

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

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.

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). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hosmer, D. H. & Lemeshow, S. (1989). *Applied Logistic Regression.* United States of America: John Wiley & Sons, Inc.

ICI Noticias ICI Febrero 1985 *E-08. Breves notas sobre el modelo de Rasch*. Disponible en: http://www.ieesa-kalt.com/E08-1.JPG Recuperado el 29 de abril de 2009.

Ihaka & Gentleman. (2007). R Software V.2.6. [Software libre]. Disponible en: http://www.r-proyect.org.

Instituto Colombiano para el Fomento de la Educación Superior. (2006). *Propósitos del Examen de Estado*. Bogotá.

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

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 regression procedure for DIF detection*. (Paper presented at the Annual Meeting of the National Council on Measurement in Education) Seattle.

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

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 2,* 101-118.

Kennedy, M. (1994). *The influence of sample size, effect size, and percentage of DIF items on the performance of the Mantel-Haenszel and logistic regression DIF* identification procedures. Tesis Doctoral. University of Ottawa.

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

Lemonte, A. J., & Vanegas, L. H. (2005). Una comparación entre la inferencia basada en las estadísticas de Wald y razón de verosimilitud en los modelos logia y probit vía Monte Carlo. *Revista Colombiana de Estadística, 28*(1), 77-96.

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

Mahadevan, S. (2005). Logistic regression and generalized linear models. University of Massachusetts [recuperado el 03 de abril de 2009]

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

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician. 54* (1), 17-24.

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

Peng, C.J., So, T. H., Stage, F. K. & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in Higher Education. 43*(3), 259-293.

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

R: a language and environment for statistical computing [Computer program]. Version 2.8.1. Vienna (Austria): R Development Core Team; 2009. Available from: URL: http://www.r-project.org

Raju, N. (1988). The area between two item characteristic curves. *Psychometrika. 53*(4), 495-502.

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.

Silva, L. C. & Barroso, I. M. (2004). *Regresión Logística* (Cuadernos de Estadística No. 27). Madrid: La Muralla

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.

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

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. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educational Researcher*, *25*(3), 26-30.

Tian, F. (1999). *Detecting DIF in polytomous item responses.* Tesis de Doctorado en Educación. University of Ottawa.

Traub, J. M. (1983). A priori considerations in choosing an item response model. *Journal of Educational Measurement, 13,* 28-34.

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

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2002). *BILOG MG3 for Windows*. Lincolnwood, USA.

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*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.

ANEXOS

Anexo 1. Script de muestreo de las 36 condiciones experimentales para la obtención de muestras de n = 65000

```
#!/usr/bin/perl

##############################################################################
# #
# # Métodos e Instrumentos para Investigación en Salud - MIIS
# # Proyecto Sesgo cultural en el examen de estado ICFES
# #
# # Muestreo de matrices de respuesta estudio Regresión Logística
# #
# # Script para perl
# #
# # Victor H Cervantes
# # vhcervantesb@unal.edu.co
# #
# # Last revision: Marzo 3 de 2008
# #
##############################################################################

use File::Find;

$USAGE = "Usage:\n\tCall with:\n\t\t\$this_script data_directory\n" ;

if ($#ARGV < 0) {die($USAGE)};
if ($#ARGV > 0) {die($USAGE)};


$dir = $ARGV[0];

find(\&edits, $dir);


# From Knuth Art of Programming
# Algortihm S(3.4.2)
# Select n records at random from a set of N records where
# 0<n<=N
#
# This algorithm is only useful when N is known in advance.
# If n=2 then the average number of elements considered is
# 2/3*N. the General formula is (N+1)n/(n+1)
# Its possible to optimise this even more.
#
# In this case we use $array a reference to an array of items
# and $num for the number of elements we want, we return an
# array of elements
#
# It should be remembered that this will be as random as
# the random number generator being used.


sub selection_sample {
   my ($array,$num)=@_;
```

```
    die "Too few elements (".scalar(@$array).") to select $num from\n"
      unless $num<@$array;
    my @result;
    my $pos=0;
    while (@result<$num) {
      $pos++ while (rand(@$array-$pos)>($num-@result));
      push @result,$array->[$pos++];
    }
    return @result
}

sub edits() {

        if ($_ =~ m/^d\d{4}r\d+\.dat$/) {

                print "File name is $_\n\t\tFull path is $File::Find::name\n";

                $_ =~ m/(^d)(\d)(\d{3})(r\d+\.dat$)/ ;

                $razon = $2;

                if ($razon == 1 | $razon == 2) {
                        print "The file comes from a condition with sample size ratio of 1000:1 or 500:1\n" ;
                        print "This conditions won't be used in the Logistic Regression study\n\n" ;
                        return ;
                } if ($razon == 3) {

                        $ref_size = 64741;
                        $foc_size = 259;

                        $ref_length = 129482;
                        $foc_length = 518;

                        print "This file comes from a condition with sample size ratio of 250:1\n\n" ;

                } if ($razon == 4) {

                        $ref_size = 64356;
                        $foc_size = 644;

                        $ref_length = 128713;
                        $foc_length = 1287;

                        print "This file comes from a condition with sample size ratio of 100:1\n\n" ;

                } if ($razon == 5) {

                        $ref_size = 61905;
                        $foc_size = 3095;

                        $ref_length = 123810;
                        $foc_length = 6190;

                        print "This file comes from a condition with sample size ratio of 20:1\n\n" ;

                } if ($razon < 1 | $razon > 5) {
```

```
                print "Unknown sample size ratio\n\n" ;
                return ;

        }

        $file =  $File::Find::name ;
        $outfile = $file ;
        $outfile =~ s/d(\d{4}r\d+\.dat$)/rl\1/ ;

        print "Output file in \n\t$outfile\n\n";

        open(FILE, $file) ;

        @lines = <FILE>;
        @ref_file = (0 .. ($ref_length - 1));
        @foc_file = ($ref_length .. ($ref_length + $foc_length - 1));


        print "Sampling data\n\n";

        print "Sampling ".$ref_size." subjects from reference group\n\n" ;
        @ref_sample = selection_sample(\@ref_file, $ref_size) ;


        print "Sampling ".$foc_size." subjects from focal group\n\n" ;
        @foc_sample = selection_sample(\@foc_file, $foc_size) ;


        print "Saving file $outfile\n\n\n";

        open(NEWFILE, ">$outfile") ;
        open(TEMP, ">temp.dat");

        print TEMP "@lines[@ref_sample]";
        print TEMP "@lines[@foc_sample]";

        close(FILE);
        close(TEMP);
        open(TEMP, "temp.dat");

        while(<TEMP>) {
                $_ =~ s/^ //;
                print NEWFILE $_ ;
        }

        close(TEMP);
        close(NEWFILE);
    }

}

print("# # The end\n");
```

Anexo 2. Script procedimiento de regresión logística con purificación bietápica para la detección de DIF

```
#############################################################################
    # #
    # # Metodos e Instrumentos para Investigación en Salud - MIIS
    # # Proyecto Sesgo cultural en el examen de estado ICFES
    # #
    # # Detección de DIF - Dos etapas
    # # Purificación de medida de equiparación y reporte de ítems que presentan DIF
    # # Procedimiento de R-L
    # # Uso de pseudo-R2 como medidas de tamaño del efecto
    # #
    # # Script para R
    # #
    # # Victor H Cervantes
    # # vhcervantesb@unal.edu.co
    # #
    # # Last revision: Feb 10, 2009
    # #
    ############################################################################

    ############################################################################
    # # Funciones generales para el procedimiento
    ############################################################################

    # # Crear variable de equiparación basada en el func_puntaje en la prueba
    func_puntaje <- function(base, ...) {
          omit <- c(...)
          if (length(omit) > 0) {
                apply(base[,-omit], 1, sum, na.rm=TRUE)
          } else {
                apply(base, 1, sum, na.rm=TRUE)
          }
    }

    # # Standardizing function
    func_z <- function(variable){
          ( (variable - mean(variable, na.rm=TRUE)) / sd(variable, na.rm=TRUE) )
    }


    ############################################################################
    # # Funciones para obtener las regresiones logísticas de los 3 modelos anidados
    ############################################################################

    # # Modelos de regresión logística
    func_RL_crit <- function(resp_var, crit, use.glm = TRUE, ...) {
      if (use.glm == TRUE) {
          rl <- glm(resp_var ~ func_z(crit), binomial, ...)
      } else {
        require(Design)
        rl <- lrm(resp_var ~ func_z(crit), ...)
        rl$y            <- resp_var
        rl$null.deviance <- rl$deviance[1]
        rl$deviance      <- rl$deviance[2]
```

```
   rl$fitted.values <- 1 / (1 + exp(-predict(rl)))
   rl$df.residual   <- rl$stats[["d.f."]]
  }
rl
}

func_RL_no_inter <- function(resp_var, crit, group, use.glm = TRUE, ...) {
  if (use.glm == TRUE) {
      rl <- glm(resp_var ~ func_z(crit) + func_z(group), binomial, ...)
  } else {
   require(Design)
   rl <- lrm(resp_var ~ func_z(crit) + func_z(group), ...)
   rl$y            <- resp_var
   rl$null.deviance <- rl$deviance[1]
   rl$deviance      <- rl$deviance[2]
   rl$fitted.values <- 1 / (1 + exp(-predict(rl)))
   rl$df.residual   <- rl$stats[["d.f."]]
  }
rl
}

func_RL_inter <- function(resp_var, crit, group, use.glm = TRUE, ...) {
  if (use.glm == TRUE) {
      rl <- glm(resp_var ~ func_z(crit) * func_z(group), binomial, ...)
  } else {
   require(Design)
   rl <- lrm(resp_var ~ func_z(crit) * func_z(group), ...)
   rl$y            <- resp_var
   rl$null.deviance <- rl$deviance[1]
   rl$deviance      <- rl$deviance[2]
   rl$fitted.values <- 1 / (1 + exp(-predict(rl)))
   rl$df.residual   <- rl$stats[["d.f."]]
  }
rl
}


#############################################################################
# # Funciones para calcular los diferentes R2
#############################################################################

func_R_Nagelkerke <- function(rl) {
  dev.m <- rl$deviance
  dev.n <- rl$null.deviance
  n     <- length(rl$y)

  num   <- 1 - exp( (dev.m - dev.n)/n )
  den   <- 1 - exp( -dev.n / n )

return(num/den)
}

func_R_McFadden <- function(rl) {
  num <- rl$deviance
  den <- rl$null.deviance
```

```
1 - (num/den)
}

func_R_OLS <- function(rl) {
  num <- sum( (rl$y - rl$fitted.values)^2 )
  den <- sum( (rl$y - mean(rl$y))^2 )

1 - (num/den)
}

func_R_Delta_Jodoin <- function(rl, crit, group, tipo = "conjunto") {
  it       <- rl$y
  pre      <- rl$fitted.values
  zgroup   <- func_z(group)
  ztotal   <- func_z(crit)

  agregated <- cbind(aggregate(it, list(zgroup, ztotal), sum)[, -c(1:2)],
                aggregate(cbind(pre, zgroup, ztotal), list(zgroup, ztotal), mean)[, -c(1:2)],
                aggregate(it, list(zgroup, ztotal), length)[, -c(1:2)]
    )

  item     <- agregated[, 1]
  pre      <- agregated[, 2]
  zgroup   <- agregated[, 3]
  ztotal   <- agregated[, 4]
  n        <- agregated[, 5]

  v        <- n * pre * (1 - pre)
  z        <- log(pre /(1 - pre))+((item - n * pre)/v)

  R2_ztotal <- summary(lm(z ~ ztotal, weights = v))$r.squared
  R2_zgroup <- summary(lm(z ~ ztotal + zgroup, weights = v))$r.squared
  R2_zinter <- summary(lm(z ~ ztotal * zgroup, weights = v))$r.squared

  if (tipo == "conjunto") {
    R2     <- list(R2crit = R2_ztotal, R2inter = R2_zinter)
    Rdelta <- list(RdG = R2_zinter - R2_ztotal)
  } else {
    R2     <- list(R2crit = R2_ztotal, R2group = R2_zgroup, R2inter = R2_zinter)
    Rdelta <- list(RdN = R2_zinter - R2_zgroup, RdU = R2_zgroup - R2_ztotal)
  }

return(list(R2, Rdelta))
}


#############################################################################
# # Obtención de los modelos
#############################################################################

# # Calcular R-L sobre los ítems de la base
func_RL_DIF <- function(base, crit, group, tipo = "conjunto", use.glm = TRUE, ...) {

  if (tipo == "conjunto") {
    cbind(lapply(base, func_RL_crit, crit, use.glm, ...),
          lapply(base, func_RL_inter, crit, group, use.glm, ...))
```

```
  } else {
    cbind(lapply(base, func_RL_crit, crit, use.glm, ...),
          lapply(base, func_RL_no_inter, crit, group, use.glm, ...),
          lapply(base, func_RL_inter, crit, group, use.glm, ...))
  }
}

#############################################################################
# # Funciones para identificación de ítems con DIF según RL
#############################################################################

# # Prueba de hipótesis sobre los modelos (Chi2, Deviance)
modeltest <- function(coupled_data, tipo = "conjunto") {
# Input coupled data should be as a line resulting from
# func_RL_DIF
  if (tipo == "conjunto") {
      LL <- abs(coupled_data[[1]][["deviance"]] - coupled_data[[2]][["deviance"]])
        df <- abs(coupled_data[[1]][["df.residual"]] - coupled_data[[2]][["df.residual"]])
      c(G = LL, df = df, p.value = pchisq(LL, df, lower.tail = FALSE))
  } else {
    LL1 <- abs(coupled_data[[1]][["deviance"]] - coupled_data[[2]][["deviance"]])
    LL2 <- abs(coupled_data[[2]][["deviance"]] - coupled_data[[3]][["deviance"]])

    df1 <- abs(coupled_data[[1]][["df.residual"]] - coupled_data[[2]][["df.residual"]])
    df2 <- abs(coupled_data[[2]][["df.residual"]] - coupled_data[[3]][["df.residual"]])

    c(G1 = LL1, df1 = df1, p.value1 = pchisq(LL1, df1, lower.tail = FALSE),
      G2=LL2, df2 = df2, p.value2 = pchisq(LL2, df2, lower.tail = FALSE))
  }
}


# # Obtener las pruebas de Chi2 para todos los ítems
func_RL_test <- function(RL_DIF, tipo = "conjunto") {
   apply(RL_DIF, 1, modeltest, tipo)
}

# # Items para los cuales la prueba es significativa
func_RL_sig <- function(RL_test, alfa_level, tipo = "conjunto") {
  if (tipo == "conjunto") {
    identified <- row(as.matrix(RL_test["p.value", ]) )[RL_test["p.value", ] < alfa_level]
  } else {
    identified1 <- row(as.matrix(RL_test["p.value1", ]) )[RL_test["p.value1", ] < alfa_level]
    identified2 <- row(as.matrix(RL_test["p.value2", ]) )[RL_test["p.value2", ] < alfa_level]
    identified <- unique( c(identified1, identified2) )
  }
}

#############################################################################
# # Funciones para calcular el R-Delta-cuadrado
#############################################################################


# # Obtener Rdelta Jodoin (Jodoin & Gierl, 2001)

func_R_Delta_JG <- function(RL_DIF, crit, group, tipo = "conjunto") {
```

```
  deltas <- sapply(RL_DIF[, ncol(RL_DIF)], func_R_Delta_Jodoin, crit, group, tipo)

}


# # Pendientes las de los otros R2



# # Classify and report item list detected with effect size measure
func_is.BC_by_R2 <- function(Rdeltas, severity = "B", tipo = "conjunto") {
  if (severity == "B") {
    thereshold = 0.035
  } else if (severity == "C") {
    thereshold = 0.07
  }

  if (tipo == "completo") {
    is.BC <- (unlist(Rdeltas[2,]) >= thereshold)
    row(as.matrix(is.BC))[is.BC > 0]
  } else {
  }
}


#############################################################################
# # Funciones para ejecutar las dos fases en la identificacion
# # basadas en el puntaje en la prueba
# # detectados tanto al nivel de 0.05 como 0.01 de significancia
# # empleando y sin emplear una medida del tama\~no del efecto (R2delta)
#############################################################################

# # Obtener los ítems detectados con DIF en la primera fase
func_RL_fase_1 <- function(base, group, alfa, tipo = "conjunto", use.glm = TRUE, pseudo.R = "Jodoin") {

  Rdelta  <- switch(pseudo.R, Jodoin = func_R_Delta_JG, Nagelkerke = func_R_Delta_Nagelkerke,
                McFadden = func_R_Delta_McFadden, OLS = func_R_Delta_OLS)

      RL_DIF  <- func_RL_DIF(base, func_puntaje(base), group, tipo, use.glm)
      tests   <- func_RL_test(RL_DIF, tipo)
      sig     <- func_RL_sig(tests, alfa, tipo)

  if (pseudo.R == "None") {
    salida  <- list(sig = sig)
  } else if (pseudo.R == "Jodoin") {
    Rdeltas <- Rdelta(RL_DIF, func_puntaje(base), group, tipo)
    r2      <- func_is.BC_by_R2(Rdeltas)
    salida  <- list(sig = sig, r2delta = r2)
  } else {
    Rdeltas <- Rdelta(RL_DIF)
  }

return(salida)
}
```

```
    # # Obtener los ítems detectados con DIF en la segunda fase
    # # junto con sus características

    func_RL_fase_2 <- function(base, group, fase_1, tipo = "conjunto", use.glm = TRUE, use.R = TRUE, pseudo.R
= "Jodoin") {

      if (use.R == TRUE) {
        items_fase_1 <- c(fase_1$sig, fase_1$r2delta)[duplicated( c(fase_1$sig, fase_1$r2delta) )]
      } else {
        items_fase_1 <- fase_1$sig
      }

      Rdelta  <- switch(pseudo.R, Jodoin = func_R_Delta_JG, Nagelkerke = func_R_Delta_Nagelkerke,
                  McFadden = func_R_Delta_McFadden, OLS = func_R_Delta_OLS)


      RL_DIF_2nd  <- func_RL_DIF(base, func_puntaje(base, items_fase_1), group, tipo, use.glm)
      tests_2nd   <- func_RL_test(RL_DIF_2nd, tipo)


      if (pseudo.R == "Jodoin") {
        Rdeltas_2nd <- Rdelta(RL_DIF_2nd, func_puntaje(base), group, tipo)
    #    r2          <- func_is.BC_by_R2(Rdeltas_2nd)

      } else {
        Rdeltas_2nd <- Rdelta(RL_DIF)
      }

      if (tipo == "conjunto") {
        tests <- c(1, 3)

        if (pseudo.R != "None") {
          R.sq   <- as.data.frame(lapply(Rdeltas_2nd[1,],
                                  function(x) c(x[[1]], x[[2]])
                                  )
                          )
          Deltas <- unlist(Rdeltas_2nd[2,])

          colnames(R.sq)   <- colnames(tests_2nd)
          names(Deltas) <- colnames(tests_2nd)
        }
      } else {
        tests <- c(1, 3, 4, 6)
        if (pseudo.R != "None") {
          R.sq   <- as.data.frame(lapply(Rdeltas_2nd[1,],
                                  function(x) c(x[[1]], x[[2]], x[[3]])
                                  )
                          )
          Deltas <-  as.data.frame(lapply(Rdeltas_2nd[2,],
                                  function(x) c(x[[1]], x[[2]])
                                  )
                          )
          colnames(R.sq)   <- colnames(tests_2nd)
          colnames(Deltas) <- colnames(tests_2nd)
        }
      }
```

```
  coeff <- as.data.frame( lapply(RL_DIF_2nd[, ncol(RL_DIF_2nd)], function(x) x[["coefficients"]]))
  colnames(coeff) <- colnames(tests_2nd)

  RLDIF <- rbind(tests_2nd[tests,], coeff)

  if (pseudo.R != "None") {
    RLDIF <- rbind(RLDIF, R.sq, Deltas)
    if (length(tests) == 2) {
      rownames(RLDIF)[7:9] <- c("R1", "R3", "DRC")
    } else {
      rownames(RLDIF)[9:13] <- c("R1", "R2", "R3", "DRU", "DRN")
    }
  }

 return(RLDIF)
 }



 func_RL_ambas_fases <- function(base, group, tipo = "conjunto", use.glm = TRUE, use.R = TRUE, pseudo.R =
"Jodoin") {

      detected <- func_RL_fase_1(base, group, 0.01, tipo, use.glm, pseudo.R)
      func_RL_fase_2(base, group, detected, tipo, use.glm, use.R, pseudo.R)
 }
```

Anexo 3. Tasas promedio de error tipo I en 0%, 10% y 20% de DIF, por condición experimental (α = 0.01)

	Error tipo I 0% DIF			Error Tipo I 10% DIF			Error Tipo I 20% DIF		
Condición*	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}
R 20:1, SI, 1 P	0.010	0.004	0.005						
R 20:1, SI, 3 P	0.010	0.004	0.004						
R 20:1, SI, 1 P				0.009	0.002	0.005			
R 20:1, SI, 3 P				0.009	0.003	0.005			
R 20:1, SI, 1 P							0.010	0.005	0.004
R 20:1, SI, 3 P							0.009	0.004	0.004
R 20:1, DM, 1 P	0.938	0.799	0.875						
R 20:1, DM, 3 P	0.651	0.482	0.404						
R 20:1, DM, 1 P				0.947	0.817	0.883			
R 20:1, DM 3 P				0.664	0.472	0.373			
R 20:1, DM, 1 P							0.938	0.806	0.863
R 20:1, DM, 3 P							0.668	0.495	0.370
R 100:1, SI, 1 P	0.009	0.004	0.004						
R 100:1, SI, 3 P	0.011	0.003	0.006						
R 100:1, SI, 1 P				0.009	0.004	0.004			
R 100:1, SI, 3 P				0.010	0.003	0.005			
R 100:1, SI, 1 P							0.011	0.004	0.006
R 100:1, SI, 3 P							0.009	0.003	0.005
R 100:1, DM, 1 P	0.365	0.200	0.277						
R 100:1, DM, 3 P	0.222	0.155	0.084						
R 100:1, DM, 1 P				0.365	0.185	0.288			
R 100:1, DM, 3 P				0.156	0.087	0.083			
R 100:1, DM, 1 P							0.392	0.205	0.310
R 100:1, DM, 3 P							0.150	0.084	0.076
R 250:1, SI, 1 P	0.011	0.005	0.006						
R 250:1, SI, 3 P	0.011	0.004	0.005						
R 250:1, SI, 1 P				0.011	0.004	0.005			
R 250:1, SI, 3 P				0.010	0.004	0.005			
R 250:1, SI, 1 P							0.010	0.004	0.004
R 250:1, SI, 3 P							0.010	0.004	0.005
R 250:1, DM, 1 P	0.113	0.047	0.069						
R 250:1, DM, 3 P	0.113	0.080	0.034						
R 250:1, DM, 1 P				0.120	0.050	0.076			
R 250:1, DM, 3 P				0.075	0.043	0.032			
R 250:1, DM, 1 P							0.130	0.052	0.083
R 250:1, DM, 3 P							0.073	0.042	0.030

* CONVENCIONES: R 20:1 = Razón de Tamaño 20:1; R 100:1 = Razón de Tamaño 100:1; R 250:1= Razón de Tamaño 250:1; SI = Sin impacto; DM = Diferencia en la media; 1P = 1 parámetro; 3 P = 3 parámetros. Las tasas promedio resaltadas en negrita no cumplen el criterio liberal de Bradley (FP > 0.015).

Anexo 4. Valor F y significación de los efectos sobre el error tipo I de la RL en la condición de 0% de DIF (α = 0.01)

	χ^2_{2df}		χ^2U_{1df}		χ^2NU_{1df}	
Factor	F	Significación	F	Significación	F	Significación
Razón	110.521	.000	81.097	.000	126.444	.000
Impacto	404.863	.000	218.374	.000	323.785	.000
Modelo	13.613	.000	7.908	.005	54.070	.000
Razón * Impacto	111.006	.000	81.030	.000	126.970	.000
Razón * Modelo	4.560	.011	7.301	.001	16.179	.000
Impacto * Modelo	13.757	.000	7.787	.006	54.070	.000

Anexo 5. Valor F y significación de los efectos sobre el error tipo I de la RL en la condición de 10% y 20% de DIF (α = 0.01)

	χ^2_{2df}		χ^2U_{1df}		χ^2NU_{1df}	
Factor	F	Significación	F	Significación	F	Significación
Razón	230.478	.000	183.506	.000	188.386	.000
Impacto	739.058	.000	397.093	.000	520.096	.000
Porcentaje de DIF	.025	.874	.044	.834	.001	.975
Modelo	43.545	.000	29.240	.000	106.031	.000
Razón * Impacto	232.177	.000	184.185	.000	188.880	.000
Razón * Porcentaje de DIF	.018	.982	.010	.990	.062	.940
Impacto * Porcentaje de DIF	.022	.882	.032	.858	.000	.983
Razón * Modelo	5.961	.003	11.705	.000	28.175	.000
Impacto * Modelo	43.220	.000	29.028	.000	106.082	.000
Porcentaje de DIF * Modelo	.044	.833	.001	.974	.018	.895

Anexo 6. Tasas promedio de detecciones correctas en 10% y 20% de DIF, por condición experimental (α = 0.01)

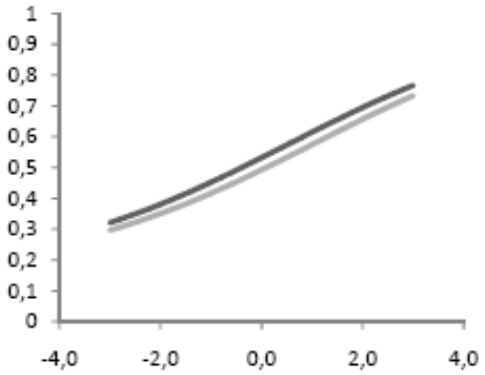
	Potencia DIF 10%			Potencia DIF 20%		
Condición [a]	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}	χ^2_{2df}	χ^2U_{1df}	χ^2NU_{1df}
R 20:1, SI, 1 P						
R 20:1, SI, 3 P						
R 20:1, SI, 1 P	0.671	1	0.012			
R 20:1, SI, 3 P	1	1	1			
R 20:1, SI, 1 P				0.673	1	0.007
R 20:1, SI, 3 P				0.977	0.914	0.982
R 20:1, DM, 1 P						
R 20:1, DM, 3 P						
R 20:1, DM, 1 P	0.997	0.994	0.93			
R 20:1, DM 3 P	0.951	0.926	1			
R 20:1, DM, 1 P				1	1	0.924
R 20:1, DM, 3 P				0.869	0.585	0.559
R 100:1, SI, 1 P						
R 100:1, SI, 3 P						
R 100:1, SI, 1 P	0.673	1	0.01			
R 100:1, SI, 3 P	0.836	0.756	0.918			
R 100:1, SI, 1 P				0.768	0.896	0.002
R 100:1, SI, 3 P				0.563	0.434	0.589
R 100:1, DM, 1 P						
R 100:1, DM, 3 P						
R 100:1, DM, 1 P	0.991	1	0.692			
R 100:1, DM, 3 P	0.738	0.608	0.926			
R 100:1, DM, 1 P				0.681	0.644	0.511
R 100:1, DM, 3 P				0.511	0.304	0.457
R 250:1, SI, 1 P						
R 250:1, SI, 3 P						
R 250:1, SI, 1 P	0.653	0.97	0.002			
R 250:1, SI, 3 P	0.617	0.594	0.43			
R 250:1, SI, 1 P				0.266	0.625	0.005
R 250:1, SI, 3 P				0.330	0.288	0.238
R 250:1, DM, 1 P						
R 250:1, DM, 3 P						
R 250:1, DM, 1 P	0.784	0.916	0.308			
R 250:1, DM, 3 P	0.589	0.543	0.484			
* R 250:1, DM, 1 P				0.433	0.501	0.168
R 250:1, DM, 3 P				0.349	0.273	0.264

CONVENCIONES: R 20:1 = Razón de Tamaño 20:1; R 100:1 = Razón de Tamaño 100:1; R 250:1= Razón de Tamaño 250:1; SI = Sin impacto; DM = Diferencia en la media; 1P = 1 parámetro; 3 P = 3 parámetros. Las tasas promedio resaltadas en negrita corresponden a potencia > 0.70.

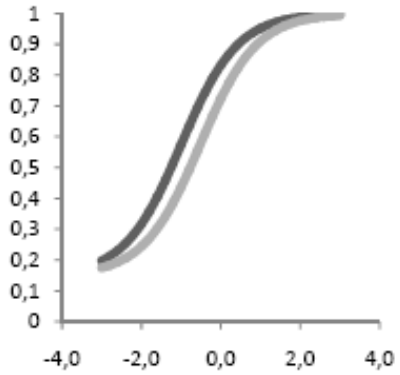
Anexo 7. Valor F y significación de los efectos sobre la potencia de la RL en la condición de 10% y 20% de DIF (α = 0.01)

	χ^2_{2df}		χ^2U_{1df}		χ^2NU_{1df}	
Factor	F	Significación	F	Significación	F	Significación
Razón	9.579	.000	8.563	.000	12.457	.000
Impacto	.979	.325	.574	.451	14.795	.000
Porcentaje de DIF	5.665	.020	6.348	.014	1.687	.198
Modelo	.088	.767	.479	.491	5.429	.022
Razón * Impacto	.169	.845	.071	.931	3.161	.047
Razón * Porcentaje de DIF	1.334	.269	.647	.526	.051	.950
Impacto * Porcentaje de DIF	.143	.706	.279	.599	.836	.363
Razón * Modelo	.914	.405	.634	.533	.173	.842
Impacto * Modelo	2.939	.090	1.318	.254	4.065	.047
Porcentaje de DIF * Modelo	.044	.834	.235	.629	.004	.950

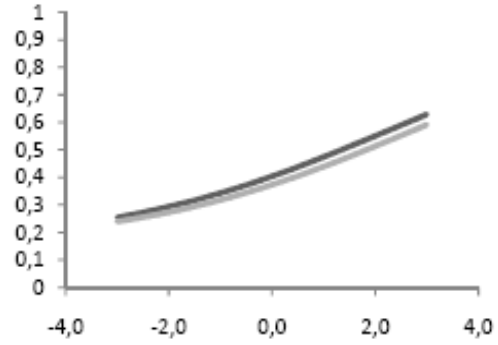
Anexo 8. CCI de los ítems con DIF uniforme y no uniforme



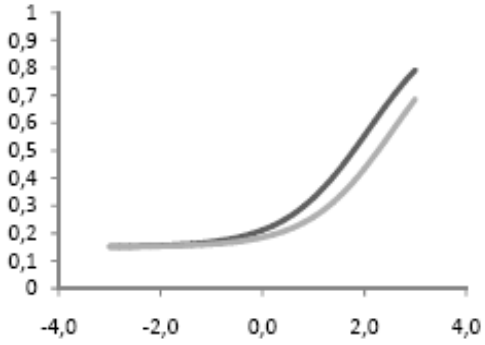
Item 3



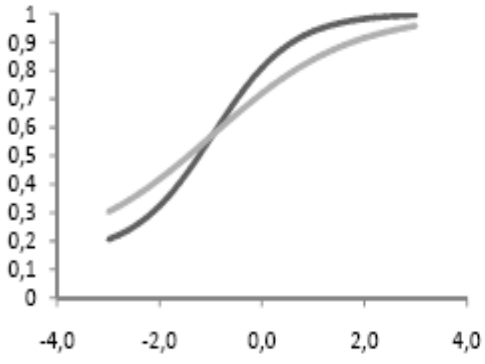
Item 10



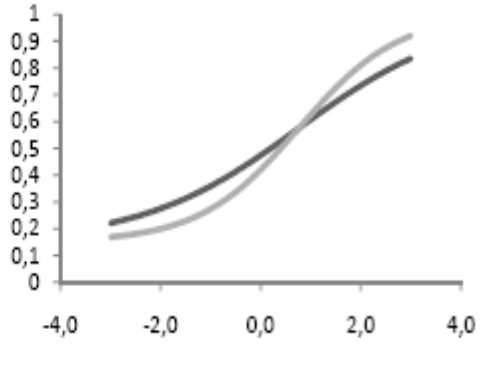
Item 21



Item 25



Item 9



Item 24