IMPACTO Y FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS RESPECTO AL GÉNERO EN EL TEST DE EFICACIA LECTORA (TECLE)

IMPACT AND DIFFERENTIAL FUNCTIONING OF THE ITEMS CONCERNING THE GENDER IN THE TEST OF READING EFICACIA (TECLE)

Daniel Costa Ball
Universidad Católica del Uruguay, Uruguay
María Gründel
Universidad Católica del Uruguay, Uruguay
Ariel Cuadro
Universidad Católica del Uruguay, Uruguay

Resumen: Este trabajo tiene como objetivo estudiar el posible funcionamiento diferencial de los ítems (FDI) que componen la Prueba de Eficacia Lectora (TECLE). Se utilizó el estadístico de Mantel-Haenszelt (MH) como procedimiento de detección del FDI para evaluar si los ítems del TECLE funcionan de forma distinta en grupos igualados en función del género. La muestra está conformada por 1159 alumnos de educación primaria privada con edades comprendidas entre 9 y 12 años (50,5% varones y 49,5% niñas). Los resultados del análisis del FDI encontraron que según el estadístico MH que el porcentaje de ítems con FDI es del 6,25%. Independientemente de que puedan existir diferencias reales en las habilidades de eficacia lectora entre niñas y niños en cuatro del los 64 ítems del TECLE, no es preciso realizar una baremación por sexo. Habría que considerar continuar con los estudios con jueces expertos en el tema para despejar la sospecha de FDI.

Palabras clave*: funcionamiento diferencial del ítem, impacto, eficacia lectora.*

Abstract: The aim of this paper is to study the possible differential item functioning (DIF) on the Test of Reading Efficiency (TECLE). Mantel-Haenszelt (MH) analyses were performed for DIF detection purposes, to test whether the TECLE items performed differentially by gender-based group. The sample (n=1159) was composed by 9 to 12 year old elementary school students (50,5 % males, 49,5 % females). Analyses yielded DIF on 6,25% of items. Although the literature suggests there were differential reading efficacy skills among male and female participants in 4 out of 64 items of the measure reported on here, gender-based standardization doesn't appear to be needed. We should consider the possibility of continue studying this subject by expert judges who would contribute to acquire further knowledge on DIF.

Keywords: *differential item functioning, impact, reading skills.*

Correspondencia: Daniel Costa Ball. Facultad de Psicología, Universidad Católica del Uruguay. Correo Electrónico: ccosta@ucu.edu.uy.

INTRODUCCIÓN

El problema a investigar concierne al *funcionamiento diferencial de los ítems* (FDI sigla en Español y DIF en Inglés) en una Prueba de Eficacia Lectora (TECLE; Cuadro, Costa, Trias & Ponce de León, 2009) para alumnos de educación primaria privada que cursan cuarto a sexto año en Montevideo.

A finales de la década de los años setenta, el estudio de la falta de equidad en las puntuaciones de los tests cobró relevancia (Martínez, Hernández & Hernández, 2006) y buena parte de la investigación en psicometría se ha centrado en el estudio sobre el posible sesgo de los test, o de parte de sus ítems, este tema ha ocupado un lugar importante en la investigación psicométrica de los últimos 30 años y es previsible que lo siga siendo en el futuro (Gómez-Benito, Hidalgo & Guilera, 2010).

En Uruguay a pesar de la importancia del tema, la metodología de análisis de funcionamiento diferencial de los ítems (FDI) no ha sido usada y no se encuentran publicaciones en el país realizando una búsqueda en la base de datos Scielo con la palabra clave *funcionamiento diferencial de los ítems.* En el país*, g*eneralmente se utilizan test adaptados para España u otrospaíses (México, Argentina, etc.), actualmente ha comenzado un proceso de construcción / adaptación de test a nuestra realidad cultural. Es en este contexto donde cobra relevancia el análisis de FDI en el proceso de adaptación de instrumentos estandarizados. Con el presente trabajo queremos hacer un aporte metodológico para nuestro país, al utilizar el procedimiento de Mantel Haenszel (MH) creado por Mantel y

Haenszel (1959) y aplicada al análisis de FDI por Holland y Thayer en 1988 (citado en Gómez & Hidalgo, 1997) para identificar el FDI en una prueba construida en Uruguay que evalúa la eficacia lectora (TECLE) y mejorar los estudios psicométricos de los análisis de validez con los procedimientos del FID (APA, AERA & NCME, 1999). Estas normas, incluyen el análisis del FDI y el sesgo en el análisis de validez con el objetivo de lograr que las evaluaciones psicológicas garanticen la "equidad y validez de las interpretaciones y decisiones adoptadas a partir de las mismas. Para ello es necesaria la utilización de instrumentos libres de sesgo, y capaces de evaluar necesidades personales y sociales de individuos con diferentes características" (Gómez-Benito et al., 2010, p.75). Tomando en cuenta que los tests se han convertido en instrumentos de medidas necesarios para la evaluación en psicología, tanto a nivel laboral como educativo, dada la importancia de los mismos, la utilización incorrecta termina resultando nefasta. En especial cuando son usados con la finalidad de tomar determinadas decisiones sobre los individuos, de tipo de selección, promoción o evaluación del sujeto evaluado o son usados para llevar a cabo estudios comparativos entre distintos grupos. Por este motivo, se debe promover la construcción de tests que respeten los *Standars for Educational and Psychological Testing* (APA et al., 1999). Estas normas intentan dar respuestas a las problemáticas que se generan en el proceso de creación/adaptación y uso de tests (Carretero-Dios & Pérez, 2007; Gómez-Benito et al., 2010; Gómez & Hidalgo, 1997; Muniz & Hambleton, 1996). Las directrices internacionales para el uso de los tests en el apartado segundo, sobre la utilización adecuada de los tests recomiendan prestar atención a los aspectos relacionados con el sesgo de los tests. Cuando el test será usado en diferentes grupos hay que asegurarse que los mismos sean imparciales y adecuados para todos los grupos evaluados; que el constructo que mide el test sea relevante o adecuado para los grupos en cuestión; estudiar los rendimientos de los grupos en el test y analizar el funcionamiento diferencial de los ítems (FDI) en los grupos que se le administra el test (International Test Commission, 2010). Aplicar las normas APA et al. (1999) en el proceso de construcción del test, garantiza que las evaluaciones psicológicas tengan equidad y validez de sus interpretaciones y decisiones que de las mismas devengan, se han convertido en un punto central para lograr estos objetivos. Analizando los instrumentos de medidas usados para asegurarse que no presenten sesgo a favor o en contra de un grupo, para alcanzar evaluaciones válidas y justas para personas de diferentes grupos. La metodología utilizada para realizar estos estudios es análisis del funcionamiento diferencial de los ítems (Gómez, septiembre 2005), el término fue introducido por Holland y Thayer en 1988 (citado en Gómez, 1997).

Es oportuno diferenciar lo que se entiende por Impacto, FDI y Sesgo, Ackerman en 1992 (citado en Fidalgo, 1996) define impacto como "una diferencia entre grupos en el desempeño en un ítem causada por una diferencia real en la variable medida" (p.377). Camilli y Shepard en 1994 definen el término impacto refiriéndose al hecho que "un instrumento de medida obtenga resultados sistemáticamente inferiores en un grupo en comparación a otro no necesariamente implica la presencia de FDI, sino que pueden existir diferencias reales entre los grupos en el rasgo medido por el test en cuestión" (citado en Gómez-Benito et al., 2010, p.76). Por otro lado, "un ítem presenta funcionamiento diferencial (FDI) cuando existen diferencias en la puntuación de medida obtenida en ese ítem por dos grupos distintos de sujetos pero con el mismo nivel en el rasgo o característica evaluada por el test" (Barbero, Vila & Suárez, 2006, p.459). Cohen, Kim y Baker en 1993 clasifican los estudios de FDI en dos tipos según los objetivos pretendidos, los primeros tratan de estudios para la detección de FDI. En este grupo están las investigaciones que emplean algún método tradicional para identificar el FDI y los segundos para la verificación del impacto del FDI, en este último, están las investigaciones realizadas para identificar las supuestas causas del FDI (citado en Andriola, 2002). Por último, la definición de sesgo "en la actualidad se ha llegado al acuerdo que el término alude a que los investigadores buscan las causas por las cuales determinados ítems presentan un comportamiento diferencial en función de ciertas variables" (Gómez-Benito et al., 2010, p.76).

En la década de 1960 comenzaron a surgir los procedimientos de detección del FDI, los cuales se han clasificado en dos categorías, la primera comprende los métodos empíricos (métodos de invariancia condicional observada o métodos condicional) que utiliza le Teoría Clásica de los Tests (TCT) y la segunda, incluye los métodos teóricos (métodos de invariancia condicional no observada) que utiliza la Teoría de Respuesta al ítem (TRI) (Andriola, 2002).

El estudio de las propiedades psicométricas de los tests, se ha vuelto una práctica común entre los profesionales de la psicología. Una consecuencia directa de este hecho, se vincula con la necesidad de perfeccionar los instrumentos de evaluación (Santisteban, 2009). Así, en este contexto, los estudios de *comportamiento diferencial de los ítems* han reunido la atención de profesionales e investigadores, tanto a nivel metodológico como a nivel aplicado. Teniendo información sobre la calidad métrica de los ítems que componen un test, proporciona información para decidir cuáles ítems han de utilizarse (o seguir utilizándose) y cuáles sería mejor desechar debido a su "baja calidad técnica". En otras palabras, analizar los ítems resulta de utilidad y contribuye a la construcción y mejoramiento de los tests, ayudando a su vez a maximizar la fiabilidad y validez del mismo (Barbero et al., 2006). Realizar el análisis cuantitativo de los ítems que componen un test, implica estudiar sus propiedades estadísticas y psicométricas. Para esto, es preciso haber contado previamente con el análisis de validez de contenido, de

constructo, formato y calidad de redacción de los ítems. Cumplido y aprobado esto, sabemos que un test nos brindará medidas de la variable que pretendemos medir con escaso poco error (Muñiz, 1996). Esto significa que un test válido brinda medidas idénticas a personas o grupos con niveles iguales en la variable medida. Cuando un test presenta ítems cuya probabilidad de acierto difiere entre subgrupos de igual nivel de la variable medida, existe FDI.

El estudio del FDI de los ítems se convierte en un instrumento importante para evaluar las diferencias en la actuación de los sujetos pertenecientes a distintos grupos y evaluados con el mismo test. Este estudio, permite distinguir en la actuación de los sujetos en distintos grupos entre diferencias reales (Impacto) y ficticias (FDI) (Escorial & Navas, 2006). Es importante que los profesionales usuarios de los tests, se aseguren de que al usar los mismos estén brindando igualdad de oportunidades a los sujetos que pasan por la administración, o en otras palabras, que el test sea justo (Gómez-Benito, Hidalgo & Guilera, 2010). En especial, el estudio del FDI, tiene como objetivo garantizar que las puntuaciones obtenidas en un test en diferentes grupos sociales tengan el mismo significado, para conseguirlo, hay que utilizar los pasos recomendados por las Normas publicadas en 1999 (APA et al., 1999), y la guía International Test Commission (2010).

El hallazgo del funcionamiento diferencial en el ítem siempre se lo tiene que interpretar con precaución, por el solo rechazo estadístico de la hipótesis nula de que ambos grupos (focal – referencia) tienen la misma probabilidad de acertar el ítem cuando están igualados por nivel de habilidad, evidencia solamente la posible aceptación de un sesgo hipotético. Es responsabilidad del investigador consultar a expertos en el área evaluada para analizar el contenido y contexto, para poder determinar así, la existencia de una posible interacción entre el contenido del ítem y alguna característica específica del grupo que puede explicar la posible contaminación del proceso de medida (Elosua, López & Tarres, 2000).

Por este motivo, en 1985 la American Educational Research Association (AERA), junto con American Psychological Association (APA) y el National Council on Measurement in Education (NCME) han publicado los Estándares para los Tests Psicológicos y Educacional proporcionando un marco teórico para la construcción y adaptación de instrumentos de evaluación psicológica (Barbero, Abad & Holgado, 2008) hasta años después, fueron sustituidas por las normas APA et al. (1999). Con el fin de garantizar que la puntuación en el test tenga el mismo resultado en los diferentes grupos que se ha aplicado o evaluado, se requieren diversos estudios de validez, recientemente la International Test Commission (2010) ha elaborado las nuevas directrices para la traducción y adaptación de tests.

En la actualidad, se puede confirmar que en los últimos años en el mundo la preocupación por el estudio del FDI continúa vigente mientras que en habla hispana se ha publicado poco. En Uruguay Gründel & Costa (2010, octubre) presentaron un Póster en el III Simposio de investigación de la Facultad de Psicología de la Universidad Católica sobre los resultados de la Memoria de grado sobre la detección del funcionamiento diferencial de los ítems en el TECLE, en aquellos ítems que previamente se había detectado impacto (Gründel, 2010).

Desde hace unos años, desde la perspectiva cognitiva, se ha avanzado en el conocimiento de los procesos implicados en la lectura, analizando sus componentes, para entender cómo se adquiere la lectura y así, aproximarse a entender las dificultades en la adquisición de la misma. En el 2009, la Facultad de Psicología de la Universidad Católica del Uruguay, publicó la prueba TECLE. Para la construcción de la misma se utilizó la Teoría Clásica de los Test (TCT), siendo una prueba cognitiva de rendimiento en eficacia lectora de ejecución máxima, conformando un tipo de prueba caracterizada por ser un test de velocidad y no de potencia. La misma presenta 64 ítems y fue creada para ser administrada a alumnos de educación primaria. El alumno, debe responder a cada ítem leyendo una frase introductoria que está incompleta. Luego, tiene que decidir entre un conjunto de 4 opciones cual es el estímulo correcto. La tarea del lector consiste en seleccionar la opción que completa la frase en el menor tiempo posible. Para esto, el sujeto ha de leer la frase, comprender su sentido y utilizar tal información para discernir entre las opciones que se le presentan. El conjunto de opciones está conformado por: la opción correcta, un distractor fonológico, un distractor ortográfico y un distractor léxico. Las frases que introducen al sujeto, varían en complejidad sintáctica (medida por el número de palabras), semántica (medida por la familiaridad de las palabras que las componen) y ortográfica (medida por la longitud y estructura silábica). Dado que el TECLE es un test de velocidad, los ítems se caracterizan por presentar un índice de facilidad elevado, todos los ítems son fáciles de contestar (Cuadro et al., 2009).

La prueba fue diseñada con el cometido de evaluar el nivel de Eficacia Lectora que presentan los alumnos de primaria al enfrentarse a una tarea que propone frases incompletas. El objetivo es que los sujetos completen la mayor cantidad de frases posibles en un tiempo de cinco minutos. Los ítems del test están dispuestos de forma tal que al inicio aparecen los menos difíciles y gradualmente va aumentando la dificultad. Los sujetos que presenten dificultades en la descodificación, requerirán mayor cantidad de tiempo para lograr elegir la opción que completa cada frase. Por ende, estos sujetos, obtendrán una puntuación total menor. Por otra parte, los sujetos que presenten

dificultades en la ruta ortográfica (dificultad en el acceso a la representación ortográfica y en consecuencia semántica de las palabras), requerirán de más cantidad de tiempo para leer. Por este motivo, estos sujetos insumirán más tiempo en la lectura y en consecuencia, la producción en la prueba será menor (Cuadro et al., 2009).

En el marco de estudio del FDI, el objetivo del presente trabajo es el análisis de los ítems que conforman la Prueba de Eficacia Lectora (TECLE, Cuadro et al., 2009) según la variable género, en alumnos escolares de colegios privados de Montevideo, con edades comprendidas entre los 9 y 12 años. Con este trabajo queremos brindar un aporte metodológico al comenzar a utilizar en nuestro país el método Mantel Haenszel para identificar el FDI, en segundo lugar con este estudio complementará los análisis psicométricos de la prueba TECLE publicada por nuestro equipo de investigación en relación a la validez, cumpliendo con los criterios de la International Test Commission (2010) y las normas APA et al. (1999).

MÉTODO

Participantes

La muestra está conformada por 1159 alumnos, 396 cursan 4° año, 388 cursan 5° y 375 en 6° año de enseñanza primaria con edades comprendidas entre los 9 y 12 años que estudian en colegios privados de Montevideo. 585 varones (50,5%) y 574 niñas (49,5%). Se utilizó el método de muestreo sistemático, con punto de arranque aleatorio e intervalo constante igual a la inversa de la fracción del muestreo, seleccionando 18 colegios. El grupo de referencia quedo conformado por el sexo femenino y el grupo control por el grupo masculino.

Medidas

Se utilizó la prueba TECLE (Cuadro et al., 2009) que evalúa la habilidad lectora, compuesta por 64 ítems. El modelo estadístico en que se basa el TECLE es el modelo psicométrico de la Teoría Clásica de los Test (TCT).

Es un tipo de prueba de velocidad, está conformado "por ítems fáciles, de tal forma que cualquier sujeto debería ser capaz de contestar correctamente a todos ellos, aunque por limitaciones de tiempo no todos los sujetos llegan a contestar el mismo número de ítems en el test" (Martínez et al., 2006, p.62). El alumno debe responder a cada ítem leyendo una frase introductoria que está incompleta porque le falta la última palabra, teniendo que decidir entre un conjunto de 4 opciones, tres incorrecta y una correcta.

El análisis de ítems según la TCT, evidenció que el índice de dificultad de los ítems oscila entre .80 y .99, mientras la formula de Stafford utilizada para medir si el TECLE es una prueba de potencia o velocidad, los resultados arrojaron valores próximos a uno, evidenciando que la prueba es un test de velocidad. El estudio de dimensionalidad a través del programa NOHARM (Fraser, 1988) evidencia que la prueba es unidimensional y su fiabilidad evaluada con el método test-retest fue de .88 (Cuadro et al., 2009).

Procedimiento

Aplicación de la prueba. Se aplicó la prueba a los 1159 alumnos, el TECLE es un test de velocidad y la prueba está diseñada para evaluar la cantidad de ítems que el alumno responde en cinco minutos. En esta oportunidad, administramos la prueba sin límite de tiempo y todos los alumnos completaron los 64 ítems. Con la base de datos con todas las respuestas de los alumnos a todos los ítems, se corrieron los análisis de dimensionalidad, impacto y FID.

Análisis

En primer lugar se realizó el análisis factorial confirmatorio con el programa NOHARM (Fraser, 1988) para comprobar la unidimensionalidad del TECLE; en segundo lugar se efectuó el análisis de Impacto y posteriormente el estudio de Funcionamiento Diferencial de los Items (FID) con el procedimiento de Mantel-Haenszel (MH).

Análisis de dimensionalidad

La unidimensionalidad se evalúa con el programa NOHARM de Fraser (1988) y el ajuste del modelo se valoró con el índice de Tanaka. La literatura recoge que valores por encima a 0,90 evidencian buen ajuste.

Análisis de Impacto

Para la detección del impacto se investigó la relación entre las variables género y la respuesta a cada uno de los 64 ítems de la prueba a través del procedimiento de contraste de hipótesis sobre dos proporciones independientes.

Análisis de FDI

Para la detección del FDI se empleó el procedimiento de Mantel-Haenszel, “desarrollado por el autor en el año 1959 y aplicado por Holland y Thayer en 1988” (citado en Andriola, 2002, p73). Para realizar estos cálculos se utilizó el programa computarizado MHDIF (Fidalgo, 1994) para detectar funcionamiento diferencial en los ítems uniforme con el procedimiento de Mantel-Haenszel. Este programa, funciona en dos etapas, en la primera se calcula la puntuación total con todos los ítems que componen la prueba y se evalúa el índice de funcionamiento diferencial. En una segunda etapa, se elimina los ítems que en la primera etapa presentaban FDI. Por último, se vuelven a realizar los cálculos de los índices de FDI (Elosúa & López, 1999). El método consiste en comparar las frecuencias observadas y esperadas de aciertos y errores entre los grupos de referencia y focal, en los distintos niveles elegidos de la habilidad estudiada por el investigador. Generalmente se llama grupo focal al grupo de interés y grupo de referencia al grupo con el que se va a comparar el de interés, normalmente el grupo mayoritario (Escorial & Navas, 2006). El método estadístico de Mantel-Haenszel (MH) es uno de los métodos más utilizados por ser económico, sencillo y se lo pueda calcular sin necesidad de utilizar un paquete estadístico (Fidalgo, 1996). Está basado en tablas de contingencia, se necesita el puntaje de cada sujeto a cada ítem en una modalidad dicotómica uno (acierto) y cero (error), y tener la puntuación total del sujeto en el test, que se puede resumir en la cantidad o suma de puntajes igual a uno en todos los ítems. Todas las personas con igual nivel de habilidad en la variable estudiada deben tener la misma probabilidad de acertar el ítem. Se divide la puntuación total de los sujetos en la prueba en un determinado número de grupos, supondremos que todas las personas que están en una determinada categoría tendrían un mismo nivel de habilidad. Se construyen K tablas de contingencia 2 x 2 (bidimensional) para cada ítems que conforman el test. La tabla queda conformada por dos filas correspondiente al grupo de referencia y la segunda para el grupo focal, y dos columnas, la primera: número de aciertos y la segunda: número de errores en cada clase para cada uno de los ítems del test, como aparecen en la Tabla 1 (Donoghue & Allen, 1993, p.133):

Tabla 1

Tabla de contingencia 2x2 para el nivel de puntuación K

		Respuesta al ítem		
Grupo		Aciertos (1)	Error (0)	Total
Referencia	(G_R)	A_K	B_K	N_{RK}
Focal	(G_F)	C_K	D_K	N_{FK}
Total		N_{1K}	N_{0K}	N_K

El procedimiento *Mantel-Haenszel* se basa en el supuesto de que si el ítem no presenta FDI, entonces, el cociente o razón entre la cantidad de personas que aciertan (A_K) el ítem y las que fracasan (B_K) debe ser el mismo que el cociente entre las personas que lo aciertan y las que fracasan en el otro grupo. Esta relación de igualdad tiene que ser cierta para las distintas clases establecidas a priori por el investigador (Fidalgo, 1996, p.385):

$$H_O : (A_K / B_K) = \alpha (C_K / D_K)$$

siendo α = 1 para todo K.

$$H_1 : (A_K / B_K) = \alpha (C_K / D_K)$$

siendo α ≠ 1 para todo K.

Cuando Alfa es igual a 1, se puede concluir que la probabilidad de acertar el ítem en el grupo de referencia y la probabilidad de fracasar el ítem en el grupo Focal son iguales, y por lo tanto no hay DIF. Cuando el alfa es distinta de 1 significa que existe DIF.

Este cociente de razones (odds ratio) puede tomar valores entre 0 e infinito. Cuando el resultado del cociente es igual a 1, podemos interpretar que no hay diferencias entre los grupos (masculino – femenino) que estamos evaluando, podemos concluir que los ítems que presenten un α = 1 no presentan DIF. Como el cociente puede tomar valores positivos o negativos, siempre que el valor absoluto sea mayor que 1 estaremos frente a un ítem que presenta DIF, en segundo lugar estudiaremos el signo. Si α > 1 evidencia que ese ítem favorece al grupo de Referencia sobre el grupo Focal, pero si α < 1 significa que ese ítem favorece al grupo Focal sobre el grupo de Referencia.

Por último, mediante el estadístico de Holland y Thayer (citado en Gómez & Navas, 1988, p.334) la hipótesis nula de no DIF se debe someter a comprobación utilizando las siguientes formulas (Fidalgo, 1996, p.386):

$$\chi^2_{MH} = \frac{\left(\left|\sum_{K=1}^{m} A_K - \sum_{K=1}^{m} E(A_K)\right| - 0.5\right)^2}{\sum_{K=1}^{K} Var(A_K)}$$

$$Var(A_K) = \frac{N_{RK} N_{FK} N_{1k} N_{0k}}{N_K^2 (N_K - 1)}$$

El estadístico χ^2_{MH} sigue una distribución χ^2 con un grado de libertad, entonces para saber si existe FDI, comparamos: $\chi^2_{MH} \geq \chi^2_{1-\alpha}$, si esto ocurre entonces el ítem presenta FDI al nivel de significación (Fidalgo, 1996, p.386):

$$\hat{\alpha}_{MH} = \frac{\sum_{K=1}^{m} A_K D_K / N_K}{\sum_{K=1}^{m} B_K C_K / N_K}$$

Por último, la métrica con la cual se expresa el indicador calculado ( ) varía entre 0 e infinito, Holland y Thayer en 1985 transforma la métrica a una escala delta (citado en Fidalgo, 1996):

$$MHD - FDI = -2035 \ln\left[\hat{\alpha}_{MH}\right]$$

Gracias a la transformación propuesta por Holland y Thayer DE 1986 la ausencia de FDI es indicada por valores próximos a cero, los valores positivos significa que el ítem favorece al grupo focal y favorece al grupo de referencia cuando el valor es negativo (Fidalgo, 1996).

A continuación se compara los índices obtenidos con la nueva escala (MHD – FDI) con las recomendaciones de la Educational Testing Service (ETS) que propone una escala jerárquica para los distintos valores del coeficiente ∆MH = MH D -FDI de acuerdo con su magnitud: “categoría A (MH D - FDI < 1: FDI despreciable o irrelevante), categoría B (MH D - FDI ≤ 1,0 o 1,5 > MH D - FDI: FDI moderado) y categoría C (MH D - FDI > 1.5: FDI severo)” (citado en Andriola, 2002, p.45). Los ítems clasificados como A según la Educational Testing Service (ETS) se consideran que exhiben poco o ningún FDI y son considerados apropiados para el uso en la construcción de prueba. Los ítems clasificados como B son usados sólo si ningún ítem clasificado como A está disponible para satisfacer los

requisitos del contenido de la prueba. Los ítems clasificados como C, si los expertos consideran que los contenidos son esenciales según las especificaciones del test, tendrían que estudiarlos cuidadosamente para decidir dejarlos en el test (Clauser & Mazor, 1998).

Por último, para Fidalgo & Ferreres (2002) los procedimientos de análisis de FDI deben cumplir con tres supuestos para poder aplicarse: 1. las muestras utilizadas para evaluar el FDI deben ser muestras representativas de las poblaciones de interés; 2. la mayor parte de los ítems que componen el test deben ser ítems válidos y 3. el costo de cometer un error tipo I es mayor que el costo de cometer un error de tipo II.

RESULTADOS

La Tabla 2 resume los resultados obtenidos con la prueba estadística *T de student* de comparación de medias, puntaje total en el TECLE según la variable género, se llega a la conclusión que no existen diferencias significativas entre las medias en el puntaje total en el TECLE en el grupo de referencia (niñas) y el grupo focal (varones). La tabla 3 resume las medias, desvíos típicos por curso y los resultados de la comparación de medias del TECLE por curso.

Tabla 2

Diferencias grupales en el Puntaje Total en el TECLE según sexo

	Femenino (Grupo Referencia)		Masculino (Grupo Focal)		
TECLE	*M*	*DE*	*M*	*DE*	*t*
Puntaje total	58.08	5.82	58.10	5.07	-0.70

$p \leq .05$

Análisis de unidimensionalidad

Se evaluó mediante el programa NOHARM de Fraser (1988) y el ajuste del modelo se valoró con el índice de Tanaka. La literatura recoge que valores por encima a .90 evidencian buen ajuste. Para el análisis de la dimensionalidad se utilizó una muestra de 1159 alumnos, compuestas por alumnos de cuarto a sexto año de primaria de ambos sexo. El análisis de la dimensionalidad utilizando el programa para ajustar modelos unidimensionales de Fraser (1988) evidencia que los 64 ítems del test miden una sola dimensión (habilidad de eficacia lectora). Se obtiene un índice Tanaka bastante alto según la literatura (Tanaka index of goodness of fit = .98). Se puede concluir que TECLE cumple con el supuesto de unidimensionalidad.

Análisis de Impacto

La Tabla 3 resume los resultados obtenidos al evaluar el impacto a nivel del ítem, muestra los valores obtenidos para el estadístico Z_i en cada uno de los ítems del TECLE. Indicando en negrita los que presentaban una relación estadísticamente significativa ($p \leq .05$) con la variable género. Este estadístico pone a prueba la hipótesis nula de igualdad de proporción de aciertos entre los grupos de referencia y focal.

Tabla 3

Resultados de ítems con impacto

Ítem	Z
6	-2.108*
10	2.359*
19	3.060*
28	-2.075*
45	-2.245*
64	2.189*

$p \leq .05$

Tal como se puede observar en la Tabla 4 los ítems 6, 10, 19, 28, 45 y 64 evidencian impacto a un nivel de significación del 5%. Mientras que los ítems 6, 28 y 45 presentaron un valor Z negativo ($Z \leq -1.64$) indicando que la proporción de acierto es mayor para los hombres que para las mujeres. A su vez, los ítems 10, 19 y 64 presentan un valor Z positivo ($Z \geq 1.64$), indicando que la proporción de éxito para las mujeres en esos ítems son mayores que en los hombres.

Tabla 4

Ítem con impacto significativo por género

ítem	Favorable a los hombres	opciones			
6	Han atrapado un castor,,,	caslor	cantos	caspor	castor
28	El capitán mando subir el ,,,	perizcopio	periférico	periscotio	periscopio
45	Esa niña pequeña no dejó de llorar en toda la ...	node	nolle	nota	noche
ítem	**Favorable a las mujeres**	**opciones**			
10	Está viendo la ...	tetevisión	teléfono	televisión	terevisión
19	Entre las flores hay un ...	tudipán	tufipán	tutora	tulipán
64	Ten mucho cuidado para que la máquina no caiga al agua, ya que no es ...	sumergible	sumengible	sunergible	sustituirle

Análisis de FDI

Ante de mostrar los resultados, es importante considerar que el estudio del FDI debe cumplir con los tres supuestos planteados por Fidalgo & Ferreres (2002). Nosotros solamente estudiamos los dos primeros supuestos: las muestras utilizadas para evaluar el FDI son representativas de las poblaciones de interés y la mayor parte de los ítems que componen el test son válidos. Ante el primer supuesto, el TECLE va dirigido a la población de alumnos de primaria de colegios católicos de Montevideo, y para obtener la representatividad se utilizó el muestreo aleatorio estratificado. En relación al segundo supuesto, el método Mantel-Haenszel utiliza para estimar el nivel de la variable medida el puntaje total en el test, por lo tanto, si el test no es unidimensional se estaría atentando contra la validez del estudio. Los resultados de los estudios de unidimensionalidad con el programa NOHARM evidencian que la prueba de eficacia lectora es unidimensional, cumpliéndose así el segundo supuesto.

La Tabla 5 presenta los estadísticos descriptivos del grupo de referencia y del grupo focal.

Tabla 5

Resumen del análisis del FDI de los ítems en el TECLE

Ítems	$\hat{\alpha}_{MH}$	$-2.35\ln[\hat{\alpha}_{MH}]$	χ^2_{MH}	ETS
6	0.25	3.27	4.84	C
19	1.80	-1.38	8.68	B
28	0.71	0.82	6.37	A
64	1.64	-1.16	4.69	B

$p \leq .05$

En la Tabla 6 presenta solamente los valores de MH en los ítems del TECLE cuando se rechaza la hipótesis nula de ausencia de FDI con un nivel de significación de .05, pudiéndose concluir que el ítem funciona diferencialmente entre los dos grupos.

Tabla 6

Estadísticos descriptivos para niños y niñas

	Total	Grupo Focal	Grupo Referencia
M	58.09	58.10	58.08
DE	5.46	5.07	5.82
N	1159	574	585

En la Tabla 6 se puede ver los resultados de la aplicación del procedimiento de Mantel-Haenszel. En este procedimiento se utilizó el programa computacional MHDIF (Fidalgo, 1994) y como variable de equiparación se utilizó la puntuación total observada en la prueba. La Tabla 6 presenta los valores de MH para carda uno de los ítems del TECLE. Los resultados del análisis del funcionamiento diferencial de los ítems con el programa MHDIF encontraron que según el estadístico MH el porcentaje de ítems con FDI es del 6,25%.

En la Tabla 7 se muestra la detección de los ítems con FDI, a partir del análisis empírico con los 64 ítems, se detecto un 6,25% de ítems con FDI (4 ítems), de los cuales la mitad resultó ser favorable al grupo de referencia de las niñas y la otra mitad al grupo focal de los varones.

Tabla 7

Ítems con FDI significativo

Ítem	Favorable a los hombres	Opciones de respuesta			
6	Han atrapado un castor…	caslor	cantos	caspor	castor
28	El capitán mando subir el …	perizcopio	periférico	periscotio	periscopio

Ítem	Favorable a las mujeres	Opciones de respuesta			
19	Entre las flores hay un …	tudipán	tufipán	tutora	tulipán
64	Ten mucho cuidado para que la máquina no caiga al agua, ya que no es …				

Los cuatros ítems con sospecha de FDI (6, 19, 28 y 64) se compararán, el estadístico MH D - FDI de Holland y Thayer en 1986 (citado en Fidalgo, 1996) con las recomendaciones de la Educational Testing Service (ETS), que propone una escala jerárquica para los distintos valores del coeficiente ∆MH = MH D -FDI de acuerdo con su magnitud (ver tabla 5): categoría: A (MH D - FDI < 1: FDI despreciable o irrelevante), categoría B (MH D - FDI ≤ 1,0 o 1,5 > MH D - FD:I FDI moderado) y categoría C (MH D - FDI > 1.5: FDI severo) (citado en Andriola, 2002, p.45).

CONCLUSIONES

El objetivo de este estudio era comprobar si las diferencias de género observadas en los ítems que evalúa el TECLE son diferencias verdaderas o están provocadas por un funcionamiento diferencial de los ítems que componen el test.

Los estudios empíricos sobre el funcionamiento diferencial de los ítems deben cumplir con tres supuestos: muestras representativas; unidimensionalidad del test y costo de cometer error tipo 1 es mayor que el de tipo 2 (Fidalgo & Ferreres, 2002), nosotros nos abocamos a los dos primeros. Los resultados mostraron que las muestras utilizadas para evaluar el FDI son estadísticamente representativas de la población de interés y fue tenido en cuenta desde la planificación del test. En relación al segundo supuesto, los procedimientos de detección de FDI necesitan comparar los puntajes de los sujetos pertenecientes a distintas poblaciones pero igualados en el nivel de habilidad evaluada. O sea, se analizan los ítems usando la puntuación total del test sin usar ningún criterio externo al mismo,

por este motivo, la mayor parte de los estudios empíricos asumen que solo se mide un constructo o que específicamente el test sea unidimensional. En relación al estudio de dimensionalidad realizado con el programa NOHARM, los resultados muestran en primer lugar que el TECLE es una prueba unidimensional confirmando que los ítems son válidos cuando miden lo que el constructo pretende medir. Una vez establecida la unidimensionalidad, pasamos a analizar los resultados de los estudios de Impacto y FDI.

Luego de realizar los análisis de Impacto - FDI, podemos observar cada uno de estos estudios en forma independiente. Podemos afirmar que los ítems 6, 10, 19, 28, 45 y 64 son candidatos a presentar impacto; y de éstos, los ítems 6, 19, 28 y 64, también pueden presentar FDI. Cuando interaccionamos ambos estudios, según "Holland y Wainer, podemos seguir la hipótesis que todos los ítems midiesen sólo la habilidad pretendida (eficacia lectora), cualquier diferencia entre grupos reflejaría sólo el impacto de ese ítem, no el sesgo" (citado en Fidalgo, 1995, p.238). Entonces, de los 6 ítems del TECLE que presentaron impacto (6, 10, 19, 28, 45 y 64), cuatro de éstos presentaron FDI (6, 19, 28 y 45), pudiéndose concluir entonces, que solamente los ítems 10 y 45 no presentan FDI pero si impacto, y en los cuatro restantes (6, 19, 28 y 64) presentan funcionamiento diferencial cuando se lo iguala por nivel de habilidad en el puntaje total del TECLE. En conclusión, los ítems 10 y 45 sin FDI presentan impacto, donde la probabilidad de acierto en esos dos ítems, viene dada por el nivel de habilidad de los sujetos. Las diferencias encontradas entre los varones y las niñas en los ítems 10 y 45 se explicarían por la diferencia de habilidad entre uno y otro sexo.

Mientras, que en los ítems 6, 19, 28 y 64 que presentan sospecha de FDI, cuando se comparan las respuestas al ítem, (probabilidades de aciertos en los cuatro ítems estudiados), entre los grupos, únicamente cuando estos han sido igualados en el nivel de habilidad mediante un criterio de igualación, no se puede atribuir a diferencias de habilidad ya que el método para calcular el FDI previamente se equipara la habilidad (Gómez-Benito et al. 2010).

Una vez detectado la posibilidad de FDI en estos cuatro ítems, según Holland y Thayer se puede inferir que las diferencias encontradas en los resultados conseguidos por distintos alumnos, nivelados en la habilidad de eficacia lectora con el TECLE según género: que los ítems 6, 9, 28 y 64 presentan funcionamiento diferencial y no sesgo (citado en Gómez-Benito et al., 2010).

Al comparar los índices obtenidos con la nueva escala (MH D – FDI) con las recomendaciones de la Educational Testing Service (ETS) que propone una escala jerárquica para los distintos valores del coeficiente ∆MH = MH D –FDI (Andriola, 2002, p.45), podemos concluir que de los cuatro ítems con sospecha de FDI, uno presenta un FDI severo (6), dos moderado (19, 64) y uno un FDI irrelevante (28). En este, punto, es necesario recordar que los procedimientos de detección del FDI fueron creados para detectar posibles sesgos, se necesita de la opinión de expertos en el área de la psicología cognitiva de la lectura, para poder explicar las posibles causas de un FDI detectado por estas técnicas o procedimientos (Hidalgo, Galindo, Inglés, Campoy & Ortiz, 1999; Elosua & Torres, 2000). No podemos afirmar que exista un FDI solamente por haber obtenido índices de funcionamiento diferencial en cuatro ítems luego de aplicar el procedimiento MH, no podemos concluir necesariamente que existe sesgo contra uno de los grupos. En este punto nos vemos obligados a continuar con análisis cualitativos con expertos en el tema de eficacia lectora, para revisar nuevamente la validez de contenido para detectar posibles errores sistemáticos de medida que puedan estar atentando contra la validez del instrumento (Elosua et al., 2000).

Dado lo anterior, se recomienda siguiendo los criterios de Hidalgo et al. (1999); Elosua et al. (2000), continuar con los estudios con jueces expertos en el tema para despejar la sospecha de FDI severo y moderado, para evaluar si hay que eliminar estos ítems de la prueba de eficacia lectora (TECLE). En especial se recomienda que el ítem 6 que presenta un FDI severo, tenga que ser estudiado por expertos (Clauser & Mazor, 1998) para evaluar su permanencia en la prueba o continuar en una investigación futura, en busca de las posibles causas que terminen confirmando o no la presencia de sesgo en ese ítem (Gómez-Benito et al., 2010).

Los resultados obtenidos evidencian que no es preciso realizar una baremación del TECLE por sexo. Al mismo tiempo habría que considerar si se justifica realizar una purificación de la escala eliminando los ítems con sospecha de FDI, dado que sólo dos ítems presentan un FDI moderado y solamente uno con FDI severo.

REFERENCIAS

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

Andriola, W. (2002). *Detecciòn del Funcionamiento Diferencial del Item (DIF) en Tests de Rendimiento. Aportaciones Teóricas y Metodológicas.* (Tesis doctoral, Universidad Complutense de Madrid, Madrid, España). Recuperada de http://www.ucm.es/BUCM/tesis/edu/ucm-t26457.pdf

Barbero, M., Vila, E. & Holgado, F. (2008). La adaptación de los tests en estudios comparativos interculturales. *Acción Psicológica*, 5 (2), p. 7-16. Recuperado el 20 de mayo de 2011 de http://e-spacio.uned.es/fez/view.php?pid=bibliuned:AccionPsicologica2008-2-0001

Barbero, M. I., Vila, E. & Suárez, J. (2006). *Psicometría.* Madrid: UNED.

Carretero-Dios, H. & Pérez, C. (2007). Normas para el desarrollo y revisión de estudios instrumentales: consideraciones sobre la selección de tests en la investigación psicológica. *International Journal of Clinical and Health Psychology, 7* (3), 863-882.

Clauser, B. E. and K. M. Mazor (1998). "Using Statistical Procedures to Identify Differentially Functioning Test Items." *Educational Measurement: Issues and Practice* 17(1): 31-44.

Cuadro, A. (1999). *Evaluación de la lectura y de una propuesta de intervención en alumnos de primaria.* Tesis de maestría no publicada, Universidad Católica del Uruguay, Montevideo, Uruguay.

Cuadro, A., Costa, D., Trías, D. & Ponce de León, P. (2009). *Evaluación del nivel lector. Test de eficacia lectora (TECLE) de J. Marín y M. Carrillo. Manual técnico.* Montevideo: Prensa Médica.

Donoghue, J. & Allen, N. (1993). Thin Versus Thick Matching in the Mantel-Haenszel Procedure for Detecting DIF. *Journal of Educational Statistics, 18* (2).

Elosúa, P. & López, A. (1999). Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales. *Psicológica, 20*, 23-40.

Elosua, P., López, A. & Torres, E. (2000). Desarrollos didácticos y funcionamiento diferencial de los ítems. Problemas inherentes a toda investigación empírica sobre sesgo. *Psicothema, 12* (2), 198-202.

Escorial, N. & Navas, M. J. (2006). Análisis de la variable género en las escalas del EDTC mediante técnicas de funcionamiento diferencial de los ítems. *Psicothema, 18* (2), 319-325.

Fidalgo, A. (1995). Differential Item Functioning [Revisión del libro Differential Item Functioning, por P.W. Holland y H. Wainer, Eds.]. *Psicothema, 7* (1), 237-241.

Fidalgo, A. (1996). *Funcionamiento diferencial de los ítems.* En José Muñiz (Coord.), *Psicometría*. Madrid: Editorial Universitas, S.A.

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

Fidalgo, A. & Ferreres, D. (2002). Supuestos y consideraciones en los estudios empíricos sobre el funcionamiento diferencial de los ítems. *Psicothema, 14* (2), 491-496.

Fraser, C. (1988). *NOHARM. Computer software and manual*. Armidale, New South Wales, Australia: author.

Gómez, J. (septiembre, 2005). *Simposio: Funcionamiento diferencial de los ítems.* Trabajo presentado en la IX Congreso de Metodologìa de las Ciencias Sociales y de la Salud, Granada, España. Resúmen recuperado de http://www.ugr.es/~cmetodo/pdf/simposio/simposio_gomez_benito.pdf

Gómez, J. & Hidalgo, M. D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: una revisión metodológica. *Anuario de Psicología. Facultad de Psicología. Universidad de Murcia, 74*, 3-32.

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

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

Gründel, M. (2010). Estudio de Impacto y Funcionamiento Diferencial del los Ítems en la Prueba de Eficacia Lectora (TECLE). Memoria de grado no publicada de psicología, Universidad Católica del Uruguay – Montevideo. Uruguay.

Gründel, M. & Costa, D. (2010, octubre). Estudio de Impacto y Funcionamiento Diferencial del los Ítems en la Prueba de Eficacia Lectora (TECLE). Póster presentado al III Simposio de Investigación en Psicología de la Universidad Católica del Uruguay, Montevideo, Uruguay.

Hidalgo, M. D., Galindo, F., Inglés, C. J., Campoy, G. & Ortiz, B. (1999). Estudio del funcionamiento diferencial de los ítems en una escala de habilidades sociales para adolescentes [Versión electrónica], *Anales de psicología, 15*, 2, p. 333-345. Recuperado el 20 de mayo del 2011, de http://www.um.es/analesps/v15/v15_2pdf/17v98_14mdhidalg.PDF

International Test Commission (2010). International Test Commission Guidelines for Translating and Adapting Tests. http://www.intestcom.org

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

Martínez, M. R., Hernández, M. J. & Hernández, M. V. (2006). *Psicometría*. Madrid: Alianza editorial.

Muñiz, J. (coordinador) (1996). *Psicometría.* Madrid: Editorial Universitas, S.A.

Muniz, J. & Hambleton, R. (1996). Directrices para la traducción y adaptación de los tests. *Papeles del Psicólogo, 66*, 75-84.

Santisteban, C. (2009) *Principios de psicometría.* Madrid: Síntesis.

Para citar este articulo: Costa Ball, D., Grundel, M. & Cuadro, A. (2011). Impacto y funcionamiento diferencial de los ítems respecto al género en el Test de Eficiencia Lectora (TECLE). *Ciencias Psicológicas V* (1): 47-57.

Recibido: 02/2011
Revisado: 03/2011
Aceptado: 04/2011