Interdisciplinaria
ISSN: 0325-8203
interdisciplinaria@fibercorp.com.ar
Centro Interamericano de Investigaciones Psicológicas y Ciencias Afines
Argentina

BOGOYA M., DANIEL; OCAÑA GÓMEZ, ADELINA; BARRAGÁN MORENO, SANDRA PATRICIA; CONTENTO RUBIO, RICARDO

FUNCIONAMIENTO DIFERENCIAL DE ITEMES: EXAMEN DE MATEMÁTICA-UNIVERSIDAD JORGE TADEO LOZANO

Interdisciplinaria, vol. 31, núm. 1, 2014, pp. 121-138
Centro Interamericano de Investigaciones Psicológicas y Ciencias Afines
Buenos Aires, Argentina

# FUNCIONAMIENTO DIFERENCIAL DE ITEMES: EXAMEN DE MATEMÁTICA - UNIVERSIDAD JORGE TADEO LOZANO*

# DIFFERENTIAL ITEM FUNCTIONING: MATH TEST UNIVERSIDAD JORGE TADEO LOZANO

DANIEL **BOGOYA M.****, ADELINA **OCAÑA GÓMEZ*****, SANDRA PATRICIA **BARRAGÁN MORENO****** Y RICARDO **CONTENTO RUBIO*******

*Trabajo realizado en el marco del proyecto de investigación *Elementos de evaluación en ciencias mediante la matemática y el lenguaje* (código 402- 08-11), financiado por la Universidad Jorge Tadeo Lozano.
**Ingeniero Químico y Magister en Ingeniería de Sistemas. Consultor Independiente. E-Mail: dbogoya@yahoo.com Carrera 4 # 22-61, Bloque 15, Oficina 201. Bogotá, Colombia.
***Licenciada en Matemática y Magister en Investigación y Docencia Universitaria. Profesora Asociada del área de Matemática del Departamento de Ciencias Básicas y miembro del Grupo de Investigación de Didáctica de las Ciencias en Evaluación de la Educación de la Universidad Jorge Tadeo Lozano (UJTL). E-Mail: adelina.ocana@utadeo.edu.co
****Magister en Ciencias Matemáticas. Profesora Asociada del área de Matemática del Departamento de Ciencias Básicas y miembro del Grupo de Investigación de Didáctica de las Ciencias en Evaluación de la Educación de la Universidad Jorge Tadeo Lozano (UJTL). E-Mail: sandra.barragan@utadeo.edu.co
*****Estadístico y Magister en Enseñanza de las Ciencias Exactas y Naturales. Profesor adscrito al Departamento de Ciencias Básicas y miembro del Grupo de Investigación de Didáctica de las Ciencias en Evaluación de la Educación de la Universidad Jorge Tadeo Lozano (UJTL). E-Mail: manuel.contento@utadeo.edu.co

## RESUMEN

La Universidad Jorge Tadeo Lozano aplica el Examen de Clasificación en Matemáticas Básicas, como *evaluación diagnóstica*, a los aspirantes y estudiantes provenientes de transferencias internas o externas, cuyo plan de estudios precise conocimientos básicos de Aritmética y Algebra Elemental. Dicho examen favorece el análisis de las condiciones académicas de los admitidos y permite a la Universidad, ofrecer opciones apropiadas para cada caso particular, al mismo tiempo que al evaluado le proporciona la posibilidad de reconocer su nivel de apropiación del conocimiento de los dominios conceptuales requeridos. Consecuentemente con el carácter decisorio del Examen de Clasificación de Matemáticas Básicas, se examinó si los ítemes utilizados presentan funcionamiento diferencial, esto es, se analizó si la diferencia de habilidades entre los evaluados podría deberse a las variables de contexto seleccionadas: sexo, edad, naturaleza jurídica del colegio de procedencia y facultad en la que el aspirante tramita su ingreso. Para ello, se procesaron 1.623 cadenas de respuestas para 61 ítemes, obtenidas en las pruebas comprendidas entre el tercer período lectivo de 2011 y el primero de 2012.

La metodología incluyó la implementación de tres técnicas: *Contraste del DIF* (diferencia entre los centros de dificultades),Contraste del DIF (diferencia entre los extremos más próximos para los intervalos de dificultad) y prueba estadística Mantel-Haenszel. La conjunción de estas técnicas permitió determinar un ítem con funcionamiento diferencial en categoría moderada a grande, para

la variable edad. Finalmente, para este ítem se exhiben sus parámetros estadísticos y su curva característica, estimados en la *calibración*.

*Palabras clave:* Funcionamiento diferencial de los ítemes; Calibración; Contraste del DIF; Calidad de ítemes; Evaluación diagnóstica.

## Abstract

The following article presents an application of *differential item functioning* (DIF), using results obtained from the qualifying test developed by the Jorge Tadeo Lozano University and taken by students to classify them at a level of mathematical knowledge and to define an academic route for them based on their cognitive status shown on the test. The analysis is part of a perspective to estimate the difficulty and others characteristics of items, and the skills and level of students through the use of the Rasch model of the one parameter item response theory (IRT) and the parameters of a sample of 1623 students taking a test composed of 61 items. The article analyzes both the statistical performance of the items in terms of the parameters of item-test correlation, misfits (infit and outfit), and discrimination, as well as the behavior of the set of items depending on the construct validity or dimensionality, reliability, internal consistency, and separation parameters. A method is then shown to examine which items display DIF, associated with the conditions of the students' origin and not of their academic ability, which could lead to bias in the results of the test.

The employed methods estimate the relative difficulty of each item, for students of similar ability but who belong to different groups, according to four variables studied: sex, age, intended major, and whether the high school of origin is public or private. The value of the difference in relative difficulty between the groups mentioned is associated with a level of DIF and recognizes whether the item in question has bias and which groups this bias is favoring. The difference in relative difficulty is graded in terms of severity according to three categories proposed by the Educational Testing Service: (1) moderate to large, if the difference in relative difficulty between groups (for students of similar ability) is greater than or equal to .64 *logits*, (2) small to moderate, if the difference is greater than or equal to .43 and less than .64 l*ogits*, and (3) not significant, if this difference is less than .43 *logits*.

In order to validate the detection of DIF, the calculations are performed using three techniques. Two are chosen from those available in the literature and the third one is a proposal by the authors of this article to consider the size of the error in the estimations of difficulty difference. The three techniques used are: (1) the measurement of the difference between the core values of the difficulty intervals, ignoring the value of the estimation error, (2) the difference between the nearest extremes of the difficulty intervals, taking into account the estimation error, and (3) the Mantel-Haenszel statistical test. Regarding databases formed for the analysis, two aspects were considered: (1) chains of responses corresponding to missing data or especially small groups, which would not have allowed an effective and reliable comparison, were omitted, and (2) random samples with uniform distribution were selected to create groups of the same size for each study variable. The analysis with the technique of difference between core values showed that two items (34 and 59) displayed DIF with moderate to large severity, regarding the age variable for item 34 and the intended major and high school of origin variables for item 59. The technique about difference between the nearest extremes confirmed DIF with moderate to large severity for item 34, with respect to the age variable. The Mantel-Haenszel test detected DIF with moderate to large severity for items 13, 20, 34, and 61 for the age variable, and for items 4, 30, 36, 43, and 59 for the major variable.

*Key words:* Differential Item Functioning; Calibrating; DIF contrast; Quality items; Diagnostic assessment.

## Introducción

Los modelos de la Teoría de Respuesta al Item (TRI) permiten realizar un análisis del desempeño estadístico de ítemes y de instru-

mentos en su conjunto, al igual que estimar la habilidad de los evaluados mediante una prueba en los dominios considerados (conceptuales y cognitivos). Estos modelos aportaron solución a problemas presentados en la Teoría Clásica de los Test (TCT), que no podía predecir cómo respondería un individuo a una prueba (conjunto de ítemes) a menos que esta hubiera sido aplicada a individuos similares. La TRI abrió caminos a dichos análisis en los cuales mostró que los puntajes obtenidos pueden independizarse de la prueba utilizada. Las curvas características de los ítemes (CCI) o funciones de respuesta ($P(\theta)$) permiten establecer la relación que existe entre la probabilidad ($P$) de que un evaluado responda correctamente un ítem y el nivel de habilidad ($\theta$) en que se ubica el individuo (Baker, 2001; Lord, 1980).

En el análisis del sesgo o Funcionamiento Diferencial de los Itemes (DIF) se comparan las respuestas a un ítem, dadas por individuos de diferentes grupos con un mismo nivel de habilidad. Se establece que un ítem exhibe funcionamiento diferencial cuando dichos individuos tienen diferente probabilidad de responder correctamente el ítem (Angoff, 1993; Attorresi, Galibert, Zanelli, Lozzia & Aguerri, 2003).

Los procedimientos estadísticos utilizados en el análisis del DIF se basan en la aplicación de los modelos de la TRI, los cuales fueron empleados en el estudio del instrumento del Examen Clasificatorio de Matemáticas Básicas que aplica la Universidad Jorge Tadeo Lozano, con el fin de detectar si los ítemes que componen dicho instrumento exhiben DIF.

El origen del estudio del sesgo en los ítemes, se remonta a los años 60, con Cardall y Coffman (1964) y posteriormente con Cleary y Hilton (1968), quienes por medio de métodos estadísticos pretendían identificar los factores que causaban las diferencias significativas entre los puntajes obtenidos en una prueba psicológica, aplicada en 1963 a grupos de blancos y negros (Angoff, 1993; Angoff & Ford, 1973).

El término *sesgo* estaba siendo utilizado simultáneamente en dos sentidos, uno social y otro estadístico[1], solo hasta la década de los 80, en la que el *Educational Testing Service* (ETS), entre otros, comenzó a usar el término *funcionamiento diferencial de los ítemes*, referido a las propiedades estadísticas que arrojaba un ítem aplicado en grupos distintos, una vez controladas las diferencias de habilidades de dichos grupos.

Entre los métodos para analizar el DIF, se cuentan aquellos que se derivan de la TCT, como el de Mantel-Haenszel, Regresión Logística y Estandarización; entre los métodos que se apoyan en la TRI se encuentran, áreas exactas con signos y áreas exactas sin signo y el estadístico *ji cuadrado* ($\chi^2$). El método de Mantel-Haenszel (Dorans & Holland, 1993) permite evaluar y describir cómo la relación entre variables se modifica por la presencia de una variable externa; la hipótesis nula a contrastar es la existencia de igualdad entre las proporciones de sujetos que responden en forma correcta o no el ítem en cada muestra y en cada nivel en que se ha dividido la puntuación. Este estadístico sigue una distribución *ji cuadrado* con un grado de libertad. El método de regresión logística (Swaminathan & Rogers, 1990) se emplea fundamentalmente en el estudio y detección del DIF no uniforme que existe cuando hay interacción entre el nivel de habilidad y los miembros del grupo, es decir, la diferencia en la probabilidad de responder correctamente, para dos grupos, no es la misma en todos los niveles de habilidad. En TRI, el DIF no uniforme está indicado por las curvas de regresión logística para dos grupos diferentes: si dichas curvas no son paralelas hay presencia de DIF no uniforme; si son paralelas pero no se superponen puede inferirse que hay DIF uniforme; y si se superponen puede concluirse que no hay DIF. El método de estandarización (Dorans &

[1] En lo social, referido a injusticia, parcialidad e inequidad contra los grupos minoritarios o menos favorecidos; en lo estadístico, asociado a la observación de un ítem que muestra propiedades estadísticas diferentes en grupos distintos que tienen la misma habilidad (Angoff, 1993).

Kulick, 1986; Dorans & Holland, 1993) establece un control para las diferencias en la habilidad de una subpoblación y en la calidad de un ítem. Usar la estandarización significa que las diferencias en una variable han sido controladas, lo cual permite efectuar comparaciones entre grupos y variables relacionadas. En el método propuesto por Raju (1990), áreas exactas con signos y áreas exactas sin signo, se utiliza como *índice* el área comprendida entre las curvas características de los ítemes de la población de referencia, definida como aquella respecto de la cual se explora la existencia de DIF y la población focal. Si el área observada es cero, se concluye que no hay DIF; a medida que aumenta el área, aumenta también el DIF. En el método de Lord (1980) se propone un estadístico para contrastar la hipótesis nula de la igualdad de los vectores que definen los parámetros de los ítemes en las poblaciones de referencia y focal, es decir, si las curvas características de las dos poblaciones para el mismo ítem son similares, se concluye que no hay funcionamiento diferencial en ese ítem.

La presencia de DIF puede derivarse de dos fuentes: defectos en los ítemes a los que personas de grupos distintos son sensiblemente diferentes, y diferencias entre los grupos, las cuales pueden o no ser detectadas por la prueba aplicada. En la primera fuente se identifican características particulares de los ítemes que conducen a estimaciones erróneas de la habilidad de las personas (Angoff & Ford, 1973; Linn & Harnish, 1981; Scheuneman & Gerriz, 1990). En la segunda fuente, las diferencias en la instrucción recibida por los integrantes de cada grupo son la principal causa del DIF. Al respecto, Miller y Linn (1988)[2] investigaron en qué grado las funciones características de los ítemes, de grupos con diferentes experiencias instruccionales, eran invariables.

Los métodos de identificación del DIF frecuentemente se enfocan en la opción correcta; sin embargo, las otras opciones o las omisiones pueden ejercer un efecto importante sobre la dificultad del ítem y es probable que el origen del DIF se encuentre allí (Angoff, 1993). De igual manera, un índice para DIF muy grande puede indicar que el ítem está midiendo un constructo adicional en uno de los grupos, por lo que la hipótesis de unidimensionalidad no se satisface: la prueba no es unidimensional para al menos uno de los grupos, o no mide la misma dimensión en los grupos (Lord, 1980).

## Métodos

La Universidad de Bogotá Jorge Tadeo Lozano desde el año 2007 ofrece el Examen de Clasificación en Matemáticas Básicas, como una evaluación diagnóstica que posibilita identificar, analizar y evaluar el nivel de los conocimientos del aspirante y de los estudiantes que realicen transferencias internas o externas. En la actualización de su Proyecto Educativo Institucional ha establecido el proceso de selección dentro de las políticas de admisión de estudiantes y es por esto que considera:

> "En razón de la heterogeneidad en la formación de los estudiantes admitidos, la Universidad aplicará exámenes de clasificación y ofrecerá asignaturas de enlace bachillerato-universidad en aquellos temas y procesos cognitivos de mayor relevancia para los estudios universitarios, con el fin de alcanzar las condiciones académicas apropiadas para avanzar en los estudios de educación superior. Entre estas asignaturas se incluyen en especial, una asignatura de fundamentación en humanidades y otra en matemática básica" (Universidad de Bogotá Jorge Tadeo Lozano, 2011, p. 143).

Por el carácter decisorio que representa esta evaluación, la Universidad adelanta entre

[2] La idea de concentrar la atención en la instrucción como posible fuente del DIF corresponde con una visión que asegura que la habilidad desarrollada por los estudiantes es función del proyecto educativo, de las prácticas de aula, de las oportunidades de aprendizaje desplegadas al interior de una institución de educación, y no del género, raza, edad u otras variables de origen de la población que es objeto de evaluación.

otros estudios, el análisis del instrumento empleado. La base de datos que se utiliza consta de 1.623 cadenas de respuestas obtenidas en diferentes momentos de aplicación comprendidos entre el tercer período lectivo de 2011 (2011-III) y el primero de 2012 (2012-I). Se seleccionaron 61 ítemes, de un banco de 231, los cuales conformaron el instrumento empleado en todas las aplicaciones; estos 61 ítemes comprenden los dominios conceptuales propuestos en la asignatura Matemáticas Básicas. Utilizando la metodología de bloques completos, se diseñaron seis cuadernillos, cada uno conformado por dos bloques de 15 ítemes; cada estudiante evaluado respondió solo un cuadernillo.

Para el análisis de ítemes y la calibración del instrumento se empleó la Teoría de Respuesta al Ítem (TRI), con la que se obtuvieron, incluyendo todas las cadenas de respuestas de los estudiantes, los siguientes valores para los parámetros estimados: confiabilidad igual a .73; coeficiente *Alpha* de Cronbach, .40; porcentaje de variancia explicada por las medidas, 22.3%, y en contraste con el primer componente, 2.3%, e intervalo de dificultad de los ítemes desde -1.80 hasta 2.22 *logits*[3].

El instrumento analizado forma parte de la prueba empleada como Examen de Clasificación en Matemáticas Básicas que tiene lugar en dos sesiones al inicio de cada período lectivo. Las variables de contexto seleccionadas: Sexo, Naturaleza jurídica del colegio de procedencia, rango de Edad y Facultad en la que el aspirante tramita su ingreso, fueron recopiladas mediante preguntas directas enunciadas en el formulario de inscripción. Se presentaron datos faltantes debido a que no era obligatorio diligenciar los campos en los que se solicitaba esta información (ver Tabla 1).

Como el objetivo de la prueba es examinar el nivel de conocimiento de los aspirantes, de forma tal que la Universidad pueda ofrecer una ruta académica adecuada al estudiante, se pretende verificar que las diferencias en habilidad no provengan de las variables seleccionadas del grupo al que pertenece el evaluado.

En la conformación de cada base de datos para el estudio del DIF de las variables elegidas (una base por cada variable de contexto) se eliminaron las cadenas de respuestas correspondientes a los datos faltantes y a los grupos de tamaño reducido que no permitían una comparación efectiva. Por otra parte y con ayuda del paquete estadístico SPSS (1998) se seleccionaron muestras aleatorias con distribución uniforme para conformar grupos del mismo tamaño dentro de cada variable. Con estas consideraciones las bases de datos se organizaron con el siguiente número de registros en cada grupo de las variables de contexto: para la variable Sexo, 728; en Naturaleza jurídica del colegio, 309 registros; para Edad, 323; y para Facultad en la que el aspirante tramita su ingreso, 476 registros.

En el proceso de equiparar los grupos conformados para llevar a cabo el análisis de funcionamiento diferencial de las variables de contexto elegidas, se eliminaron en total 167 cadenas de respuestas para la variable Sexo, 1.005 para Naturaleza jurídica del colegio, 195 para Facultad y 654 para Edad.

Considerando estudiantes con similar habilidad, en grupos distintos según las variables antes señaladas, se inició un estudio del DIF, con el fin de detectar y retirar aquellos ítemes que presenten una diferencia en la dificultad estimada mayor que .64 *logits*[4], y así garantizar equidad en la asignación del puntaje correspondiente para los diferentes grupos de aspirantes. Si se confirma la presencia de DIF, se afecta la validez del instrumento

[3] Los valores utilizados en la escala recogidos por Zwick usan *Unidades Delta*. Se empleó la equivalencia 1 *logit* = 2.35 Unidades Delta, en razón de que la unidad trabajada con el *software* WINSTEPS Versión 3.73 es el *logit* (Linacre, 2008).

[4] Si un ítem muestra una diferencia de dificultad mayor que .64 *logits*, entre dos poblaciones, se dice que presenta funcionamiento diferencial de moderado a grande, de acuerdo con la clasificación de la severidad del DIF sugerida mediante la escala del ETS (Linacre, 2008).

para predecir la habilidad de los aspirantes evaluados.

Para garantizar un alto poder de discriminación en los ítemes y que suministren información con un nivel apreciable de confiabilidad, se usó la clasificación de la severidad del DIF según la escala propuesta por ETS (Linacre, 2008; Zwick, Thayer & Lewis, 1999) que se indica en la Tabla 2.

Se aplicaron tres métodos para detectar DIF: el primero, diferencia entre los centros de las dificultades, el segundo, diferencia entre los extremos más próximos de los intervalos de las dificultades y finalmente, la prueba estadística de Mantel y Haenszel.

## Resultados

### Prueba 1: Contraste del DIF - Diferencia entre los centros de las dificultades

El procesamiento de la información se llevó a cabo con el *software* WINSTEPS Versión 3.73 que estima la dificultad de cada ítem, dentro de los grupos de la población considerada por variable de contexto. En la Tabla 3 se ilustra el valor calculado del contraste del DIF, para la variable Sexo en el Item 1; los valores encontrados muestran que el ítem es más difícil para las mujeres en .20 *logits*, lo cual indica que tiene un DIF con severidad grado A, considerado no significante según el criterio propuesto por ETS que se muestra en la Tabla 2. De igual manera, el Item 59 tiene una dificultad de 2.34 para los aspirantes que cursaron su bachillerato en un colegio oficial, en tanto que para los de colegios no oficiales la dificultad es de 1.73. Así este ítem es más difícil en .61 *logits*, para quienes vienen de un colegio oficial. Esta situación clasifica al Item 59 con una severidad grado B que se interpreta como de ligera a moderada, de acuerdo con el mismo criterio aplicado al Item 1.

El Item 34 exhibe una dificultad de .44 para los aspirantes cuya Edad se encuentra entre los 15 y los 17 años y de -.68 para los que tienen edades entre 21 y 30 años; de este modo el ítem es más difícil para el primer grupo. Como el valor que representa la diferencia entre las dos dificultades es de 1.12 *logits*, el ítem se clasifica en Categoría C. Debido a que la severidad grado C es considerada de moderada a grande, es necesario realizar el análisis del DIF en las otras tres variables consideradas, antes de tomar la decisión de retirar el Item 34 del examen.

El número de ítemes, por categoría de severidad del DIF, en cada una de las variables de contexto se registra en la Tabla 4. En la categoría leve a moderada se ubicaron 10 ítemes tanto en el análisis combinado de la variable Edad, 18-20 vs 21-30 años, como en la de la variable Facultad, FCEA - FCNI. En moderada a grande se ubicaron 7 ítemes en el análisis combinado de la variable Edad, 15-17 vs 21-30 años y en la categoría no significante, entre 47 y 58 ítemes en cada uno de los análisis combinados de las variables de contexto.

Avanzando en el análisis del DIF para la variable Edad, en la cual se presentó el mayor número de ítemes clasificados en categoría C, se representan en la Figura 1, utilizando el *software* Geogebra 4.2, las dificultades de los 61 ítemes en dos dimensiones: para el grupo 15-17 años en el eje horizontal y para 21-30 años en el eje vertical. En la franja gris claro, se encuentran los ítemes que fueron clasificados en categoría no significante de acuerdo al criterio de severidad de ETS; esta franja se encuentra delimitada por las rectas $y = x + .43$ e $y = x - .43$. En la franja oscura, comprendida entre las rectas $y = x + .64$ e $y = x + .43$ y las rectas $y = x - .43$ e $y = x - .64$, se encuentran los ítemes que fueron clasificados en categoría ligera a moderada. Finalmente los ítemes que exhiben DIF, categoría C, se encuentran fuera de las bandas sombreadas, representados con un triángulo y el número del ítem correspondiente. El ítem que se encuentra más alejado de las rectas $y = x + .64$ ó $y = x - .64$ es el 34 con una dificultad promedio para el grupo 15-17 años de .44 y para el grupo 21-30 de -.68 y con una diferencia entre los centros de dificultad de 1.12.

La Tabla 5 ilustra los ítemes que presentan DIF con un grado de severidad de moderada a grande en al menos una de las variables estudiadas. Se observa que el Item 34 exhibe grado de severidad C en dos análisis combi-

nados de la variable Edad, entre los grupos de 15-17 con 21-30 años y los de 18-20 con 21-30 años; de igual manera el Item 59 exhibe grado de severidad C en tres de los análisis combinados, en las variables Colegio, oficial con no oficial y en Facultad, en los análisis combinados FCEA - FCNI y FCHAD - FCNI.

Al usar el método diferencia entre los centros de las dificultades, se resalta el hecho de que ningún ítem presentó DIF con severidad C para la variable Sexo, ni para el análisis combinado de la variable Edad entre los grupos de 15-17 y 18-20 años.

### Prueba 2: Contraste del DIF - Diferencia entre los extremos más próximos de los intervalos de las dificultades

Considerando el número de registros relativamente pequeño en cada base de datos y el consecuente incremento en el tamaño del error generado en la estimación de la dificultad de los ítemes, se definieron intervalos para la dificultad, dentro de cada grupo, del siguiente modo: el límite inferior, calculado como la dificultad menos el error estándar y el límite superior, como la dificultad más este error estándar.

Cuando los intervalos así definidos no se superponen, es decir, si el límite inferior de la dificultad para el grupo de menor desempeño[5] es mayor que el límite superior de esta dificultad para el grupo de mayor desempeño, se revela la existencia de DIF. En la Figura 2 se ilustra el mapa que se siguió para realizar los cálculos en la variable Sexo. Al identificar la diferencia entre las dificultades que presentaban los 61 ítemes para los hombres menos las que presentaban para las mujeres, se evidenciaron dos casos: ítemes relativamente más difíciles para las mujeres (diferencia negativa) o relativamente más fáciles para ellas (diferencia positiva). Para cada caso, el límite superior del intervalo de la menor de las dos dificultades puede ubicarse bien a la izquierda o a la derecha del límite inferior del intervalo con centro en la mayor de las dificultades, dando lugar a intervalos *disyuntos* o con intersección no vacía respectivamente. Si los ítemes presentan DIF, los intervalos para la dificultad entre los dos grupos comparados resultan *disyuntos*.

En la Tabla 6, las entradas representan el número de ítemes clasificados en cada categoría de severidad del DIF propuestas por la ETS y aplicadas en la Prueba 1, de acuerdo con el análisis combinado correspondiente, al realizar la diferencia entre los extremos más próximos de los intervalos de las dificultades; para ilustrar lo anterior, en la Categoría B quedaron 3 ítemes al hacer el análisis combinado en la variable facultad FCEA - FCNI, y en la categoría C se ubicó un ítem, tanto en el análisis combinado 15-17 vs 21-30 como en 18-20 vs 21-30, en la variable Edad. Cabe anotar que este ítem es el 34.

### Prueba 3: Prueba estadística de Mantel y Haenszel

El método de Mantel-Haenszel es un procedimiento estadístico para detectar DIF basado en la comparación de la proporción de respuestas correctas $p$ e incorrectas $q$ entre dos grupos: el grupo focal y el grupo de referencia. Los evaluados objeto de análisis componen el grupo focal y aquellos que sirven de base para la comparación están en el grupo de referencia. La prueba hace uso del cociente entre la proporción de respuestas correctas sobre respuestas no correctas ($odds = p / q$) del grupo de referencia y el *odds* en el grupo focal. Un ítem presenta DIF si para los integrantes del grupo de referencia se revela sistemáticamente un mayor valor del cociente mencionado (*odds*) respecto del valor para el grupo focal, dado un nivel similar de habilidad. Este método proporciona un estimador

[5] Para el grupo que muestra un menor desempeño relativo frente al ítem que se analiza, la dificultad resulta ser mayor que para el grupo con un desempeño relativo más alto.

de la magnitud del DIF llamado *cociente de razones* o también *odds-ratio* de *MH* ($\alpha_{MH}$) así como una prueba de significación estadística conocida como *ji cuadrado MH* ($\chi^2_{MH}$) con un grado de libertad. $\alpha_{MH}$ toma valores positivos siempre; cuando $\alpha_{MH} > 1$ indica que el ítem favorece al grupo de referencia. Si $0 < \alpha_{MH} < 1$, revela que el grupo focal presenta un mejor desempeño. Por último, si $\alpha_{MH} = 1$ no hay DIF (Dorans & Holland, 1993).

Es posible hacer una transformación de $\alpha_{MH}$ a una escala logarítmica: cuando $\alpha_{MH} > 1$, el logaritmo del *odds-ratio* es positivo y confirma la presencia de DIF a favor del grupo de referencia; si $0 < \alpha_{MH} < 1$, su logaritmo es negativo e indica DIF a favor del grupo focal.

Se dispone de un criterio para categorizar la severidad del DIF con base en la magnitud y en la significancia alcanzada por la prueba estadística. Este criterio fue propuesto por ETS (Dorans & Holland, 1993), para diversos valores del logaritmo del *odds-ratio* en una métrica *delta* $\Delta_{MH}$ conocido con el nombre de *Mantel-Haenszel delta difference* y denotado *MH D-DIF*. El criterio se concentra en dos aspectos: el valor absoluto de MH D-DIF y la significancia asociada a la hipótesis de que difiera de algunos niveles preestablecidos; ambos son importantes dado que valores pequeños de MH D-DIF pueden ser significativos y señalar la presencia de DIF.

Las categorías utilizadas, A, B y C indican en su orden DIF leve, moderado y severo:
- Categoría A: ítemes con |MH D-DIF| < 1 (*unidad delta*) o con valores de MH D-DIF que no sean significativamente diferentes de cero serán considerados con DIF leve.
- Categoría C: ítemes con |MH D-DIF| ≥ 1.5 y con valores de MH D-DIF que sean significativamente mayores que uno serán considerados con DIF severo.
- Categoría B: aquí están aquellos ítemes que no pertenezcan a las anteriores categorías. No obstante, se pueden especificar las siguientes opciones: la primera, compuesta por ítemes con valores de MH D-DIF que sean significativamente diferentes de cero pero con |MH D-DIF| ≥ 1 y valores de MH D-DIF que no sean significativamente mayores de uno. La segunda la componen ítemes con 1 ≤ |MH D-DIF| < 1.5.

Un criterio equivalente al de ETS pero en unidades *logit* denotando la magnitud ahora con DIF (ver Tabla 2) establece las categorías en esta métrica así:

- Categoría A: ítemes con |DIF| < .43 o con valores de DIF que no sean significativamente diferentes de cero serán considerados con DIF leve.
- Categoría C: ítemes con |DIF| ≥ .64 y con valores de DIF que sean significativamente mayores de .43 serán considerados con DIF severo.
- Categoría B: ítemes que no pertenecen a las anteriores categorías, también con dos opciones: la primera, compuesta por ítemes con valores de DIF que sean significativamente diferentes de cero pero con |DIF| ≥ .43 y valores de DIF que no sean significativamente mayores de .43; la segunda, con ítemes donde .43 ≤ |DIF| < .64.

Puesto que WINSTEPS 3.73 utiliza este último criterio, al tomar como base el reporte del DIF obtenido con este *software*, se detecta si el valor absoluto del tamaño del estadístico Mantel y Haenszel es mayor que .64 y el *valor-p* asociado a la correspondiente prueba estadística es menor que .05. Los hallazgos referidos a la severidad en categoría C, moderada a grande, se relacionan en la Tabla 7. Las celdas vacías indican que no se cumplieron las dos condiciones simultáneamente.

En la implementación del método de Mantel-Haenszel, el Item 20 fue clasificado en categoría C al realizar dos análisis combinados en la variable edad; los otros 8 ítemes fueron clasificados en esta categoría al realizar un análisis combinado en las variables Edad y Facultad. Al realizar los análisis de las variables Sexo y Edad, ningún ítem quedó clasificado en Categoría C, moderada a grande.

Después de la aplicación de los tres tipos de procedimientos: Diferencia entre centros, Diferencia entre extremos más próximos y Prueba Estadística de Mantel y Haenszel, se comparan los hallazgos en la Figura 3.

El Item 34 presenta consistentemente funcionamiento diferencial, pues aparece en la intersección de los grupos de ítemes encontrados con los tres procedimientos utilizados.

### Análisis del Item 34 en el Examen de Clasificación en Matemáticas Básicas

El Item 34 fue vinculado al bloque 3 del Examen de Matemática, y en las aplicaciones consideradas lo contestaron 705 aspirantes, de los cuales 93 tenían edades entre 21-30 años. En dicha prueba se lo formuló con el siguiente texto:

*34) Una asociación de actores tiene 28.000 miembros, pero sólo el 30% de ellos tiene empleo. El número de actores que se encuentra desempleado es:*

*a. 4.000*
*b. 8.400*
*c. 9.333*
*d. 19.600 (Clave)*

Los resultados obtenidos fueron los siguientes: 12 estudiantes marcaron la opción A, 280 la B, 82 la C y 327 la clave. Únicamente se presentaron 4 omisiones. El proceso de calibración con las 1.623 cadenas de respuestas reporta como parámetros para el Item 34: dificultad, .11; error, .08; ajuste próximo, 1.04; ajuste lejano, 1.10; discriminación, .81; correlación .31 y porcentaje de respuesta correcta 46.4%.

El Item 34 presentó menor dificultad para los aspirantes con edades entre 21-30 que para los que tienen edades entre 15-17 y 18-20: para estudiantes con edad de 15-17, la dificultad resultó igual a .44 *logits*; con edad de 18-20, igual a .47 *logits*; y para edad de 21-30, igual a -.68.

### Curva Característica del Item 34

Para analizar el desempeño estadístico del Item 34 se graficó con el *software* Excel su curva característica CCI, en donde se representa la probabilidad de respuesta correcta para el ítem como una función de la habilidad. Se observa que el grupo de evaluados en el rango 21 a 30 años tiene consistentemente una mayor probabilidad de respuesta correcta con respecto a los otros dos grupos, controlando por habilidad. Los estudiantes con habilidad media del grupo 18 a 20 años tienen mayor probabilidad de respuesta correcta que los de 15 a 17, pero en aquellos con mayor habilidad tienen mayor probabilidad de respuesta correcta quienes están en el grupo de 15 a 17 años. Esto evidencia la presencia de funcionamiento diferencial en el Item 34 (ver Figura 4).

## Conclusiones

Se analizó el instrumento empleado en el Examen de Clasificación en Matemáticas Básicas de los estudiantes que ingresan a la Universidad de Bogotá Jorge Tadeo Lozano con el fin de establecer si algunos de los ítemes incluidos en dicha prueba presentaban DIF en cuatro variables de contexto: Sexo, Naturaleza jurídica del colegio de procedencia, rango de Edad y Facultad en la que el aspirante tramita su ingreso. El estudio se realizó teniendo como base 1.623 registros y se seleccionaron muestras con una distribución uniforme de modo que en cada base se garantizara el mismo tamaño de los grupos en cada variable de contexto.

Al realizar la Prueba 1 (Contraste del DIF-Diferencia entre los centros de las dificultades) se encontró que los ítemes 34 y 59 exhibieron un grado de severidad grande, Categoría C, en dos de los análisis combinados de la variable Edad para el primer ítem y para el segundo ítem, en el análisis de la variable Colegio y en dos análisis de la variable Facultad.

Con la Prueba 2 (Contraste del DIF - Diferencia entre los extremos más próximos de los intervalos de las dificultades) se clasificó únicamente el Item 34 en Categoría C, al realizar en la variable Edad, el análisis combinado entre el grupo de 15 a 17 años con el de 21 a 30 y entre los de 18 a 20 años con los de 21 a 30. Teniendo en cuenta el criterio de se-

veridad propuesta por ETS, con sus respectivos valores, todos los ítemes quedaron en categoría A en el análisis de la variable Edad.

Al realizar la Prueba 3 (Prueba estadística de Mantel y Haenszel) se clasifican en categoría C los ítemes 13, 20, 34 y 61 en la variable Edad y los ítemes 4, 30, 36, 43 y 59 para la variable Facultad; los valores considerados para esta asignación en categoría C corresponden al estadístico Mantel y Haenszel mayor que .64 y el *valor-p* asociado a la correspondiente prueba estadística menor que .05.

Una vez realizado el estudio utilizando las tres pruebas se concluye que ningún ítem presenta DIF al hacer la comparación en la variable Sexo, en tanto que la variable Edad es la que muestra mayor número de ítemes con clasificación en categoría C en dos de los análisis combinados. Al comparar los resultados encontrados con los tres procedimientos utilizados, se observa que consistentemente el Item 34 exhibe funcionamiento diferencial, razón por la cual se recomienda retirarlo en el procesamiento, antes de proceder a estimar las habilidades de los estudiantes evaluados.

TABLA 1
DISTRIBUCIÓN PORCENTUAL Y DE FRECUENCIA DE LOS ESTUDIANTES EVALUADOS DE 2011-III A 2012-I

| Variable | Grupo | Frecuencia absoluta | Frecuencia relativa % |
|---|---|---|---|
| Sexo | Femenino | 888 | 54.7 |
| | Masculino | 728 | 44.9 |
| | No hay información | 7 | .4 |
| Naturaleza jurídica del colegio | Oficial | 1072 | 54.7 |
| | No Oficial | 309 | 44.9 |
| | No hay información | 242 | .4 |
| Rango de edad | 15 - 17 | 403 | 24.8 |
| | 18 - 20 | 857 | 52.8 |
| | 21 - 30 | 323 | 19.9 |
| | 31 - 54 | 30 | 1.8 |
| | No hay información | 10 | .6 |
| Facultad | FCEA | 488 | 30.1 |
| | FCHAD | 641 | 39.5 |
| | FCNI | 476 | 29.3 |
| | FRICJP | 6 | .4 |
| | Convenios colegio-proyecto enlace | 5 | .3 |
| | No hay información | 7 | .4 |

*Nota:*
*N* = 1.623 cadenas de respuestas
FCEA: Facultad de Ciencias Económico Administrativas
FCHAD: Facultad de Ciencias Humanas Arte y Diseño
FCNI: Facultad de Ciencias Naturales e Ingeniería
FRICJP: Facultad de Relaciones Internacionales y Ciencias Jurídicas y Políticas

TABLA 2
CRITERIO DE SEVERIDAD DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMES SEGÚN ETS

| Categoría para DIF | Contraste DIF en *logits* | Significancia estadística del DIF |
|---|---|---|
| C: de Moderado a Grande | \|DIF\| ≥ .64 | $p(\|DIF\|) = .43) < .05$ |
| B: de Ligero a Moderado | .43 ≤ \|DIF\| < .64 | $p(\|DIF\|) > 0) < .05$ |
| A: No significante | \|DIF\| < .43 | - |

TABLA 3
CLASIFICACIÓN DEL FUNCIONAMIENTO DIFERENCIAL DE UN ÍTEM EN CATEGORÍA A

| Item | Dificultad F (Femenino) | Dificultad M (Masculino) | F-M (Contraste del DIF) | ABS (F-M) | Categoría |
|---|---|---|---|---|---|
| 1 | .29 | .49 | -.20 | .20 | A |

TABLA 4
NÚMERO DE ÍTEMES CLASIFICADOS EN CADA CATEGORÍA DE SEVERIDAD DEL DIF MEDIANTE EL MÉTODO DIFERENCIA ENTRE LOS CENTROS DE LAS DIFICULTADES

| Variable | Análisis combinado | Severidad | | |
|---|---|---|---|---|
| | | No significante | Leve a Moderada | Moderada a Grande |
| Sexo | Femenino - Masculino | 58 | 3 | 0 |
| Edad | 15 - 17 vs 18 - 20 | 53 | 8 | 0 |
| | 15 - 17 vs 21 - 30 | 47 | 7 | 7 |
| | 18 - 20 vs 21 - 30 | 50 | 10 | 1 |
| Colegio | Oficial - No Oficial | 53 | 5 | 3 |
| Facultad | FCEA - FCHAD | 55 | 4 | 2 |
| | FCEA - FCNI | 50 | 10 | 1 |
| | FCHAD - FCNI | 57 | 3 | 1 |

TABLA 5
ITEMES CLASIFICADOS CON GRADO DE SEVERIDAD DE MODERADA A GRANDE (C) EN ANÁLISIS DEL DIF MEDIANTE EL MÉTODO DIFERENCIA ENTRE LOS CENTROS DE LAS DIFICULTADES, SEGÚN LAS VARIABLES DE CONTEXTO

| Item | Sexo | Edad | | | Colegio | Facultad | | |
|---|---|---|---|---|---|---|---|---|
| | Femenino<br>Masculino | 15 - 17<br>18 - 20 | 15 -17<br>21-30 | 18 - 20<br>21 - 30 | No Oficial -<br>Oficial | FCEA<br>FCHAD | FCEA<br>FCNI | FCHAD<br>FCNI |
| 1 | A | B | C | A | A | A | A | A |
| 2 | A | B | C | A | A | A | A | A |
| 6 | A | A | C | A | B | A | A | A |
| 20 | B | A | C | B | A | A | A | A |
| 34 | A | A | C | C | A | A | A | A |
| 53 | A | A | C | A | A | A | B | A |
| 57 | A | B | C | A | A | A | B | A |
| 11 | B | A | A | A | C | A | A | A |
| 29 | A | A | B | A | C | A | B | B |
| 59 | A | A | A | A | C | A | C | C |
| 30 | A | A | A | A | A | C | B | A |
| 36 | A | A | A | A | B | C | B | A |

TABLA 6
NÚMERO DE ÍTEMES CLASIFICADOS EN CADA CATEGORÍA DE SEVERIDAD DEL DIF MEDIANTE EL MÉTODO DIFERENCIA ENTRE LOS EXTREMOS MÁS PRÓXIMOS DE LOS INTERVALOS DE LAS DIFICULTADES

| Variable | Análisis combinado | Severidad | | | No hay DIF |
|---|---|---|---|---|---|
| | | No significante<br>A | Leve a Moderada<br>B | Moderada a Grande<br>C | |
| Sexo | Fem. - Masc. | 18 | 0 | 0 | 43 |
| Edad | 15 -17 vs 18 - 20 | 12 | 0 | 0 | 49 |
| | 15 -17 vs 21 - 30 | 20 | 2 | 1 | 38 |
| | 18 -20 vs 21 - 30 | 9 | 0 | 1 | 51 |
| Colegio | Oficial - No Oficial | 11 | 1 | 0 | 49 |
| Facultad | FCEA - FCHAD | 19 | 1 | 0 | 41 |
| | FCEA - FCNI | 18 | 3 | 0 | 40 |
| | FCHAD - FCNI | 30 | 0 | 0 | 31 |

TABLA 7
ITEMES CLASIFICADOS CON GRADO DE SEVERIDAD C DE ACUERDO CON EL ESTADÍSTICO DE MANTEL Y HAENSZEL

| Item | Categoría en las variables de contexto | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Sexo | Edad | | | Colegio | Facultad | | |
| | Fem.<br>Masc. | 15 - 17<br>18 - 20 | 15 - 17<br>21 - 30 | 18 - 20<br>21 - 30 | No oficial - Oficial | FCEA<br>FCHAD | FCEA<br>FCNI | FCHAD<br>FCNI |
| 13 | | | | C | | | | |
| 20 | | | C | C | | | | |
| 34 | | | | C | | | | |
| 61 | | C | | | | | | |
| 4 | | | | | | C | | |
| 30 | | | | | | | C | |
| 36 | | | | | | | C | |
| 43 | | | | | | | | C |
| 59 | | | | | | | | C |

FIGURA 1
DIAGRAMA DE LAS DIFICULTADES DE LOS ÍTEMES EN LOS GRUPOS CONSIDERADOS: 15 - 17 AÑOS (EJE *x* ) Y 21 - 30 AÑOS (EJE *y*)

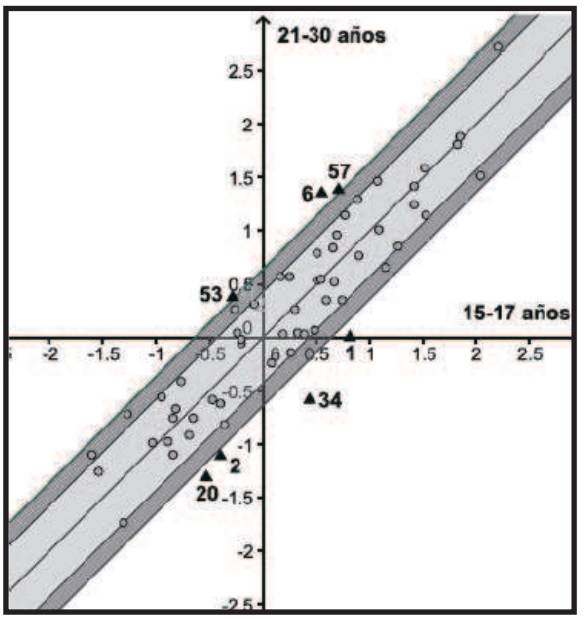


FIGURA 2
MAPA PARA DIFERENCIA ENTRE LOS EXTREMOS MÁS PRÓXIMOS DE LOS INTERVALOS DE LAS DIFICULTADES. VARIABLE SEXO

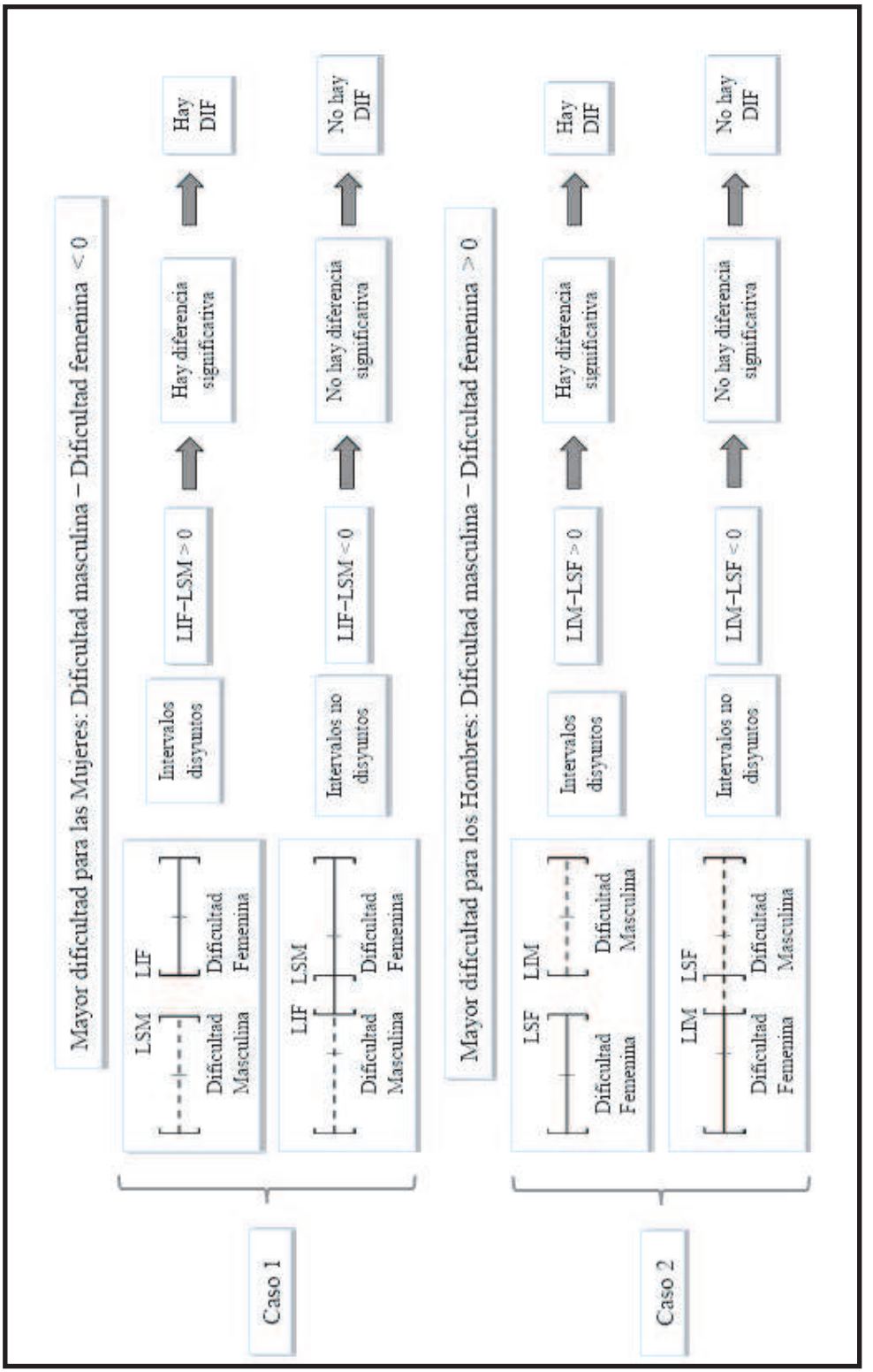


FIGURA 3
ITEMES EN CATEGORÍA C EN CADA UNO DE LOS PROCEDIMIENTOS IMPLEMENTADOS

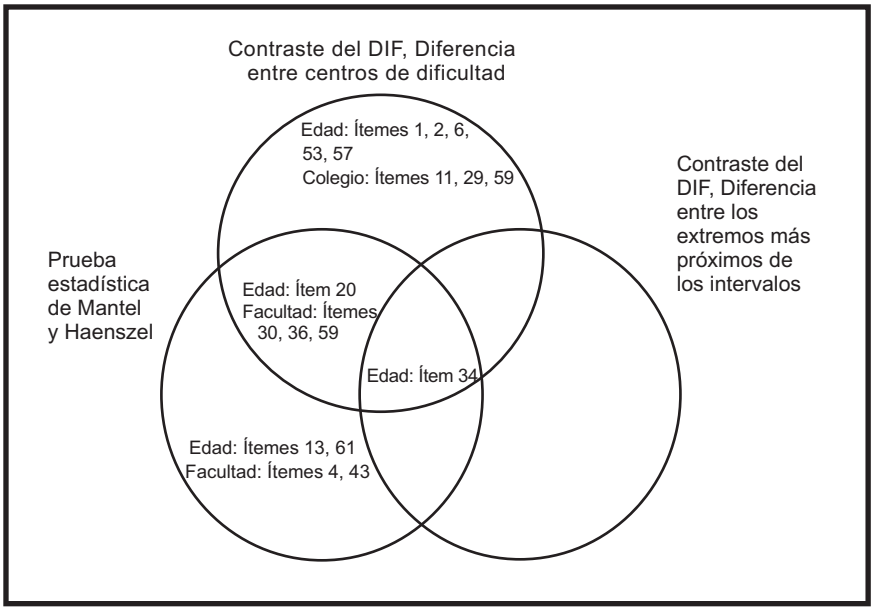
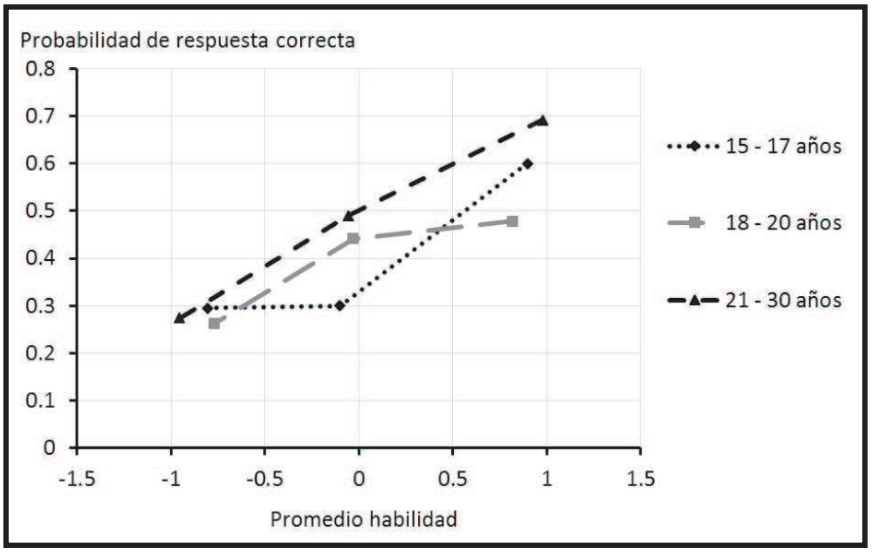



FIGURA 4
CURVA CARACTERÍSTICA DEL ITEM 34



## Referencias bibliográficas


Angoff, W. (1993). Perspectives on differential item functioning methodology. En P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). New Jersey: Erlbaum.

Angoff, W. & Ford, S. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*(2), 95-106. http://dx.doi.org/10.1111/j.1745-3984.1973.tb00787.x

Attorresi, H., Galibert, M., Zanelli, M., Lozzia, G. & Aguerri, M. (2003). *Error tipo I en el análisis del funcionamiento diferencial del ítem basado en la diferencia de los parámetros de dificultad* [Type I error in the analysis of differential item functioning based on the difference in the difficulty parameters]. *Psicológica, 24* (002), 289-306.

Baker, F. (2001). *The basics of item response theory* (2da. ed.). Wisconsin, USA: ERIC Clearinghouse on Assessment and Evaluation.

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

Cleary, T. & Hilton, T. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28,* 61-75. http://dx.doi.org/v10.1177/001316446802800106

Dorans, N.J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355-368. http:// dx.doi.org/10.1111/j.1745-3984.1986.tb00255.x

Dorans, N. & Holland, P. (1993). DIF detection and description: Mantel-Haenszel and standardization. En P. Holland & H. Wainer (Ed.), *Differential item functioning* (pp. 35-66). New Jersey: Erlbaum.

Linacre, J. (2008). *A user's guide to WINSTEPS Rasch-Model computer programs.* Chicago: John M. Linacre.

Linn, R. & Harnisch, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*(2), 109-118. http://dx.doi.org/10.1111/j.1745-3984.1981.tb 00846.x

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Michigan: Erlbaum.

Miller, M. & Linn, R. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*(3), 205-219. http://dx.doi.org/10.1111/j.1745-3984.1988.tb00303.x

Raju, N. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207. http://dx.doi.org/10.1177/014662169001400208

Scheuneman, J. & Gerritz, K. (1990). Using differential items functionig procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*(2), 109-131.http://dx.doi.org/10.1111/j.1745-3984.1990.tb00737.x

SPSS Base 8.0 for Windows (1998). *User's guide.* Chicago IL: SPSS Inc.

Swaminathan, H. & Rogers, H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. http://dx.doi.org/10.1111/j.1745-3984. 1990.tb00754.x

Universidad de Bogotá Jorge Tadeo Lozano. (2011). *Proyecto educativo institucional PEI* [Institutional educational project PEI]. Bogotá: Universidad de Bogotá Jorge Tadeo Lozano.

http://www.utadeo.edu.co/files/collections/documents/field_attached_file/pei_2012.pdf

Zwick, R., Thayer, D. & Lewis, C. (1999). An empirical bayes approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement, 36*(1), 1-28. http://dx.doi.org/10.1111/j.1745-3984.1999.tb00543.x

*Universidad Jorge Tadeo Lozano (UJTL)*
*Bogotá - Colombia*

Fecha de recepción: 1 de octubre de 2013
Fecha de aceptación: 4 de febrero de 2014