Funcionamiento Diferencial de los ítems de la prueba Ser Bachiller 2017, según sexo

Resumen

El Instituto Nacional de Evaluación Educativa, Ineval utiliza pruebas estandarizadas para evaluar los logros académicos de los estudiantes que participan en el Sistema Nacional de Educación. Las mismas son insumos que deben ser equitativos sin embargo muchas veces esto no se cumple ya que los ítems de las pruebas pueden presentar Funcionamiento Diferencial (DIF) perjudicando a un grupo de sustentantes en específico. En este estudio se evalúa el instrumento de evaluación estandarizado del Ineval, Ser Bachiller 2017 a través del método Mantel-Haenszel y regresión logística. La detección del DIF se realiza en los ítems utilizados en los regímenes Costa y Sierra del Ecuador. Los resultados de la evaluación Ser Bachiller 2017 corroboran la hipótesis de una brecha entre los puntajes obtenidos por los y las estudiantes en todos los dominios considerados; no obstante, el modelo refleja que solo el 2,3% de los ítems en la Costa y 4,1% de los ítems en la Sierra presentan DIF elevado o moderado. Se concluye que la brecha no responde a la estructura interna del instrumento estandarizado del Ineval.

Palabras claves: Pruebas estandarizadas, Educación, sesgo, Funcionamiento Diferencial del Item, Mantel-Haenszel, Regresión logística, logro académico.

Abstract

The National Institute for Educational Assessment, Ineval makes use of standardized tests when assessing the student's performance in the Ecuadorian Educational System. The latter are considered as inputs which are meant to be unbiased and provide equal opportunities. However, this is not always true as standardized tests might include Differential Item Functioning (DIF) favouring a specific group of students. This document evaluates DIF at Ineval's standardized assessment (Ser Bachiller 2017) using two methods: logistic regression and Mantel-Haenszel. The DIF detection was performed for the items used in the assessment of students in the Andean and Coastal Region. Using the national assessment test data, Ser Bachiller 2017, the results corroborate the gender gap in math test scores; nevertheless, the model shows that only 2,3% of the items in the Coastal Region and 4,3% of the items the Andean Region present high or moderate DIF. The study concludes the gap is not related to the internal structure of Ineval's standardized test.

Key words: Standardized tests, Education, bias, Differential Item Functioning, Mantel-Haenszel, Logistic regression, student's performance.

1. Introducción

En Ecuador, la Ley Orgánica de Educación Intercultural (LOEI) en el artículo 68, establece que el Instituto Nacional de Evaluación Educativa (Ineval) realizará las evaluación integral interna y externa del Sistema Nacional de Educación y establecerá los indicadores de calidad de la educación; que serán aplicados a través de la evaluación de componentes como el rendimiento académico de estudiantes, autoridades educativas, docentes, etc. En este marco, con la finalidad de determinar el nivel de logro de los aprendizajes de los estudiantes, el Ineval implementa la evaluación nacional Ser Bachiller que tiene como propósito medir "… el desarrollo de las aptitudes y destrezas que los estudiantes deben alcanzar al culminar la educación obligatoria" (Ineval, 2017).

Así, el Ineval con el fin de asegurar la calidad dentro de sus evaluaciones utiliza pruebas educativas estandarizadas. Las mismas son un instrumento de evaluación diseñado para obtener información comparable y objetiva sobre el rendimiento de los sustentantes. Esta prueba se aplica, corrige y puntúa de manera homogénea inter-temporalmente, para asegurar la comparabilidad de los resultados (MIDE UC, 2012). En este sentido, los resultados que se desprenden de estas pruebas son insumos cada vez más aceptados y utilizados por la imparcialidad, transparencia y facilidad logística para medir con alto grado de precisión la eficacia de las políticas públicas educativas (López, Sánchez, Espinosa, & Carmona, 2013).

Debido a la relevancia y frecuencia de uso de los instrumentos estandarizados en la toma de decisiones, es imprescindible generar evidencia empírica que proporcione sustento sobre la calidad de las interpretaciones de sus resultados (i.e validez) (Ravela, Wolfe, Valverde, & Esquivel, 2000). En este sentido, unos de los problemas que puede comprometer la validez de una prueba estandarizada es la posibilidad de que las diferencias en la calificación, de los sustentantes que provienen de un grupo específico, no responden a diferencias auténticas del conocimiento, sino a un diseño de los ítems que afecta su probabilidad de acierto (Camilli & Shepard, 1994). A este fenómeno se lo conoce como Funcionamiento Diferencial del Ítem.

Las consecuencias de que un ítem presente DIF son numerosas y afectan a los sustentantes, a la comunidad educativa y a los hacedores de política pública. Los efectos más visibles son: bajos rendimientos de las minorías sociales, inequitativo acceso a la educación superior y diseño equivocados de políticas educativas que no responden a las necesidades poblacionales (Rojas, 2002).

Los análisis de validez de pruebas estandarizados se han centrado en la detección del sesgo a través del análisis del funcionamiento diferencial de los ítems. Por lo

general, estos se concentran en las diferencias en el rendimiento por la variable sexo (Matus, Stevenson, & Mirella, 2010; Silva, 2014).
Esta brecha entre los resultados de hombres y mujeres ha sido ampliamente estudiada en los campos matemático y verbal por lo que resulta interesante realizar un análisis que se concentre en los mismos dominios. (Dalit Conitni, 2016)

Por tanto, este estudio busca analizar si las pruebas implementadas contienen en sus ítems Funcionamiento Diferencial, es decir, si el instrumento de evaluación Ineval garantiza las mismas oportunidades de obtener una buena calificación a todos los participantes, o por el contrario, existe la presencia de sesgo en los ítems de evaluación que inadvertidamente favorecen o perjudican a los o las estudiantes. Los resultados de este estudio por una parte permitirán la afinación de ítems e instrumentos para garantizar la igualdad y equidad en los resultados de los sustentantes (Allalouf, 1998) y por otra parte repercutirán en la generación de información de calidad para la toma de decisiones de política pública.

En la sección 2 de este estudio se describen los objetivos propuestos para este estudio. En la sección 3 se realiza una breve revisión de la literatura sobre el Funcionamiento Diferencial de ítems, experiencias en otros estudios y otros casos de aplicación. Igualmente, en la sección 4 se describe de manera más profunda en qué consiste la técnica DIF, seguida de la sección 5 donde se encuentran los resultados del estudio. En la sección 6 se ha preparado una discusión sobre las metodologías utilizadas considerando sus fortalezas y debilidades. Finalmente, en la sección 7 y 8 se incluyen conclusiones y recomendaciones.

2. Objetivos

2.1 Objetivo general

Determinar la existencia del Funcionamiento Diferencial de los ítems (DIF) de la prueba Ser Bachiller 2017, según sexo.

2.2 Objetivos específicos

Identificar aquellos ítems de la prueba Ser Bachiller 2017 que presentan DIF para la variable sexo.

Determinar la tipología del DIF, uniforme y no uniforme, de los ítems de la Prueba Ser bachiller 2017 para la variable sexo

Categorizar el efecto del DIF, si lo hubiera, en los ítems de la Prueba Ser Bachiller 2017 para la variable sexo.

3. Revisión de literatura

A partir de los años ochenta en los Estados Unidos de Norte América, los investigadores comienzan a interesarse por la disparidad de resultados de evaluaciones estandarizadas entre estudiantes blancos y estudiantes negros, hispanos y judíos mismos que presentaban un menor rendimiento. Los estudios planteados trataban de detectar la presencia de sesgo en contra de las minorías. Es así que a partir de la década de los noventa se han ido desarrollando diversas técnicas para detectar la presencia de sesgo en las evaluaciones estandarizadas. Actualmente, el análisis de la presencia de sesgo es un componente clave en la validez de las pruebas estandarizados (Moreira, 2008).

Cuando las pruebas estandarizadas presentan algún tipo de error sistémico, entre ellas, el sesgo significa que de alguna manera perjudican a un determinado grupo de la población, por lo cual los análisis de validez son fundamentales. Dado que la calibración de los ítems se la realiza con toda la población evaluada, estos pueden presentar parámetros aceptables de dificultad y discriminación; sin embargo, si estos presentan sesgo medirán una característica o rasgo diferente entre los grupos evaluados (Hidalgo & Gómez, 2000).

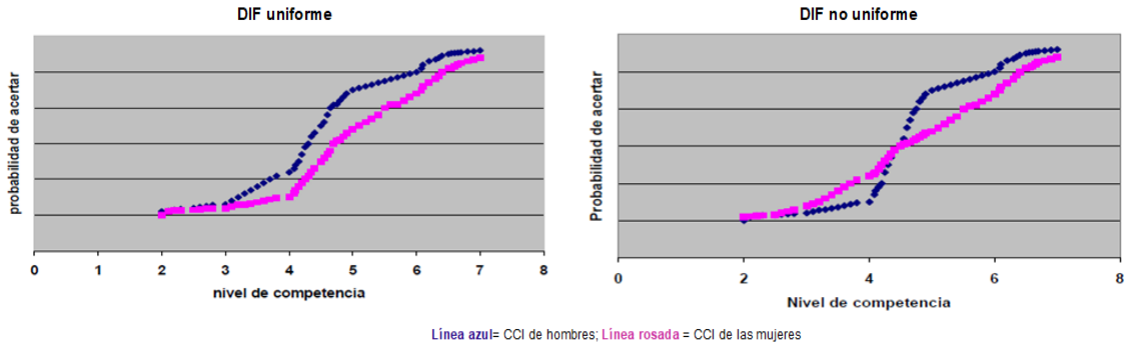
Es así que se entiende al sesgo como una fuente de invalidez y un error sistémico que puede estar presente en todo tipo de evaluación estandarizada. Cuando una prueba presenta sesgo medirá de distinta manera a un determinado grupo sociodemográfico con un mismo nivel de habilidad y creará distorsión de los resultados del rendimiento (Camilli & Shepard , 1994).

Osterlind (1989) enfatiza dos tipos de sesgo: el externo y el interno. El sesgo externo hace referencia al grado en que los resultados de la evaluación obedecen a variables externas que no se miden en la prueba. Por otra parte, el sesgo interno se relaciona con la validez de los ítems de la evaluación para diferentes grupos sociodemográficos. Por otra parte, el sesgo interno se considera un error sistemático porque crea distorsión de los resultados y se lo denomina como Funcionamiento diferencial del Ítem (DIF, por sus siglas en inglés). En este sentido, se dice que un ítem funciona diferencialmente cuando la probabilidad de acertar correctamente es diferente para grupos que tienen igual nivel de habilidad. Por lo general, el grupo mayoritario (grupo de referencia) es el beneficiado y el grupo minoritario (grupo focal) es el perjudicado (Hidalgo & Gómez, 2000).

Según (Mellenbergh, 1982) se distingue dos tipos de DIF, el uniforme y el no uniforme. El funcionamiento diferencial uniforme se da cuando la probabilidad de comprobar correctamente un ítem es mayor para un grupo poblacional que para otro

considerando un mismo nivel de habilidad (Pérez Gil, 2004).En cuanto al DIF no uniforme, se produce cuando la diferencia en la probabilidad de acertar correctamente a un ítem entre el grupo de referencia y el focal no es la misma en todos los niveles de habilidad (Pérez Gil, 2004). A continuación se presenta una representación gráfica del DIF uniforme y no uniforme (ver gráfico 1).

Gráfico 1. Representación gráfica del DIF uniforme y DIF no uniforme



Línea azul= CCI de hombres; Línea rosada = CCI de las mujeres

Fuente y elaboración: DIED-INEVAL, 2018

Como se observa en el gráfico de DIF uniforme, las mujeres con igual nivel y habilidad que los hombres, tienen una menor probabilidad de acertar el ítem. En la figura de DIF no uniforme se observa que las mujeres en los niveles bajos de habilidad tienen mayor probabilidad de acertar correctamente al ítem que los hombres y en los niveles altos de habilidad los hombres tienen mayor probabilidad de acertar al ítem que las mujeres.

Por otra parte, dentro de los análisis de validez relacionados al DIF las variables cambian según el tipo de estudio. La presencia de DIF en los ítems es un fenómeno general que está presente en evaluaciones educativas, psicológicas y profesionales (ingreso o permanencia laboral). Sin embargo, la presencia de sesgo más común suele darse en función del sexo, etnia, idioma, cultura u otra característica sociodemográfica (Gonzales, Rivera, & Padrós, 2015). Un estudio realizado por Gonzáles (2015) muestra, por ejemplo, que la presencia de DIF se da mayormente en la variable edad y cuando se analizó por sexo, las mujeres suelen ser mayormente afectadas; sin embargo, en cuanto al nivel educativo se detectó la menor cantidad de ítems con funcionamiento diferencial.

En las últimas décadas se ha desarrollado varios procedimientos para detectar el funcionamiento diferencial. Andriola (2002) clasifica estos por la dimensionalidad de los ítems; los métodos unidimensionales incluyen el Método del Delta Gráfico, Método del cálculo del área entre las CCI's, Método de las Probabilidades, Método de la comparación de los parámetros de los ítems, Método del Chi-cuadrado de Lord, Método del Chi-cuadrado de Sheuneman, Método del Chi-cuadrado de Pearson o Total, Método de la regresión logística, Método de Mantel-Haenszel, Método

Estandarizado, Método Logístico Iterativo, Método del análisis de las estructuras de covarianza – MAEC y; entre los basados en la multidimensionalidad de los ítems se destacan Método Multidimensional de DIF – MMD, Método de la regresión logística (Swaminathan & Rogers, 1990).

Entre todo ese conjunto de procedimientos los que destacan por su poder de detección del DIF son el método SIBPRUEBA, la regresión logística, el procedimiento de Mantel-Haenszel, los métodos basados en la TRI y los métodos de comparación de áreas (Goodman, Willse, Allen, & Klaric, 2011).

En un estudio de detección del DIF realizado por Ferreres et al. (2000), se evidencia la eficacia de la regresión logística y el procedimiento de Mantel-Haenszel, modificado por Mazor, Clauser y Hambeton (1994), para detectar DIF no uniforme cuando la prueba incluye ítems de dificultad media y de alta discriminación. De esta manera, la regresión logística muestra una alta potencia en la detección de ítems con alta dificultad y con bajos niveles de discriminación. Esto será muy útil en la presente investigación ya que permite cuantificar la medida del DIF y detectar la tipología del mismo (uniforme o no uniforme). Por otra parte, los métodos de detección del DIF y sesgo pueden ser útiles para desarrollar nuevos instrumentos de medida, adaptar pruebas existentes a nuevos contextos de evaluación u otras poblaciones que no se consideraron al crear un instrumento (Juana Gómez-Benito, 2010). En este sentido, el método de Mantel-Haenszel se destaca por su relativa facilidad de cálculo y la regresión logística tiene a su favor la detección del DIF por su tipología.

En el 2012 en la evaluación de matemáticas de largo plazo aplicada a los niños de nueve años en la comparación blancos contra hispanos se encontró que 10 ítems presentaban una indicación débil de DIF que favorecía al grupo focal, 6 presentaban una fuerte indicación de DIF a favor de los hispanos y 7 ítems mostraron mayor dificultad para estos (National Center for Education Statistics, 2016).

Este estudio es un primer paso para analizar los instrumentos del Ineval y se concentrará en un análisis diferencial por sexo. Se debe considerar que el desarrollo de la presente investigación se inspira en el estudio: “Funcionamiento Diferencial de los ítems de la prueba PISA 2009, según sexo para las pruebas de Matemática y Ciencias” realizado por la Agencia de Calidad de educación SIMCE. En este estudio se comprueba que no existe comportamiento diferencial de los items mediante regresiones logísticas en la evaluación internacional PISA, en consecuencia la brecha del rendimiento entre hombres y mujeres se debe a otros factores de contexto de los y las estudiantes (Matus, Stevenson, & Mirella, 2010).

4. Metodología

4.1 Base de datos

Como insumo para el análisis se utilizaron las bases de datos del proceso Ser bachiller 2017 por régimen Costa y Sierra. Se decidió realizar la detección del DIF por separado puesto que la calibración y calificación de los ítems es realizada por régimen y como resultado se tiene un número distinto de ítems para Sierra y Costa.

La base de datos del Régimen Costa cuenta con 256.826 observaciones referentes a estudiantes escolares y no escolares con estado evaluado y que presenten nota de Postulación a la Educación Superior (PES). Se observa que en la base de datos existen resultados de 131.202 (51%) mujeres y 125.624 (49%) hombres. En esta existen un total de 1907 ítems, los cuales se distribuyen de la siguiente manera, 431 para aptitud abstracta, 254 para dominio científico, 479 para el dominio lingüístico, 463 para el dominio matemático y 280 en el dominio social.

En la base del Régimen Sierra se contabilizan 241.629 observaciones que poseen las características de evaluado, son escolares o no escolares y cuentan con puntaje PES. En esta base existen 131.202 (51.09%) mujeres y 125.624 (48.91%) hombres. La base considera un número mayor de ítems que la del Régimen Costa. En la misma, 3840 ítems se utilizó para el Ser bachiller 2017 de Sierra. Por dominios estos se distribuyen de tal manera que 833 pertenecen a la aptitud abstracta, 514 al dominio científico, 1000 al dominio lingüístico, 998 al dominio matemático y 495 al dominio social.

4.2 Método

Para la detección del DIF en los ítems de la evaluación Ser Bachiller 2017 se utilizó dos procedimientos estadísticos, la regresión logística y el método de Mantel-Haenszel. Los mismos buscan asegurar la correcta detección del funcionamiento diferencial de los ítems en la prueba estandarizada.

Al aplicar la regresión logística existen tres estrategias de análisis de la detección de DIF (Montesinos, Benito, & García, 2005):

- La primera es introducir las variables sucesivamente en el modelo, hasta llegar el modelo 3 y comparar los valores pseudo R^2
- La segunda es detectar simultáneamente el DIF uniforme y no uniforme, esta hipótesis se la comprueba comparando el valor de verosimilitud del modelo 1 con el 3.

- La tercera consiste en aplicar solamente el modelo 3 y comprobar la significación del modelo con el estadístico de Wald.

En la presente investigación se utiliza la primera estrategia ya que esta permite cumplir con el segundo objetivo propuesto: clasificar al DIF según su tipología (Uniforme y No uniforme).

Según (Zumbo B. , 1999) se incluyen las variables en la regresión siguiendo un orden específico; seguir una secuencia al introducir las variables es importante puesto que permite determinar el efecto de la variable dicotómica del grupo sociodemográfico y la variación adicional por la interacción entre la puntuación total obtenida por grupo sociodemográfico. Este procedimiento detecta la mejora en el factor de ajuste al incluir las variables en el modelo de regresión logística (Elosua & López-Jáuregui, 2007).
Para la detección del DIF mediante regresión logística se utilizará el siguiente procedimiento (Zumbo B. , 1999):

Modelo 1: En el modelo 1 se introduce la variable condicional (Puntaje total obtenido en el test), y sirve como línea base para comparar con el modelo 2.

$$Ln\frac{P}{1-P} = \beta_0 + \beta_1 TOT$$

Modelo 2 (DIF uniforme): En este se incluye la variable condicional más la variable de agrupación, esta es dicotómica y se la codificará 1 para el grupo de referencia y 2 para el grupo focal) En este paso se evalúa el efecto de la variable del grupo sociodemográfico manteniendo constante el efecto de la variable condicional.

$$Ln\frac{P}{1-P} = \beta_0 + \beta_1 TOT + \beta_2 VI$$

Modelo 3 (DIF no uniforme): Este contiene la variable condicional más la variable de agrupación y el término de interacción (puntuación total por grupo sociodemográfico). En este paso se describe si la diferencia entre el grupo de referencia y el grupo focal en su puntuación obtenida en el ítem varía a través del continuo de la variable latente.

$$Ln\frac{P}{1-P} = \beta_0 + \beta_1 TOT + \beta_2 VI + \beta_3 TOT\ x\ VI$$

Según Zumbo (1999) la identificación del tipo de DIF se puede detectar a partir de las diferencias de los pseudo R^2 obtenidos en el ajuste de los tres modelos, el pseudo R^2 explica la proporción de variación de las respuestas al ítem dado la interacción de las variables independientes en el modelo logístico (Elosua & López-Jáuregui, 2007):

$$\Delta R^2 = R^2(modelo\ 1) - R^2(modelo\ 2)$$
$$\Delta R^2 = R^2(modelo\ 1) - R^2(modelo\ 3)$$
$$\Delta R^2 = R^2(modelo\ 2) - R^2(modelo\ 3)$$

La interacción entre el modelo 2 y el modelo 3 indicaría si existe **DIF no uniforme** ya que este es el resultado de la interacción entre el puntaje total y el grupo sociodemográfico, para la detección del **DIF uniforme** es necesario usar los modelos 1 y 2 como indicadores, este tipo de DIF es un efecto principal significativo para los grupos sociodemográficos (Arias, Arias, Gómez, & Inmaculada, 2013).

Según la literatura existen dos metodologías para detectar la medida del nivel de DIF cuando se usa regresión logística y los pseudo R^2, la de (Zumbo & Thomas, 1997) y la de (Jodoin & Gierl, 2001). Se utilizará la primera categorización en la detección de DIF:

Tabla 1. Interpretación del modelo de regresión logística en base a χ^2 y los pseudo R^2

Categoría	Criterio
A (Ausencia de DIF)	$Si\ el\ test\ \chi^2$ no es significativo al 0,05 o $\Delta R^2 < 0{,}13$
B (DIF Moderado)	$Si\ el\ test\ \chi^2$ es significativo al 0,05 y $0{,}13 < \Delta R^2 < 0{,}26$
C (DIF elevado)	$Si\ el\ test\ \chi^2$ es significativo al 0,05 y $\Delta R^2 \geq 0{,}26$

Fuente: (Zumbo & Thomas, 1997)

El otro método estadístico que se utilizará para confirmar la presencia del DIF es el de Mantel-Haenszel y así también este nos servirá para cuantificar y categorizar el tamaño del efecto DIF. En este procedimiento se compara si la probabilidad de responder correctamente a un ítem es igual en ambos grupos o no, el grupo estándar de comparación que es el grupo de referencia y el grupo focal (minoría) que es el que se supone el afectado por el DIF (Andriola, 2002).

El estadístico se calcula bajo el siguiente procedimiento:

- Se definen dos grupos de comparación, el grupo de referencia (GR) y el grupo focal (GF).
- Se forman grupos de sustentantes en “i” intervalos de acuerdo al puntaje obtenido por estos en la evaluación estandarizada.
- Para cada intervalo “i” se contruye una tabla de contingencia como se muestra a continuación:

Tabla 2. Tabla de contingencia para el intervalo i.

Grupo	**Aciertos (1)**	**Errores (0)**	**Total**
De referencia	a_i	b_i	$a_i + b_i$
Focal	c_i	d_i	$c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	$T_i = a_i + b_i + c_i + d_i$

Fuente: (Chávez & Saade, 2009)

Para el cálculo del coeficiente, Mantel y Haenszel propusieron la siguiente formula $\propto_{MH}$:

$$\propto_{MH} = \frac{\frac{\sum_{i=1}^{S} a_i d_i}{\sum_{i=1}^{S} T_i}}{\frac{\sum_{i=1}^{S} b_i c_i}{\sum_{i=1}^{S} T_i}}$$

Donde:

$a_i =$ Número de sustentantes en el grupo de referencia que contestaron correctamente el ítem

$b_i =$Número de sustentantes en el grupo de referencia que contestaron incorrectamente el ítem

$c_i =$Número de sustentantes en el grupo focal que contestaron correctamente el ítem

$d_i =$ Número de sustentantes en el grupo focal que contestaron incorrectamente el ítem

$T_i =$ Es el total de sujetos en el nivel i de la puntuación observada.

El estadístico $\propto_{MH}$ toma valores de cero a infinito positivo. Por lo tanto se transforma el mismo a una escala simétrica (Andriola, 2002). Esta escala se conoce como delta (δ) y se define de la siguiente manera:

$$\delta = -2{,}35 * Ln(\propto_{MH})$$

El estadístico de Mantel-Haenszel se transforma a una escala logarítmica porque esta permite representar valores de diferentes magnitudes a única escala de medida (Jiménez, 2018). Al calcular el delta, la escala revierte la interpretación del estadístico, donde valores cercanos a cero indican la ausencia de DIF, valores positivos de delta revelan que el ítem favorece al grupo focal y valores negativos que el ítem perjudica al grupo focal o minoritario (Andriola, 2002).

El Educational Testing Service (ETS) interpreta los valores delta de la siguiente manera (Zieky , 2003):

Tabla 3. Interpretación del estadístico delta de Mantel- Haenszel para DIF.

Categoría	Valor de delta	Interpretación
Categoría A	$\|\delta\| < 1$	Ítems con DIF despreciable o irrelevante
Categoría B	$1 \leq \|\delta\| < 1{,}5$	Ítems con DIF moderado
Categoría C	$1{,}5 \leq \|\delta\|$	Ítems con DIF severo

Fuente: (Zieky , 2003)

Entre las principales limitaciones de los procedimientos metodológicos para detectar DIF se encuentran la posibilidad de cometer Error tipo I (falsos positivos) o Error tipo II (no detectar un ítem con DIF), el tamaño de la muestra es otro factor relevante al

momento de detectar DIF. Para no cometer errores tipo I o II el análisis DIF necesita ir acompañado de otros tipos de análisis como el de estructura factorial para descartar como una posible fuente la multidimensionalidad del ítem, el de confiabilidad para garantizar la consistencia y estabilidad de las puntuaciones; así como la técnica de expertos en el área que ayuden a determinar las posibles causas de porque un ítem funciona diferencialmente (Moreira, 2008).

5. Resultados

En la presente sección se presentan los resultados para los regímenes Costa y Sierra, por separado. Al aplicar el análisis estadístico DIF, a través de los métodos propuestos se puede observar los siguientes hallazgos:

5.1 Régimen Costa

5.1.1 Análisis general

Para el régimen Costa los resultados mostraron que de los 1.907 ítems 1.065 (56%) ítems de la prueba Ser Bachiller del año 2.017 presentaron algún tipo de DIF (irrelevante, moderado o severo). Los dominios con mayor número de ítems detectados con DIF son el dominio de aptitud abstracta con 293 ítems (28%), el dominio lingüístico con 291 ítems (27%) y el dominio matemático con 202 ítems (19%).

Tabla 4 Ítems que presentan DIF para el Régimen Costa

Dominio	Ítems con DIF	Porcentaje
Aptitud Abstracta	293	28%
Dominio científico	132	12%
Dominio lingüístico	291	27%
Dominio matemático	202	19%
Dominio social	147	14%
TOTAL	1.065	100%

Fuente y elaboración: DIED-INEVAL, 2018

A partir del análisis empírico en los 1.065 reactivos con DIF, se observó que en todos los dominios evaluados el tipo de DIF mayormente detectado es el de tipo uniforme, es así que el dominio aptitud abstracta presenta el mayor porcentaje de ítems con DIF Uniforme (80%) seguido por el dominio social (78%). Por tipo de DIF No uniforme los dominios científico (37%) y matemático (33%) presentan el mayor porcentaje de ítems con esta tipología de funcionamiento diferencial.

Tabla 5 Ítems que presentan DIF para el Régimen Costa según su tipología

Dominio	No uniforme	Uniforme	Total

Aptitud Abstracta	20%	80%	100%
Dominio científico	37%	63%	100%
Dominio lingüístico	29%	71%	100%
Dominio matemático	33%	67%	100%
Dominio social	22%	78%	100%
TOTAL	27%	73%	100%

Fuente y elaboración: DIED-INEVAL, 2018

Al categorizar los ítems con DIF mediante el estadístico de Mantel-Haenszel con la escala utilizada por la Educational Testing Service (ETS), se observa en todos los dominios predominan los ítems con DIF irrelevante. Los campos evaluados en el Ser bachiller 2017 que presentan mayor porcentaje de ítems con efecto irrelevante son el dominio científico (97%) y aptitud abstracta (96.9%). Con respecto al DIF de efecto moderado el dominio social (4.8%) y lingüístico (4.5%) son los que muestran los mayores porcentajes de ítems con este tipo de funcionamiento diferencial. El dominio lingüístico (0.7%) es el campo con mayor porcentaje de ítems con DIF severo.

Tabla 6 Ítems que presentan DIF para el Régimen Costa según el tamaño

Dominio	DIF irrelevante	DIF Moderado	DIF severo	Total
Aptitud Abstracta	96.9%	2.7%	0.3%	100.0%
Dominio científico	97.0%	3.0%	-	100.0%
Dominio lingüístico	94.8%	4.5%	0.7%	100.0%
Dominio matemático	96.0%	3.5%	0.5%	100.0%
Dominio social	95.2%	4.8%	-	100.0%
TOTAL	**96.0%**	**3.7%**	**0.4%**	**100.0%**

Fuente y elaboración: DIED-INEVAL, 2018

5.1.2 Análisis por dominio

Siguiendo la metodología planteada, los ítems que se analizaron por dominio cumplen los siguientes criterios:

- Que con el método de regresión logística se haya detectado la presencia de DIF.
- Que con el estadístico delta de Mantel-Haenszel los ítems presenten DIF moderado o elevado.
- Que exista significancia del estadístico de Mantel-Haenszel.

En el análisis del DIF, independientemente de la estrategia utilizada, el nivel de confianza se estableció al 95%.

En el Régimen Costa se evaluaron 1.907 ítems y en 43 de ellos que representan el 2% presentaron DIF, los dominios con mayor número de ítems con funcionamiento

diferencial fueron el dominio lingüístico (15 ítems), seguido por aptitud abstracta (9 ítems) y el dominio matemático (8 ítems).

En la tabla 7 se presentan los resultados del análisis DIF por cada uno de los dominios. En el dominio aptitud abstracta, se observa que existe 1 ítem con DIF uniforme de efecto severo y 8 ítems con efecto moderado de los cuales 1 es de tipo No Uniforme y 7 de tipo Uniforme. Para el dominio científico, se observa 4 ítems de efecto moderado de los cuales 2 son de tipo No uniforme y así también 2 son de tipo Uniforme. Para el dominio lingüístico se identifican 13 ítems con efecto moderado de los cuales 9 son de tipo Uniforme y 4 de tipo No Uniforme, así también en este dominio puede observarse que existe 2 ítems con DIF Uniforme de efecto severo. Para el dominio matemático existen 7 ítems con funcionamiento diferencial moderado de los cuales 2 muestran DIF No Uniforme y 5 DIF Uniforme, en este se presenta un solo ítem con DIF Uniforme de efecto severo. Finalmente, en el dominio social, se encontraron 7 ítems de efecto severo de los cuales 1 es de tipo No Uniforme y 6 de tipo Uniforme.

Tabla 7 Ítems que presentan DIF según el efecto y tipo para el Régimen Costa

Dominio	Efecto Moderado		Efecto Severo	Total General
	DIF No uniforme	DIF Uniforme	DIF Uniforme	
Aptitud Abstracta	1	7	1	9
Dominio científico	2	2	-	4
Dominio lingüístico	4	9	2	15
Dominio matemático	2	5	1	8
Dominio social	1	6	-	7
TOTAL	10	29	4	43

Fuente y elaboración: DIED-INEVAL, 2018

5.2 Régimen Sierra

5.2.1 Análisis general

Los resultados mostraron que de los 3.849 ítems evaluados en el Régimen Sierra 1.395 (36%) ítems de la prueba Ser Bachiller del año 2017 presentaron funcionamiento diferencial. De los 1.395 ítems con DIF, 423 (30%) corresponden a la aptitud abstracta, 403 (29%) al dominio lingüístico y 221 (16%) al dominio matemático.

Tabla 8 Ítems que presentan DIF para el Régimen Sierra

Dominio	Ítems con DIF	Porcentaje
Aptitud Abstracta	423	30%
Dominio científico	180	13%
Dominio lingüístico	403	29%
Dominio matemático	221	16%

Dominio social	168	12%
TOTAL	1395	100%

Fuente y elaboración: DIED-INEVAL, 2018

El Régimen Sierra sigue la misma tendencia del tipo DIF detectado en Costa, en este sentido se observa que de los ítems con funcionamiento diferencial la mayoría presenta DIF uniforme, en aptitud abstracta se observa un 82% de ítems con este tipo de DIF y así también en el dominio matemático y lingüístico se evidenció un 74% de ítems de tipología uniforme, respectivamente.

Tabla 9 Ítems que presentan DIF para el Régimen Sierra según su tipología

Dominio	No uniforme	Uniforme	Total
Aptitud Abstracta	18%	82%	100%
Dominio científico	38%	62%	100%
Dominio lingüístico	26%	74%	100%
Dominio matemático	26%	74%	100%
Dominio social	33%	67%	100%
TOTAL	26%	74%	100%

Fuente y elaboración: DIED-INEVAL, 2018

En la prueba Ser bachiller del Régimen Sierra se detectó un mayor número de ítems con DIF irrelevante que en el Régimen Costa. En cuanto al tamaño del DIF para el Régimen Sierra se evidencia que el dominio científico (97.8%) y matemático (91.4%) son los campos evaluados que muestran mayor porcentaje de ítems con DIF de efecto irrelevante. Con respecto al funcionamiento diferencial de tamaño moderado aptitud abstracta (13.2%) y el dominio lingüístico (8.7%) son los reactivos con mayor porcentaje con este tipo de efecto. En relación al DIF severo el dominio aptitud abstracta (3.5%) presenta el mayor porcentaje de ítems con este efecto

Tabla 10 Ítems que presentan DIF para el Régimen Sierra según el tamaño

Dominio	**DIF irrelevante**	**DIF Moderado**	**DIF severo**	**Total**
Aptitud Abstracta	83.2%	13.2%	3.5%	100.0%
Dominio científico	97.8%	1.1%	1.1%	100.0%
Dominio lingüístico	88.1%	8.7%	3.2%	100.0%
Dominio matemático	91.4%	5.4%	3.2%	100.0%
Dominio social	91.1%	7.1%	1.8%	100.0%
TOTAL	**88.7%**	**8.4%**	**2.9%**	**100.0%**

Fuente y elaboración: DIED-INEVAL, 2018

5.2.2 Análisis por dominio

Los ítems que se analizaron por dominio para el Régimen Sierra cumplen los mismos criterios considerados para el Régimen Costa.

En el Régimen Sierra se evaluaron 3.840 ítems, en 157 ítems (4%) se observó la existencia de DIF. Los dominios con mayor número de ítems con funcionamiento diferencial fueron Aptitud abstracta con 71 ítems, dominio lingüístico con 48 ítems y dominio matemático con 19 ítems.

En la tabla 11 se presentan los resultados DIF para cada dominio. En el dominio aptitud abstracta, se detectaron 56 ítems con efecto moderado de los cuales 7 muestran DIF No Uniforme y 49 DIF Uniforme, así también se evidenció 15 ítems de efecto severo de los cuales 1 es No Uniforme y 14 de tipo No Uniforme. El dominio científico es el que presenta menor cantidad de ítems con funcionamiento diferencial, en este se observa 2 ítems con DIF No Uniforme de efecto moderado y 2 ítems con DIF Uniforme con efecto severo. En el dominio lingüístico es el campo con mayor número de ítems con DIF, en este se observa 35 ítems con DIF de efecto moderado y 13 con efecto severo. En el dominio matemático se encontró 12 ítems de efecto moderado de los cuales 1 presenta DIF No Uniforme y 11 DIF Uniforme, así también en este campo existen 7 preguntas con DIF Uniforme de efecto severo. Finalmente, en el dominio social los resultados muestran mayor número de ítems con DIF moderado que severo.

Tabla 11 Ítems que presentan DIF según el efecto y tipo para el Régimen Sierra

Dominio	Efecto moderado		Efecto severo		Total General
	DIF No uniforme	DIF Uniforme	DIF No uniforme	DIF Uniforme	
Aptitud Abstracta	7	49	1	14	71
Dominio científico	2	-	-	2	4
Dominio lingüístico	4	31	2	11	48
Dominio matemático	1	11	-	7	19
Dominio social	4	8	2	1	15
TOTAL	18	99	5	35	157

Fuente y elaboración: DIED-INEVAL, 2018

6. Discusión

La creciente necesidad de evaluar la calidad educativa del Sistema Nacional de Educación que se ha venido dando en los últimos años en Ecuador, exigen que las evaluaciones tengan un alto nivel de precisión en los resultados de las mismas. Estas deben ser confiables, validas, oportunas y objetivas, (Ineval, 2017) y su metodología

de aplicación debe responder a la necesidad de comparar resultados en un determinado grupo sociodemográfico con un mismo nivel de habilidad, sin crear distorsión (sesgo) en los resultados del rendimiento (Camilli & Shepard , 1994).

Cuando el sesgo es interno se considera un error sistémico porque crea distorsión de los resultados y se lo denomina como Funcionamiento diferencial del Ítem (DIF por sus siglas en inglés). No obstante, que los ítems presenten DIF, es un fenómeno general que está presente en evaluaciones educativas, psicológicas y profesionales (ingreso o permanencia laboral), y suele darse comúnmente en función del sexo, etnia, idioma, cultura u otra característica sociodemográfica (Gonzales, Rivera, & Padrós, 2015).

Las consecuencias de que un ítem presente DIF son numerosas y afectan a los sustentantes, a la comunidad educativa y a los hacedores de política pública. Las consecuencias más visibles son: bajos rendimientos de las minorías sociales, inequitativo acceso a la educación superior y diseño equivocados de políticas educativas que no responden a las necesidades poblacionales (Rojas, 2002).

Es así que, en este estudio se busca contrastar la presencia del DIF según sexo para los ítems de la prueba Ser Bachiller 2017, a través de los métodos estadísticos de Mantel-Haenszel y Regresión logística, puesto que, destacan por su alto poder de detección (Goodman, Willse, Allen, & Klaric, 2011; Ferreres et al., 2000) de ítems con DIF por magnitud (irrelevante, moderado y severo) y tipología (uniforme y el no uniforme) respectivamente.

Según la literatura existente se conoce que entre el 10% y 15% de los ítems de los tests estandarizados podrían presentar DIF (Narayanan & Swaminathan, 1994). No obstante, no se aclara que porcentaje de estos ítems presenta DIF moderado o severo.

En tal sentido, en la presente investigación se comprobó que para el Régimen Costa el 2,3% y para Régimen Sierra el 4,1% de ítems presentaron funcionamiento diferencial moderado o severo. Estos porcentajes son aceptables para respaldar la validez de la prueba estandariza “Ser Bachiller”, por lo tanto, no se puede afirmar que los mismos favorezcan o perjudiquen a hombres o mujeres y en consecuencia la brecha en el rendimiento por sexo se atribuiría a factores de contexto que rodean a los estudiantes, ya que factores como el proceso de enseñanza, material didáctico, expectativas de padres y docentes y la malla curricular ayudarían a explicar las causas de este fenómeno (Matus, Stevenson, & Mirella, 2010).

Por otra parte, en el presente estudio se detectó que para el DIF Uniforme existe un mayor número de ítems que favorecen a los hombres al igual que el estudio realizado

por Gonzáles (2015). Caso contrario sucede con el dominio lingüístico que a nivel nacional, las mujeres resultan ser mayormente beneficiadas por los ítems con funcionamiento diferencial de efecto moderado o severo.

7. Conclusiones

El análisis del Funcionamiento Diferencial del ítem es un mecanismo necesario para garantizar evaluaciones más justas, ya que con este procedimiento se asegura que el test estandarizado mide de igual manera a toda la población evaluada y que los mismos tienen la misma probabilidad de acertar al ítem.

Según la literatura revisada se debe utilizar varios métodos estadísticos en la detección del Funcionamiento diferencial con la finalidad de no detectar DIF en ítems correctamente calibrados, esto se comprobó en la presente investigación puesto que cada uno de los métodos difiere en el número de ítems detectados. En este sentido, en la presente investigación los ítems que se recomienda revisar para indagar las posibles causas del DIF, cumplieron los siguientes criterios:

- Que con el método de regresión logística se haya detectado la presencia de DIF.
- Que con el estadístico delta de Mantel-Haenszel los ítems presenten DIF moderado o elevado.
- Que exista significancia del estadístico de Mantel-Haenszel.

Los ítems que cumplieron los criterios antes mencionados representaron un 2,3% para el Régimen Costa y un 4,1% para el Régimen Sierra. En ambos regímenes los dominios con mayor número de ítems detectados con DIF fueron el dominio lingüístico y aptitud abstracta.

Una de las ventajas del procedimiento de regresión logística es que permite detectar la tipología del DIF, en los ítems que se detectó funcionamiento diferencial el tipo mayormente detectado es el DIF Uniforme.

En lo referente a la categorización del tamaño del DIF se detectó que existe un mayor número de ítems con DIF moderado que con DIF severo. Los dominios que presentaron mayor número de ítems con efecto severo fueron el dominio lingüístico y aptitud abstracta para los regímenes Costa y Sierra.

En este estudio se evidenció la existencia de DIF en varios ítems de la prueba Ser bachiller 2017 según la variable sexo tanto para el Régimen Costa como Sierra, en este sentido el elemento sexo se debería tomar en cuenta para el diseño de la prueba y la creación de ítems.

Los resultados de esta investigación pueden dar lugar a un protocolo a seguir en el proceso de creación y calibración de los ítems que permita ahorrar tiempo y dinero a la par que incremente la validez de las pruebas estandarizadas aplicadas por el Ineval.

La baja presencia de ítems con DIF en el instrumento de medición utilizado por el Ineval, no favorece en particular a hombres o mujeres, esto sirve como evidencia suficiente para respaldar la validez del test estandarizado según la variable sexo.

8. Recomendaciones

Para los ítems que presentan funcionamiento diferencial, en especial los de efecto severo se sugiere buscar las posibles fuentes del DIF en la prueba Ser bachiller del año 2017 a través de expertos que confirmen la presencia de elementos dentro de un reactivo que dificulten a alumnos de cierto sexo utilizar toda su habilidad para responderlo correctamente.

Siguiendo las buenas prácticas aplicadas en pruebas como la Evaluación Nacional del Progreso Educativo (NAEP), Ineval debería contar con un comité de desarrolladores de pruebas y especialistas capacitados para emitir juicios sobre si el funcionamiento diferencial de un ítem está o no relacionada injustamente con la pertenencia a un determinado grupo. Los comités deben examinar cuidadosamente cada ítem para determinar si el lenguaje o los contenidos tienden a hacer el ítem más difícil para un grupo identificado de evaluados.

La detección del Funcionamiento diferencial de ítem debería suponer una fase añadida al proceso de calibración de las pruebas estandarizadas del Ineval y así también este debería ser un procedimiento obligatorio cuando se desarrollan nuevos instrumentos de medida, se adaptan pruebas existentes a un nuevo contexto de evaluación o a otras poblaciones que no se tuvieron en cuenta en el momento de crear el instrumento, se adaptan pruebas ya existentes a otras lenguas o culturas, o se validan las inferencias derivadas de las puntuaciones de la evaluación.

Una futura investigación debería considerar un análisis de la existencia de DIF según etnia, tipo de sostenimiento y tipo de área que permita asegurar la validez de los ítems de la prueba estandarizada para diferentes grupos sociodemográficos.

Se recomienda aplicar el análisis DIF al Ser Bachiller 2018 con la finalidad de contrastar si se detecta los mismos ítems con funcionamiento diferencial del Ser Bachiller 2017.

9. Referencias bibliográficas

Andriola, W. (2002). *Detección del funcionamiento diferencial del ítem (DIF) en tests de rendimiento: aportaciones teóricas y metodológicas.* Madrid.

Arias, B., Arias, V., Gómez, L., & Inmaculada, M. (2013). Funcionamiento diferencial del ítem en la evaluación de la sintomatología TDAH en función del género y el formato de calificación. Bogotá, Colombia.

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

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Chávez, C., & Saade, A. (2009). Procedimientos básicos para el análisis de reactivos Cuaderno técnico 8. Ciudad de México, México.

Dalit Conitni, M. L. (2016). The gender gap in mathematics achievement: Evidence from Italian data. *Carlo Alberto Notebooks*, 492.

Elosua, P., & López-Jáuregui, A. (2007). Aplicación de cuatro procedimientos de detección del funcionamiento sobre ítems politómicos. Psicothema, 19(2), 329–336.

Gonzales, F., Rivera, M., & Padrós, F. (2015). Invarianza por Sexo en la Escala de Detección del Trastorno de Ansiedad General. *Actualidades en Psicología*, 141-151.

Goodman, J., Willse, J., Allen, N., & Klaric, J. (2011). Identification of differential item functioning in assessment booklet designs with structurally missing data. Educational and Psychological Measurement, 71(1), 80-94.

Hidalgo, M., & Gómez, J. (2000). Comparación de la eficacia de regresión logística politómica y análisis discriminante logístico en la detección del DIF no uniforme. Psicothema, 12(SUPPL. 2), 298–300.

Ineval. (2017). Ficha técnica y conceptual "Ser bachiller". Quito, Pichincha, Ecuador.

Jiménez, F. (2018). *Universidad de Granada.* Obtenido de http://www.ugr.es/~jmolinos/files/elaboraciondediagramasdebode.pdf

Jodoin, M., & Gierl, M. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection.

Juana Gómez-Benito, M. D. (2010). El Sesgo de los Instrumentos de Medición. Tests Justos. *Papeles del Psicólogo*, Vol. 31(1). pp.75-84.

López, A., Sánchez, H., Espinosa, J., & Carmona, M. (2013). *Elaboración de ítems de opción múltiple.* Quito, Ecuador: Instituto Nacional de Evaluación Educativa. Obtenido de https://www.educar.ec/servicios/1-manualelaboracionitems-ineval.pdf

Matus, C., Stevenson, M., & Mirella, V. (2010). Funcionamiento Diferencial de los ítems de la prueba PISA 2009, según género para las pruebas de Matemática y Ciencias.

Mellenbergh, G. J. (1982). *Contingency Table Models for Assessing Item Bias.* Washington: Journal of Educational Statistics.

MIDE UC. (11 de Enero de 2012). *El uso de evaluaciones estandarizadas en el aula: La experiencia de las pruebas SEPA.* Obtenido de http://mideuc.cl/wp-content/uploads/2012/01/Uso-de-evaluaciones-estandarizadas-en-el-aula.pdf

Moreira, E. (2008). El Funcionamiento diferencial del ítem: un asunto de validez y equidad. *Avances en medición*, 6, 5-16.

National Center for Education Statistics. (2016). *Results of NAEP Differential Item Functioning (DIF) Analysis for Mathematics Long-Term Trend Assessment in 2012*. Obtenido de https://nces.ed.gov/nationsreportcard/tdw/analysis/2012/scaling_avoidviolat_results_mathltt2012.aspx

Pérez Gil, J. A. (2004). Desarrollos actuales de la medición: Aplicaciones en evaluación psicológica. Sevilla, España.

Ravela, P., Wolfe, R., Valverde, G., & Esquivel, J. (2000). *Lod próximos pasos: ¿Hacia dónde y cómo avanzar en la evaluación de aprendizajes en América Latina?* Lima: Programa de Promoción de la Reforma Educativa en América Latina y el Caribe.

Rojas, A. (2002). Reflexiones críticas sobre la investigación en medición mediante tests en España. *Apuntes de Psicología*, 81-95.

Swaminathan, H., & Rogers, J. (1990). *Detecting Differential Item Functioning Using Logistic Regression Procedures.* Obtenido de http://www.jstor.org/stable/1434855

Zieky , M. (2003). *A DIF primer.* Obtenido de https://www.ets.org/Media/Tests/PRAXIS/pdf/DIF_primer.pdf.

Zumbo, B. (1999). A Handbook on the Theory and Methods of Differential Item Functioning (DIF): : Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores. Ottawa: ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B., & Thomas, D. (1997). A measure of effect size for a model-based approach for studying DIF (Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science). Canada.