Revista Educación Vol. 20, Núm. 20(2022), 26-38

Evaluación psicométrica del "Ser Bachiller 2020-Régimen Costa" aplicando el Funcionamiento Diferencial de los Ítems

Psychometric evaluation of the "Ser Bachiller 2020-Régimen Costa" applying the Differential Functioning of the Items

Ajila Sanmartín, Jhon Paul
Instituto Nacional de Evaluación Educativa, Quito, Ecuador
jhon.ajila@evaluacion.gob.ec
https://orcid.org/0000-0002-2823-5480.

Núñez Wong, Juan Andrés
Instituto Nacional de Evaluación Educativa, Quito, Ecuador
juan.nunez@evaluacion.gob.ec
https://orcid.org/0000-0003-3670-2488.

Recibido 12 de julio 2021 Aprobado 03 de junio de 2022

Resumen

Este estudio muestra los resultados del análisis de detección de Funcionamiento Diferencial de Ítems (DIF, por sus siglas en inglés) de los ítems de la evaluación Ser Bachiller del ciclo Costa, que se aplicó en 2020 a bachilleres y aspirantes a ingresar a las universidades públicas de Ecuador. Se utilizó el método de Mantel-Haenszel, método muy conocido y usado por el Educational Testing Service. Se exploró la presencia de DIF a partir de cuatro ejes de análisis: sexo del sustentante, área de asentamiento, grupo de autoidentificación étnica y tipo de escolaridad. Se concluye que los ítems del Ser bachiller Costa 2020 no se encuentran afectadas por funcionamiento diferencial. Esto apoya la hipótesis de que los reactivos de las pruebas aplicadas por el Instituto Nacional de Evaluación Educativa (Ineval) no favorecen o perjudican a determinados grupos poblacionales específicos, por lo cual los diferentes grupos poblacionales evaluados no se ven afectados de ninguna manera por el diseño de la evaluación Ser Bachiller.

Palabras clave: Funcionamiento diferencial del Ítem, Pruebas estandarizadas, Mantel-Haenszel, Validez, sesgo.

Abstract

This study shows the results of the Differential Item Functioning (DIF) detection analysis of the items of the Ser Bachiller assessment of the Costa cycle, which was applied in 2020 to high school graduates and applicants to public universities in Ecuador. The Mantel-Haenszel method, a well-known method used by the Educational Testing Service, was used. The presence of DIF was explored based on four axes of analysis: sex of the subject, area of settlement, ethnic self-identification group and type of schooling. It is concluded that the items of the Ser bachiller Costa 2020 are not affected by differential functioning. This supports the hypothesis that the test items applied by Ineval do not favor or disadvantage certain specific population groups, so the different population groups evaluated are not affected in any way by the design of the Ser Bachiller assessment.

Keywords: Differential Item Functioning, Standardized Tests, Mantel-Haenszel, Validity, bias.

Introducción

En las evaluaciones implementadas por el Ineval en Ecuador como la prueba Ser Bachiller (SBAC) y Ser Estudiante (SEST), los resultados de las mismas evidencian patrones consistentes: en promedio, los alumnos de escuelas privadas logran mayores puntajes que aquellos de instituciones públicas, así también los estudiantes de escuelas de zonas urbanas obtienen mayores logros de aprendizaje que los de zonas rurales (Ineval, 2018a, 2019a, 2020). De igual manera los estudiantes indígenas presentan los resultados más bajos en relación a los estudiantes blancos o mestizos (Ineval, 2019b).

En la misma línea, en la Prueba Ser Bachiller 2020 del ciclo Costa, que se aplicó a los estudiantes de tercero de bachillerato y las personas que desean acceder a la universidades públicas de Ecuador, se observa que en la nota de examen de grado, los alumnos de instituciones educativas privadas obtuvieron un promedio de 8,01 puntos, y los de fiscal de 7,51 puntos, mientras que en los campos de matemáticas y lengua y literatura se observa que los estudiantes de instituciones privadas (8,09 y 7,91 puntos) obtienen mejores logros de aprendizaje que los chicos de instituciones de sostenimiento fiscal (7,56 y 7,31 puntos). Así también con respecto al área de asentamiento de las instituciones de zonas urbanas (7,63 puntos) obtienen mejores promedios que las zonas rurales (7,61 puntos), tendencias que se repiten en las pruebas aplicadas por el Ineval en el lapso del 2018 al 2020 (Ineval, 2018a, 2019b, 2020).

Estos comportamientos se los puede interpretar como evidencia de que los factores de contexto influyen en los logros de aprendizaje de los estudiantes, donde las condiciones de las escuelas tienden a reforzar estos patrones en lugar de revertirlos (Blanco, 2007). No obstante, también se podría considerar la hipótesis que los patrones presentados en las evaluaciones estandarizadas aplicadas por Ineval evidencian algún tipo de sesgo que afectaría la validez de la prueba.

Este tipo de sesgo que se presenta en las evaluaciones de logro se lo denomina Funcionamiento Diferencial del ítem (DIF, por sus siglas en inglés), este ocurre cuando evaluados de diferentes grupos, con igual nivel de habilidad medido por la prueba, presentan diferentes probabilidades de acertar al mismo ítem (Penfield & Camilli, 2007). Este fenómeno genera invalidez de la prueba, lo que conlleva a que se formulen conclusiones erradas de ciertos grupos según su sexo, nivel socioeconómico, área de asentamiento, etnicidad, entre otras; así también esto deriva en diagnósticos erróneos y en la eventual formulación de políticas públicas poco apropiadas al contexto de los grupos vulnerables o que históricamente han sido relegados (García-Medina et al., 2016).

En este sentido, el propósito de este análisis es comprobar la existencia de Funcionamiento Diferencial de los ítems, según sexo del sustentante, área de asentamiento, grupo de autoidentificación étnica y tipo de escolaridad, para generar evidencia sobre la validez de los ítems de la evaluación estandarizada Ser Bachiller del ciclo Costa aplicada en el 2020.

Preguntas de investigación

- ¿Existe Funcionamiento Diferencial de los ítems (DIF), según sexo del sustentante, área de asentamiento, grupo de autoidentificación étnica y tipo de escolaridad de la evaluación Ser Bachiller del ciclo Costa aplicada en el 2020?

- ¿Qué porcentaje de ítems de la evaluación Ser Bachiller del ciclo Costa aplicada en el 2020 presentan DIF según las variables sexo del sustentante, área de asentamiento, grupo de autoidentificación étnica y tipo de escolaridad?
- ¿Cuál es el efecto del DIF, si lo hubiera, en los ítems de la evaluación Ser Bachiller del ciclo Costa aplicada en el 2020 según las variables sexo del sustentante, área de asentamiento, grupo de autoidentificación étnica y tipo de escolaridad?
- La validez es uno de los criterios más relevantes para valorar la eficacia de una evaluación estandarizada y se la define como "el grado en el que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba y los usos que se pretende hacer de ellos" (AERA-APA-NCME, 2014:11).

El DIF presente en los ítems suele afectar la validez de las pruebas, por lo cual es relevante que durante el diseño y elaboración de las mismas se considere el lenguaje empleado, contexto explicado, las figuras, en otras palabras, para que una prueba sea válida no debe presentar obstáculos y todos los sustentantes deben tener la misma probabilidad de éxito (AERA-APA-NCME, 2014).

Ciertas situaciones o rasgos podrían influir en la dificultad de la prueba, evaluando rasgos ajenos o secundarios a lo que realmente se pretende medir, por ejemplo en un ítem que pretende evaluar la habilidad de multiplicar (factor principal) implícitamente también está evaluando comprensión lectora (factor secundario), en este caso el ítem ya estaría presentado funcionamiento diferencial (García-Medina et al., 2016).

La presencia de DIF en instrumentos evaluativos implica no poder conocer los verdaderos resultados del nivel de rendimiento de los sustentantes, ya que esta variable latente se encuentra medida con un instrumento con características psicométricas deficientes. Esta situación podría generar consecuencias dentro del Sistema Nacional de Educación, ya que a partir de estas medidas se clasifica a los estudiantes en determinados niveles de logro (Resino, 2018).

El funcionamiento diferencial del ítem puede afectar la equidad de género, ya que potencialmente causa sesgos en las pruebas y en la interpretación de los resultados. Wedman (2018) realizó un estudio sobre el DIF en relación al género con datos de 800 reactivos contestados por 250.000 sustentantes, las técnicas estadísticas utilizadas en el análisis fueron el método Mantel-Haenszel y el procedimiento de regresión logística; el autor encontró que los ítems de matemáticas presentaban DIF irrelevante y los ítems verbales exhibían DIF moderado. Asimismo, los reactivos de vocabulario favorecían a las mujeres si se tomaban muestras de dominios que tradicionalmente favorecían al género femenino, pero no al revés, si se tomaban muestras de dominios donde los chicos mostraban mayor éxito. Finalmente, se observó que el formato del ítem para completar oraciones en la subprueba de comprensión de lectura en inglés benefició a los hombres independientemente del contenido.

Cozby y Burcu (2020), en su estudio el cual tuvo por objetivo analizar las respuestas de los estudiantes de tres países latinoamericanos (Chile, Costa Rica y México) a 16 reactivos de ciencias de respuesta dicotómica de la prueba PISA 2015, encontraron que alrededor de un 25% de los reactivos en cada par de países presentaba ítems con funcionamiento diferencial severo.

Para Ecuador, en un estudio realizado por Ineval sobre el DIF a la evaluación Ser Bachiller 2017 mediante los métodos de Mantel-Haenszel y de regresión logística, los resultados evidenciaron

que el 2,3% de los reactivos en el régimen Costa y el 4,1% de los ítems en el régimen Sierra mostraban funcionamiento diferencial elevado o moderado (Ineval, 2018b).

En la misma línea Ineval realizó la detección del funcionamiento para la prueba Ser Bachiller del ciclo 2018 para las variables sexo, área, autoidentificación étnica y financiamiento. Considerando las características analizadas, se evidenció que el campo evaluado que exhibe mayor porcentaje de ítems con DIF es el Dominio científico; seguido por el Dominio lingüístico. Por otra parte, la variable de autoidentificación étnica indígenas vs blancos/mestizos y la variable financiamiento privado vs instituciones públicas mostraron los más altos porcentajes de reactivos con DIF tanto para Sierra como para el régimen de Costa (Ineval et al., 2021).

Método

Participantes

Para el análisis DIF se usó la información de la evaluación Ser Bachiller Costa aplicada en el 2020, a sustentantes ordinarios de la población general, que son aquellos que rindieron la prueba en las fechas originales de programación.

En la base de datos existen 254.761 sustentantes de los cuales el 52,03% son mujeres y el 47,97% son hombres. Para la variable área, el 82,47% son sustentantes de instituciones educativas del área urbana y el 17,53% del área rural. En relación a la variable etnia, el grupo con mayor representación es el de blanco/mestizo, con 80,63% de participación y el grupo más pequeño son los indígenas con un porcentaje del 2,89%. Para la variable población (escolares y no escolares), el 64,96% son escolares y el 35,04% son no escolares, es decir, escolares son aquellos que dan la prueba para graduarse del bachillerato e ingresar a las universidades y los no escolares son aquellos individuos ya graduados, que postulan para acceder a las universidades públicas de Ecuador.

Tabla 1
Estadísticos descriptivos de los sustentantes de la evaluación Ser Bachiller

Variable	Grupo	Respuestas Válidas	Porcentaje
Sexo	Mujer	132.541	52,03%
	Hombre	122.220	47,97%
Área de asentamiento	Rural	44.653	17,53%
	Urbano	210.108	82,47%
Autoidentificación étnica	Blanco/mestizo	205.418	80,63%
	Otros	49.343	19,37%
Población	Escolar	165.483	64,96%
	No Escolar	89.278	35,04%

Fuente: Ineval, 2020. Elaborado por el Instituto Nacional de Evaluación Educativa–Dirección de Análisis Psicométrico.

Instrumento

La información analizada corresponde a la evaluación Ser Bachiller del ciclo Costa aplicada en el 2020, la cual está dirigida a alumnos de tercer año de bachillerato y a ciudadanos que desean acceder

a las universidades públicas del Ecuador. En esta prueba se evaluaron los campos de biología, ciencias sociales, física, lengua y literatura, matemática y química.

Procedimiento y análisis

En este análisis se utilizó el método estadístico de Mantel-Haenszel (MH) para la detección del funcionamiento diferencial, este nos sirvió para identificar y cuantificar el tamaño del efecto DIF en los reactivos de la Prueba Ser Bachiller del ciclo Costa aplicada en el 2020. Este es un método no paramétrico muy utilizado actualmente para la detección estadística de DIF, propuesto por primera por Holland y Thayer en 1988.

La metodología de MH comprende procedimientos estadísticos muy flexibles para valorar el grado de asociación entre dos variables categóricas, ya sean estas nominales u ordinales, mientras se controlan otras variables. La versatilidad de los enfoques analíticos de Mantel-Haenszel los ha hecho muy populares en la detección del DIF de ítems dicotómicos y politómicos (Fidalgo & Madeira, 2008).

El procedimiento de MH calcula la razón de probabilidades entre el grupo focal y el grupo de referencia, estadísticamente es una proporción de las probabilidades de éxito del grupo de referencia sobre las probabilidades de éxito del grupo focal.

El tamaño de muestra necesario para ejecutar el procedimiento MH es de 200 a 500 participantes con al menos 100 participantes en el grupo focal (Brown, 1996). En este estudio se aplicó la prueba MH a los datos de la evaluación Ser Bachiller del ciclo Costa aplicada en el 2020, como se muestra en la Tabla 1, el número de sustentantes superó los requisitos para todas las variables analizadas. Para el cálculo del estadístico de MH, se utilizó el paquete "dicho Dif" del software R.

Para el análisis se consideró un grupo focal, el cual se cree que está siendo afectado por el sesgo como mujeres, indígenas, personas de zonas rurales; y otro de referencia el cual sería el beneficiado, por ejemplo, hombres, personas blancas y mestizas, individuos de zonas urbanas. Así también en el procedimiento de MH se agrupó a los sustentantes según el nivel de logro (Insuficiente, elemental, satisfactorio y excelente) y para cada uno de ellos se conforma una tabla de contingencia como se muestra a continuación:

Tabla 2

Tabla de contingencia para el intervalo i.

	Aciertos (1)	Errores (0)	Total
Grupo de referencia	a_i	b_i	$a_i + b_i$
Grupo 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 y elaboración: (Chavez & Saade, 2010)

Luego se procedió a calcular la probabilidad de acertar al ítem entre el grupo de referencia y el focal, en base a la siguiente fórmula:

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

Donde:
a_i= Número de evaluados del grupo de referencia que acertaron al reactivo
b_i= Número de evaluados s del grupo de referencia que no acertaron al reactivo
c_i= Número de evaluados del grupo focal que acertaron al reactivo
d_i= Número de evaluados del grupo focal que no acertaron al reactivo
T_i= Es el total de evaluados en el nivel i de la puntuación observada.

Posteriormente, una vez calculado el estadístico de MH, se convierte a una escala logarítmica, con el fin de representar valores de diferentes medidas a una única escala de medición (Jiménez, 2018). Esta nueva escala la nombramos como delta (δ) y se obtiene de la siguiente manera:

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

Finalmente, los valores calculados del delta lo interpretamos en base a la tabla 3 (Zwick & Ercikan, 1989, Dorans & Holland, 1992; Hidalgo & López, 2004; Zieky, 2003):

Tabla 3

Interpretación del valor delta de Mantel- Haenszel para la detección de funcionamiento diferencial.

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

Fuente y elaboración: (Zieky, 2003)

Para el análisis no se tomó en cuenta las siguientes cuatro poblaciones:

- Población general excluidas las poblaciones PPL, aulas hospitalarias, domicilio, exterior, CAI (Centros de Adolescentes Infractores), CETAD (Centros Especializados en Tratamiento a Personas con Consumo Problemático de Alcohol y otras drogas).
- Población con discapacidad auditiva.
- Población con discapacidad visual.
- Población con discapacidad intelectual.

Resultados

El análisis DIF se realizó para la Prueba Ser Bachiller 2020 del ciclo Costa. En este sentido, la presentación de resultados es realizada de manera individual para cada una de las variables que componen el análisis DIF. El orden de presentación es área de asentamiento de la institución educativa, sexo del sustentante, escolaridad y finalmente, autoidentificación étnica.

A continuación, se presentan consideraciones generales para las variables analizadas; en el caso de la variable escolaridad, se toma como grupo referencial a la población de escolares y como grupo focal a la población de no escolares. Para la variable autoidentificación étnica se considera como grupo referencial al segmento poblacional autoidentificado como mestizos y blancos; y como grupo focal a individuos autoidentificados como afroecuatorianos, montubios, indígenas y otros. Para la variable sexo, se consideró al grupo referencial como hombres y el grupo focal como mujeres. Para la variable área se consideró como grupo focal al grupo urbano y como grupo referencial al grupo rural.

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

- Que exista significancia del estadístico ji-cuadrado de Mantel-Haenszel.
- En el análisis del DIF, el nivel de confianza se estableció al 95%.

DIF por área de asentamiento de la institución educativa

El análisis por área de asentamiento de la institución educativa, diferencia las instituciones asentadas en el área urbana (grupo de referencia) y en el área rural (grupo focal).

En cuanto al análisis del funcionamiento diferencial del ítem por la variable área solo se encontró ítems con DIF de categoría irrelevante. A partir de este análisis 634 ítems presentaron funcionamiento diferencial, donde, en los dominios de lengua y literatura (53,0%), y biología (51,7%) se evidenciaron los mayores porcentajes de reactivos con este fenómeno. En los dominios de ciencias sociales (50,5%), física (50,0%) y química (50,0%) se registra los mayores porcentajes de ítems sin DIF.

Tabla 4

DIF en SBAC Costa aplicada en el 2020 por área de asentamiento de la institución educativa

Dominio		Ítems con DIF irrelevante	Ítems sin DIF	Total ítems por dominio
Biología	Absolutos	46	43	89
	Porcentaje	51,7%	48,3%	100,0%
Ciencias Sociales	Absolutos	150	153	303
	Porcentaje	49,5%	50,5%	100,0%
Física	Absolutos	74	74	148
	Porcentaje	50,0%	50,0%	100,0%
Lengua y Literatura	Absolutos	152	135	287
	Porcentaje	53,0%	47,0%	100,0%
Matemática	Absolutos	131	127	258
	Porcentaje	50,8%	49,2%	100,0%
Química	Absolutos	81	81	162
	Porcentaje	50,0%	50,0%	100,0%
TOTAL	Absolutos	634	613	1247
	Porcentaje	50,8%	49,2%	100,0%

Fuente: Ineval, 2020 – Dirección de Análisis Psicométrico. Elaborado por el Instituto Nacional de Evaluación Educativa-Dirección de Análisis de la Evaluación Educativa

DIF por sexo

La siguiente variable para la cual se realiza el análisis es para sexo del sustentante. En este caso, el total general de ítems con DIF irrelevante no supera el 5%; específicamente en los campos de física (7,4%) y ciencias sociales (6,3%) se muestran los mayores porcentajes de DIF irrelevante para la variable analizada.

Si consideramos aquellos reactivos donde no se detecta DIF, los mismos representan el 95,6% del total general. A nivel de dominio se evidencia que Lengua y Literatura con el 97,6% es donde se registra el mayor porcentaje de reactivos sin DIF.

Tabla 5
DIF en SBAC Costa aplicada en el 2020 por sexo del sustentante

Dominio		Ítems con DIF irrelevante	Ítems sin DIF	Total ítems por dominio
Biología	Absolutos	5	84	89
	Porcentaje	5,6%	94,4%	100,0%
Ciencias Sociales	Absolutos	19	284	303
	Porcentaje	6,3%	93,7%	100,0%
Física	Absolutos	11	137	148
	Porcentaje	7,4%	92,6%	100,0%
Lengua y Literatura	Absolutos	7	280	287
	Porcentaje	2,4%	97,6%	100,0%
Matemática	Absolutos	8	250	258
	Porcentaje	3,1%	96,9%	100,0%
Química	Absolutos	5	157	162
	Porcentaje	3,1%	96,9%	100,0%
TOTAL	Absolutos	55	1192	1247
	Porcentaje	4,4%	95,6%	100,0%

Fuente: Ineval, 2020 – Dirección de Análisis Psicométrico. Elaborado por el Instituto Nacional de Evaluación Educativa–Dirección de Análisis de la Evaluación Educativa

DIF por autoidentificación étnica

Para la variable de autoidentificación étnica se observa la presencia de DIF irrelevante en 277 de los 1247 ítems de la evaluación Ser Bachiller del ciclo Costa aplicada en el 2020, de los cuales en química representan el 26,5% y en física el 26,4%.

En relación, a los ítems sin DIF, 970 reactivos de 1247 no presentan funcionamiento diferencial; en los dominios de lengua y literatura (84,3%), y matemáticas (80,2) se evidencia el mayor porcentaje de ítems sin DIF.

Tabla 6
DIF en SBAC Costa aplicada en el 2020 por autoidentificación étnica

Dominio		**Ítems con DIF irrelevante**	**Ítems sin DIF**	**Total ítems por dominio**
Biología	Absolutos	20	69	89
	Porcentaje	22,5%	77,5%	100,0%
Ciencias Sociales	Absolutos	79	224	303
	Porcentaje	26,1%	73,9%	100,0%

Física	Absolutos	39	109	148
	Porcentaje	26,4%	73,6%	100,0%
Lengua y Literatura	Absolutos	45	242	287
	Porcentaje	15,7%	84,3%	100,0%
Matemática	Absolutos	51	207	258
	Porcentaje	19,8%	80,2%	100,0%
Química	Absolutos	43	119	162
	Porcentaje	26,5%	73,5%	100,0%
TOTAL	**Absolutos**	**277**	**970**	1247
	Porcentaje	**22,2%**	**77,8%**	**100,0%**

Fuente: Ineval, 2020 – Dirección de Análisis Psicométrico. Elaborado por el Instituto Nacional de Evaluación Educativa – Dirección de Análisis de la Evaluación Educativa

DIF por escolaridad

En la variable escolaridad, 1138 ítems fueron detectados con funcionamiento diferencial en la categoría irrelevante. De estos ítems el mayor porcentaje se concentra en los dominios de física y biología, con 96,6% y 93,3%, respectivamente. Así también se detectaron ítems con DIF moderado en los campos de física y lengua y literatura, entre ambos suman un total de 0,2% en relación al total general. Reactivos de ciencias sociales, física y matemática mostraron DIF severo, estos representan el 0,2% del total de ítems evaluados en la prueba Ser Bachiller del ciclo costa 2020.

Tabla 7

DIF en SBAC Costa aplicada en el 2020 por escolaridad

Dominio		Ítems con DIF irrelevante	Ítems con DIF moderado	Ítems con DIF severo	Ítems sin DIF	Total ítems por dominio
Biología	Absolutos	112	0	0	8	120
	Porcentaje	93,3%	0,0%	0,0%	6,7%	100,0%
Ciencias Sociales	Absolutos	268	0	1	34	303
	Porcentaje	88,4%	0,0%	0,3%	11,2%	100,0%
Física	Absolutos	143	1	1	3	148
	Porcentaje	96,6%	0,0%	0,7%	2,0%	99,3%
Lengua y Literatura	Absolutos	258	1	0	28	287
	Porcentaje	89,9%	0,0%	0,0%	9,8%	99,7%
Matemática	Absolutos	233	0	1	24	258
	Porcentaje	90,3%	0,0%	0,4%	9,3%	100,0%
Química	Absolutos	119	0	0	12	131
	Porcentaje	90,8%	0,0%	0,0%	9,2%	100,0%
TOTAL	**Absolutos**	**1133**	**2**	**3**	**109**	**1247**
	Porcentaje	**90,9%**	**0,2%**	**0,2%**	**8,7%**	**100,0%**

Fuente: Ineval, 2020 – Dirección de Análisis Psicométrico. Elaborado por Instituto Nacional de Evaluación Educativa – Dirección de Análisis de la Evaluación Educativa

Discusión

El impacto de las variables de contexto incide no solo en los niveles de logro de los estudiantes sino también en la forma en que el estudiantado contesta las pruebas (Woitschach & Ortiz, 2019). América Latina es de las regiones con mayor inequidad social y educativa (UNESCO-OREALC, 2013), en este sentido ignorar el DIF puede resultar en comparaciones de puntajes no válidos para los sustentantes que rindieron la prueba (Cho et al., 2016), es así que con el fin de garantizar la validez de las pruebas y de que las mismas no presentan sesgos se realiza el análisis del Funcionamiento Diferencial del Ítem.

Aplicando la metodología de Mantel Haenszel, el análisis estadístico dio como resultado la identificación de ítems con DIF irrelevante en las variables área de asentamiento de la institución educativa, sexo del sustentante, autoidentificación étnica. Cabe mencionar que para la variable escolaridad 2 ítems presentaron DIF moderado y 3 reactivos mostraron DIF severo.

En relación a los resultados de este estudio se observa que, aunque algunos reactivos presentan cierta medida de sesgo irrelevante en las variables analizadas, el funcionamiento diferencial es tan imperceptible que no incide a nivel de la prueba; por consiguiente, las medidas de rendimiento del Ser bachiller Costa 2020 no se encuentran afectadas por DIF para las variables área de asentamiento de la institución educativa, sexo del sustentante, autoidentificación étnica del alumnado.

Para los ítems que presentaron DIF moderado y DIF severo en la variable escolaridad debería corroborarse mediante el análisis crítico de expertos las posibles razones por las que los estudiantes contestaron de diferente forma, ya que identificar las causas del funcionamiento diferencial en esos reactivos ayudaría a comprender que constructos realmente se están midiendo. En consecuencia, la validez de las conclusiones extraídas de las notas de las pruebas podría aumentar.

Para la variable sexo del sustentante se presenta semejantes resultados con el estudio de Resino (2018) y al de García-Medina et al., (2016), donde se detectaron ítems con funcionamiento diferencial de categoría irrelevante de acuerdo a los criterios de Zumbo y Thomas (1997), y de Jodoign & Gierl (2001), en resumen, este tipo de DIF no afecto la validez de las pruebas analizadas. Los resultados de esta investigación para la variable autoidentificación étnica es consistente con el estudio de Maddox et al., 2015, donde no se encontró DIF significativo por la variable analizada. Sin embargo, estos no concuerdan con los del estudio realizado por Taylor y Lee (2011), donde se analizó la prueba de lectura de cuarto, séptimo y decimo, encontrándose que los ítems de respuesta construida beneficiaban al estudiantado de grupos minoritarios y los reactivos de opción múltiple a los alumnos blancos, especialmente en lectura.

En relación a la variable de asentamiento de la institución educativa los resultados concuerdan con los análisis realizados por la Agencia de Calidad de Chile en el informe técnico Simce (Sistema de Medición de Logros de Aprendizaje) de la prueba 2013, donde se descartó la presencia de DIF en base a las variables género y ruralidad (Agencia de Calidad de la Educación, 2015).

Respecto a la variable escolaridad en la literatura no se encontraron estudios realizados en la región con los cuales contrastar los resultados realizados evidenciados en este estudio.

En resumen, para garantizar la equidad de las pruebas estandarizadas a gran escala es necesario la atención a la diversidad desde el diseño de las mismas (Solano-Flores, 2011) tomando en cuenta el género, ruralidad, autoidentificación étnica y nivel socioeconómico de la población evaluada. En este

sentido, al momento de elaborar los ítems es necesario contar con un equipo multidisciplinario como sociólogos, lingüistas, antropólogos, y especialistas en determinados contextos sociales, culturales y económicos, esto con el fin de que estas pruebas favorezcan la mayor participación de alumnos y que las inferencias extraídas de los resultados no presentes sesgos.

Conclusiones

La validez en evaluaciones a gran escala es uno de los criterios más relevantes para valorar la eficacia de las mismas, que exista invalidez conlleva a que se formulen conclusiones erradas de ciertos grupos según su sexo, nivel socioeconómico, área de asentamiento, etnicidad, entre otras; así también esto deriva en diagnósticos erróneos y en la eventual formulación de políticas públicas poco apropiadas al contexto de los grupos vulnerables o que históricamente han sido relegados.

El Funcionamiento Diferencial del ítem, ocurre cuando evaluados de diferentes grupos, pero con igual nivel de habilidad medido por la prueba, presentan diferentes probabilidades de acertar al mismo ítem.

El DIF presente en los ítems suele afectar la validez de las pruebas, por lo cual es relevante que, durante el diseño y elaboración de los ítems de la evaluación, los técnicos a cargo de esta tarea consideren el lenguaje empleado, contexto explicado, las figuras, redacción, colocación espacial, entre otros, esto con el fin de no incurrir en funcionamiento diferencial.

Dado que mayormente se ha detectado la presencia de DIF irrelevante en las variables analizadas, se puede concluir que no existe evidencia de que el instrumento Ser Bachiller del ciclo Costa 2020, afecte la probabilidad de acierto de los ítems por área de asentamiento, sexo del sustentante, autoidentificación étnica y por escolaridad. Esto genera evidencia que sirve para contrastar la hipótesis de que los reactivos de las pruebas de Ineval no favorecen o perjudican a determinados grupos poblacionales específicos, por lo cual los diferentes grupos poblacionales evaluados no se ven afectados de ninguna manera por el diseño de la evaluación Ser Bachiller.

En cuanto al posible sesgo por escolaridad, para confirmar que los cinco elementos de la prueba señalados en la tabla 7 tienen realmente sesgo se deberá indagar con jueces expertos, para que validen la presencia de sesgo que dificulte la probabilidad de responder correctamente a estos ítems.

Referencias

AERA-APA-NCME. (2014). Standards for Educational and Psychological Testing, Washington, dc: American Educational Research Association- American Psychological Association-National Council on Measurement in Education.

Agencia de Calidad de la Educación (2015). Informe Técnico Simce 2013. http://archivos.agenciaeducacion.cl/documentos-web/InformeTecnicoSimce_2013.pdf

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

Blanco, E. (2007). Eficacia escolar en México: factores escolares asociados a los aprendizajes en la educación primaria, Ciudad de México: flacso-Sede México. http://flacsoandes.edu.ec/dspace/handle/10469/1247.

Brown, P. J. (1996). Using differential analysis to determine differential item functioning of survey questions (Unpublished doctoral dissertation). University of Illinois, Urbana, Champaign.

Cozby Dzul-Garcia & Burcu Atar (2020) Investigation of possible item bias on PISA 2015 science items across Chile, Costa Rica and Mexico (*Estudio de los posibles sesgos entre los ítems de ciencias de la prueba PISA 2015 de Chile, Costa Rica y México*), Culture and Education, 32:3, 470-505, DOI: 10.1080/11356405.2020.1785158

Chávez, C., & Saade, A. (2010). Procedimientos básicos para el análisis de reactivos. *México: Centro Nacional de Evaluación para la Educación Superior*.

Cho, S.-J., Suh, Y., & Lee, W. (2016). *After Differential Item Functioning Is Detected. Applied Psychological Measurement, 40(8), 573–591.* doi:10.1177/0146621616664304

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization 1, 2. *ETS Research Report Series*, *1992*(1), i-40.

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement*, *68*(6), 940-958.

García-Medina, A. M., Martínez Rizo, F., & Cordero Arroyo, G. (2016). Análisis del funcionamiento diferencial de los ítems del Excale de Matemáticas para tercero de secundaria. *Revista mexicana de investigación educativa*, 21(71), 1191-1220.

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *In H. Wainer & H. I. Braun* (Eds.), Test validity (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates

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

Jodoin, M. G., y Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329–349.

Maddox, B., Zumbo, B. D., Tay-Lim, B., & Qu, D. (2015). An Anthropologist Among the Psychometricians: Assessment Events, Ethnography, and Differential Item Functioning in the Mongolian Gobi. *International Journal of Testing*, 15(4), 291–309.doi:10.1080/15305058.2015.1017103

Mora, T. E. M. (2008). Funcionamiento diferencial del ítem en pruebas de matemática para educación media. *Actualidades en psicología*, *22*(109), 91-113.

Ineval. (2018a). Informe de resultados nacional - Ser bachiller Año lectivo 2017-2018.

Ineval. (2018b). Funcionamiento Diferencial de los Ítems de la prueba Ser Bachiller 2017, según sexo

Ineval. (2019a). Informe de resultados nacional - Ser bachiller Año lectivo 2018-2019.

Ineval. (2019b). La educación en Ecuador: logros alcanzados y nuevos desafíos.

Ineval. (2020). Informe de resultados: Evaluación Costa 2019-2020.

Ineval., Ajila, J. & Levy, E. (2021). Estudio del funcionamiento diferencial de los ítems de la evaluación Ser Bachiller 2018, según las variables sexo, área, autoidentificación étnica y financiamiento. http://evaluaciones.evaluacion.gob.ec

Penfield, R., & Camilli, G. (2007). Test fairness and differential item functioning. *Handbook of statistics*, 26, 125-167.

Resino, D. A. (2018). Funcionamiento diferencial del ítem por sexo en alumnos de Educación Secundaria. *Experiencias educativas en el aula de infantil, primaria y secundaria*, 66.

Solano-Flores, G. (2011). "Assessing the cultural validity of assessment practices", en Cultural Validity in Assessment Nueva York: Routledge, pp. 3-21.

Taylor, C., & Lee, Y. (2011). Ethnic DIF in reading tests with mixed item formats. *Educational Assessment*, 16(1), 35-68.

UNESCO-OREALC. (2013). Situación educativa de América Latina y el Caribe: Hacia la educación de calidad para todos al 2015. Santiago de Chile: UNESCO.

Wedman, J. (2018). Reasons for Gender-related Item Functioning in a College Admissions Test, *Scandinavian Journal of Educational Research*, 62: 6, 959-970, DOI: 10.1080 / 00313831.2017.1402365

Woitschach, P., & Ortiz, L (2019). Funcionamiento diferencial del ítem en la evaluación educativa a nivel América Latina y el Caribe.

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

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, 26, 55-66.

Zumbo, B. D., y Thomas, D. R. (1997) A measure of effect size for a model-basedapproach for studying DIF. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science. University of Northern British Columbia: PrinceGeorge, B.C