



DOI: <https://doi.org/10.70208/3007.8245.v6.n1.335>

Análisis del Funcionamiento Diferencial de los ítems en una prueba estandarizada de ingreso a carrera

Kevin Gómez Guillen¹

kevin.gomezguillen@ucr.ac.cr

<https://orcid.org/0009-0002-5415-4729>

Universidad de Costa Rica
Costa Rica

Graciela Ordóñez Gutiérrez

graciela.ordonez@ucr.ac.cr

<https://orcid.org/0000-0002-0762-3759>

Universidad de Costa Rica
Costa Rica

RESUMEN

El objetivo de este estudio es evaluar el funcionamiento diferencial de los ítems de una prueba estandarizada que mide las habilidades de razonamiento cuantitativo en un grupo de examinados de los años 2021, 2022 y 2023. El diseño fue no experimental *ex post facto* y se empleó el método estadístico Mantel-Haenszel. Dentro de los principales resultados se obtuvo que para la prueba del 2021 el IT27, el IT5 y IT39 de la prueba del 2022 y el IT38 de la prueba del 2023 favorecen a los hombres. En cuanto al tipo de colegio, solo para la prueba del año 2023 el IT10 favoreció a las personas provenientes de colegios privados. Finalmente, es indispensable realizar estos estudios y determinar la causa de sesgo, de esta manera mejorar el instrumento de medida y proporcionar mayor equidad y justicia.

Palabras clave: Funcionamiento diferencial de los ítems, prueba estandarizada, imparcialidad, sesgo, validez

¹ Autor principal.

Correspondencia: kevin.gomezguillen@ucr.ac.cr



Analysis of the Differential Functioning of Items in a Standardized Entrance Test for a Degree Program

ABSTRACT

The objective of this study is to evaluate the differential functioning of the items in a standardized test that measures quantitative reasoning skills for a group of examinees from 2021, 2022, and 2023. The design was non-experimental ex post facto, and the Mantel-Haenszel statistical method was employed. Among the main results, it was found that for the 2021 test, items IT 27, IT 5, and IT 39 from the 2022 test, as well as item IT 38 from the 2023 test, favored males. Regarding the type of school, only for the 2023 test did IT10 favor individuals coming from private schools. Ultimately, it is crucial to conduct these studies and identify the source of bias, thereby enhancing the measurement instrument and promoting greater equity and justice

Keywords: Differential item functioning, Standardized test, Fairness, Bias, Validity



INTRODUCCIÓN

La AERA, APA & NCME (2014) explican que una de las evidencias de validez fundamentales en el campo de la medición es aquella relativa a la estructura interna de un instrumento que mide un constructo determinado, en la cual la validez corresponde al grado en que la evidencia y la teoría permiten realizar interpretaciones de las puntuaciones relativas a una prueba.

En algunos estudios (Ordóñez Gutiérrez, 2025; Pérez-Gil et al., 2000; Gómez Guillén, 2025) sobre la estructura interna de las pruebas se analizan relaciones estadísticas sobre los comportamientos, actitudes y disposiciones clave del constructo de interés, ofreciendo evidencias para la validez de la medida según la adecuación del contenido de la prueba con el marco teórico subyacente al constructo. Otros estudios (Moreira Mora, 2008; Montero Rojas & Moreira-Mora, 2017; Martinková et al., 2017) respecto a la estructura interna de las pruebas se enfocan en demostrar si ítems en particular pueden funcionar de manera diferente para grupos de examinados identificables, por ejemplo, grupos raciales/étnicos o de género. Cuando diferentes grupos de examinados con capacidad general similar, o nivel similar en un constructo adecuado, tienen, en promedio, respuestas sistemáticamente diferentes en un ítem en particular, o bien tras considerar diferencias en el constructo las variaciones en las respuestas se asocian con la pertenencia al grupo; esto es lo que se llama *Funcionamiento Diferencial de los ítems* (DIF).

Ahora bien, el DIF no siempre es una falla o debilidad según la AERA, APA & NCME (2014), esto porque subconjuntos de ítems que tienen una característica específica en común, por ejemplo, contenidos o tareas, pueden funcionar de manera diferente para distintos grupos de examinados con puntuaciones semejantes, lo cual indica una clase de multidimensionalidad que puede esperarse o ajustarse al marco teórico de la prueba; por lo que la presencia de DIF en un ítem no garantiza que sea injusto (Martinková et al., 2017). Es por este motivo que el DIF del ítem se ha confundido con sesgo, especialmente por el doble sentido de la palabra sesgo: uno social y otro estadístico (Martínez Arias et al. 2014).

Por otra parte, el análisis del DIF es esencial para establecer la imparcialidad en las pruebas educativas y psicológicas (Albano et al., 2024), ya que al identificar DIF estadísticamente, la revisión del contenido



de los ítems puede evidenciar términos o contextos que no son esenciales para medir el constructo y que son menos familiares o relevantes para un grupo en comparación con otro. Además, el análisis de DIF en los ítems de pruebas de alto impacto permite garantizar que la medición realizada sea justa, equitativa y válida para todos los grupos de examinados, de esta manera prevenir conclusiones incorrectas sobre el rendimiento de las diferentes poblaciones y asegurar que las evaluaciones no discriminen por factores irrelevantes como el género, idioma o contexto cultural (Martinková et al., 2017).

Análisis de DIF en pruebas educativas

Dentro de los estudios encontrados en los que se aplica el análisis de DIF en pruebas Educativas se puede destacar el realizado por Rodríguez Jiménez et al., (2021) aplicado a los ítems de la prueba EXIM, en Colombia, esto mediante el método de regresión logística. Este grupo de investigación encontró que 16 ítems presentaban DIF, tanto uniforme como no uniforme, pero con tamaño del efecto moderado. En el estudio se menciona que aquellos ítems que estaban relacionados con entender algún concepto presentaron DIF no uniforme y favoreció a las mujeres; mientras que aquellas preguntas que favorecen a los hombres y que consideraban posibles factores socioculturales, presentaban DIF uniforme. En el estudio se concluye que el manejo de conceptos teóricos y la influencia de los roles de género en la capacidad cognitiva de transferir conocimiento en diferentes contextos, así como las habilidades cognitivas en la lectura de gráficas, fue un aspecto clave en el comportamiento diferencial de los ítems, por lo que es necesario tener claridad de conceptos básicos para poder aplicarlos eficazmente en la resolución de los ítems presentados.

Por su parte, García-Medina et al., (2016) realizaron un estudio de DIF a los ítems de la prueba EXCALE, la cual es un examen de la calidad y el logro educativo, que se aplica a estudiantes de tercero de secundaria en México. El procedimiento empleado para el análisis fue el de Mantel-Haenszel, tomando como salida alternativa el modelo de Rasch. El análisis de DIF se empleó para las variables sexo y nivel socioeconómico. Dentro de los hallazgos presentados se indica que ninguno de los ítems de la prueba EXCALE de matemática presentó DIF severo entre hombres y mujeres. Además, en cuanto al



posible sesgo por nivel socioeconómico 18 ítems sí resultaron con sesgo y sugieren analizar cada uno de estos mediante otros estudios para determinar el porqué de la ocurrencia.

Al igual que en el estudio de Rodríguez Jiménez et al., (2021), García-Medina et al., (2016) encontraron que la presencia de DIF en los reactivos de matemática se debe a la inclusión de términos que no están relacionados con el desempeño en matemática, pero, además, son desconocidos para los subgrupos de estudiantes poblacionales donde se presentó el DIF o bien en el contexto donde se sitúa el problema, el cual puede ser ajeno a las personas examinadas. Para determinar la raíz del problema, las personas investigadoras recomiendan realizar protocolos verbales o entrevistas cognitivas a las personas examinadas.

Además, se destaca el estudio de Sánchez Mendiola et al., (2020) la cual analizaron el funcionamiento diferencial de los ítems del examen de ingreso a la Universidad Nacional Autónoma de México, esto según sexo, con la finalidad de aportar evidencias de validez en la prueba. Para el análisis se empleó el modelo de Rasch, encontrando muy pocos reactivos con DIF leve-moderado (lógitos entre .43 y .64). Los investigadores indican que no identificaron ítems con niveles de DIF importante y que pudieran comprometer los resultados y las inferencias sobre la población. Sin embargo, no se establece cuál sería el motivo de que ciertos ítems salieran con DIF leve-moderado para su respectiva interpretación. Por otro lado, Arroyo Resino (2018) evaluó el DIF en ítems de una prueba de razonamiento matemático de la educación secundaria, esto según el sexo. El análisis lo llevó a cabo mediante una regresión logística y una prueba de verosimilitud, en la que detectó DIF en algunos de los ítems. Este autor concluye que, aunque algunos ítems se encuentran sesgados en cierta medida a favor de uno de los dos sexos, el DIF es imperceptible y, por lo tanto, no afecta el nivel del test y las medidas de rendimiento obtenidas no se encuentran afectadas por el DIF.

Desde otra perspectiva, Finch & Hernández Finch (2013) analizaron DIF en una prueba de lenguaje y matemática aplicadas a estudiantes de tercero de primaria en USA, considerando variables como adecuación, tipo de escuela, género, etnia, raza, entre otras, en un contexto de varios niveles. Con el estudio se detectaron ítems que mostraron DIF uniforme entre las diferentes clases latentes, esto a nivel de estudiantes como a nivel de escuela, especialmente en reactivos relacionados con análisis de



datos, estadística y relaciones numéricas, lo que indica que estos ítems afectan de manera diferente a los distintos grupos. Además, los hallazgos reportados muestran que la presencia de DIF puede ser multicausal. Cabe mencionar que la población considerada en el estudio fueron estudiantes con discapacidades de aprendizaje específico y con adaptaciones en la evaluación. Igualmente, Silva Morales & Santelices Etchegaray (2016) estudiaron el DIF de los ítems de una evaluación estandarizada según las necesidades Educativas Especiales Transitorias, y en las áreas de lenguaje y matemática.

Silva Morales & Santelices Etchegaray (2016) encontraron que en las pruebas de lenguaje y matemática aplicada en los niveles desde cuarto hasta octavo muestran que los ítems con DIF encontrados en quinto, sexto y octavo (en lenguaje) y cuarto y octavo (en matemática) no superan el 10%, y los ítems encontrados con DIF, tanto en la prueba de lenguaje como de matemática tuvieron un tamaño de efecto moderado que, en algunos casos, favoreció a los niños con necesidades Educativas Especiales Transitorias y en otros los perjudicaban; siendo insuficiente para probar la existencia de sesgo en los resultados de las pruebas.

Por otra parte, Moreira Mora (2008) y Montero-Rojas & Moreira-Mora (2017) realizaron estudio de DIF con datos de pruebas costarricenses. En el primer caso, Moreira (2008) comprueba la presencia de DIF en una prueba de matemática aplicada al estudiantado de enseñanza media según trastorno de déficit atencional con hiperactividad (TDAH), encontrando que la prueba presenta un 18% de ítems con DIF de tamaño irrelevante pero estadísticamente significativo, explicando que los ítems con DIF presentan vocabulario impreciso y redacción poco clara, entre otros, lo que perjudicaría a las personas con TDAH.

Al igual que Moreira Mora (2008), Montero-Rojas & Moreira-Mora (2017) establecieron hipótesis sobre posibles fuentes de DIF, en las que compararon estudiantes reportados con el trastorno de déficit atencional con hiperactividad (TDAH, grupo focal) y estudiantes sin esas características (grupo de referencia). En ambos casos se usaron los métodos de la diferencia p estandarizada y el método Mantel Haenszel para identificar los ítems con DIF en las pruebas de bachillerato de español y matemática aplicadas en el año 2004 en colegios públicos académicos de Costa Rica. Los hallazgos indican que la



extensión y la redacción de las opciones y los textos literarios y no literarios incluidos en el encabezado pueden ser causa de DIF, y en la prueba de matemática las fuentes son el vocabulario, la redacción, la transición del lenguaje verbal al matemático, la organización visoespacial, el dibujo de gráficas o figuras geométricas. Sin embargo, se indica que, del total de ítems, se detectó un 12.5% con DIF en la prueba de español, mientras que un 17.9% en la de matemática, pero el nivel fue insignificante.

A pesar de la relevancia del tema del *Funcionamiento Diferencial de los ítems* en pruebas educativas, son escasos, si no es que nulos, los estudios que se pueden encontrar en las distintas bases de datos indagadas, sobre todo en aquellas pruebas que miden habilidades de razonamiento cuantitativo, por lo que el **objetivo de este estudio** consiste en evaluar el DIF de una prueba estandarizada que mide las habilidades de razonamiento cuantitativo, para un grupo de examinados en los años 2021, 2022 y 2023.

METODOLOGÍA

Este estudio es cuantitativo de alcance no experimental con diseño *ex post facto*, esto porque el estudio se efectuó luego de aplicar la prueba; además, las personas investigadoras no pueden cambiar las variables y los sujetos seleccionados poseen las características de interés (Portell & Vives, 2024). También, es transversal debido a que los datos fueron obtenidos una única vez con la aplicación de la prueba.

Participantes

Las personas participantes son aquellas que realizaron la prueba de habilidades cuantitativas en los años 2021, 2022 y 2023. Por lo que el tamaño de la muestra corresponde a la totalidad de la población que efectuaron la PHC. La distribución de la cantidad de personas se presenta en la tabla 1 según sexo.

Tabla 1. Cantidad de personas que realizaron la prueba estandarizada en los años 2021 al 2023, según sexo

Año	Sexo		Total
	Femenino	Masculino	
2021	1193	1484	2677
2022	997	1063	2060
2023	1222	1005	2227



Total	3412	3552	6964
--------------	-------------	-------------	-------------

Es importante indicar que las personas examinadas provienen de diferentes tipos de colegio que pueden ser públicos, privados, subvencionados, educación abierta (ED. Abierta-CR), y del extranjero. La distribución de las personas por año según tipo de colegio se proporciona en la tabla 2.

Tabla 2. Distribución de personas que realizaron la PHC en los años 2021 al 2023, según tipo de colegio

Año	Tipo de colegio					Total
	Público	Privado	Subvencionado	ED.Abierta-CR	Extranjero	
2021	1909	475	227	47	19	2677
2022	1415	394	190	43	18	2060
2023	1556	439	195	25	12	2227
Total	4880	1308	612	115	49	6964

Técnica e instrumento de recolección de datos

El instrumento empleado fue la Prueba de Habilidades Cuantitativas, la cual es una prueba estandarizada que mide el constructo *razonamiento cuantitativo* y se aplica para efectos de admisión en la Universidad de Costa Rica, en aquellas carreras que la tienen como requisito de ingreso. La prueba consta de 40 ítems distribuidos de acuerdo con las dimensiones del constructo (relacionar, clasificar, ejemplificar, validar y generalizar) y según área de contenido (álgebra, análisis de datos, geometría y aritmética). Cabe mencionar que, de los 40 ítems, 36 son de banco y 4 son experimentales. Esto significa que 36 ítems cumplen con las medidas psicométricas idóneas para la evaluación, mientras que los otros 4 son nuevos y no se conocen las medidas correspondientes. Además, la prueba se construye aplicando el modelo de medición Teoría de Respuesta al ítem (TRI), y se toma en cuenta el nivel de dificultad alrededor de cero (0) y el nivel de discriminación mayor que 0.5.

Proceso de análisis de datos

La información se obtuvo de las bases de datos de la aplicación de la Prueba de Habilidades Cuantitativas para los años 2021, 2022 y 2023. El análisis de los datos se realizó empleando el método Mantel-Haenszel, esto porque es aplicable tanto a la Teoría Clásica de los Tests (TCT) como a la TRI,



y utilizarse para detectar DIF en ítems politómicos y dicotómicos, permitiendo la comparación simultánea de dos o más grupos (Fidalgo y Scalón, 2012). Además, la prueba de *Mantel-Haenszel* es una extensión de la prueba *chi-cuadrada* para tablas de contingencias que clasifica a las personas examinadas en grupos según sus puntuaciones totales (Martinková et al., 2017). Es importante señalar que las aplicaciones correspondientes a los años 2021, 2022 y 2023 no utilizaron una forma idéntica de la prueba, sino que cada una fue construida a partir de un banco amplio y robusto de ítems, manteniendo el mismo constructo subyacente. En consecuencia, los análisis de DIF se realizaron de manera independiente para cada año, y los resultados deben interpretarse como evidencia de equidad a nivel del banco de ítems y no como un seguimiento longitudinal de ítems específicos.

Para el análisis se consideraron las variables sexo (mujer, hombre) y tipo de colegio (público, privado). Para cada una de estas se tomaron dos grupos denominados como *grupo focal* y *grupo de referencia*. El grupo focal es aquel que se considera perjudicado por la prueba, normalmente el grupo minoritario, y el de referencia es el grupo de comparación, normalmente mayoritario (Martínez Arias et al., 2014). En el caso de la *variable sexo*, el grupo focal es hombre y el grupo de referencia es mujer, mientras que para tipo de colegio el grupo focal es privado. Cabe indicar que para el caso de la variable tipo de colegio la agrupación se realizó de la siguiente manera: Privado = (Privado + Subvencionado + Extranjero), Público = (Público + ED. Abierta-CR). Además, aquellas personas que no reportaron el tipo de colegio o bien cuyo colegio era desconocido, fueron eliminadas de la base de datos para su respectivo análisis. Por otra parte, la detección e interpretación del DIF se realizó a partir de tres indicadores complementarios siguiendo las recomendaciones del *Educational Testing Service* (ETS), estos son: el estadístico MH, el valor de significancia estadística asociado (*p-value*) y el índice de magnitud ΔMH . El estadístico MH evalúa si existen diferencias estadísticamente significativas en la probabilidad de responder correctamente a un ítem entre el grupo focal y el grupo de referencia, una vez controlado el nivel de habilidad de las personas examinadas. Este control se logra mediante la estratificación de la muestra en función de la puntuación total en la prueba, lo que permite comparar a individuos de ambos grupos con niveles de habilidad equivalentes. Un valor de *p* inferior a .05 se interpreta como evidencia estadística de la presencia de DIF. No obstante, la significancia estadística por sí sola no se considera



suficiente para tomar decisiones sobre la calidad psicométrica de un ítem, especialmente en muestras grandes, ya que pequeñas diferencias pueden resultar estadísticamente significativas sin tener relevancia práctica. Por esta razón, el análisis se complementa con un índice de tamaño del efecto.

El índice ΔMH corresponde a una transformación logarítmica del *odds ratio* estimado por el método MH y permite interpretar tanto la magnitud como la dirección del DIF. Este índice indica en qué medida la probabilidad de acierto en un ítem difiere entre el grupo focal y el grupo de referencia, una vez controlada la habilidad. En este estudio, valores positivos de ΔMH indican DIF a favor del grupo focal, es decir, que el grupo focal presenta una mayor probabilidad de responder correctamente al ítem que el grupo de referencia a igual nivel de habilidad. Por el contrario, valores negativos de ΔMH indican DIF a favor del grupo de referencia, lo que implica una desventaja relativa para el grupo focal, y valores cercanos a cero sugieren ausencia de funcionamiento diferencial relevante.

Para la interpretación de la magnitud (ΔMH) el DIF se adoptaron los criterios desarrollados por el ETS, basados en Holland & Wainer (1993) que consideran tres categorías, DIF despreciable (A), DIF intermedio (B) y DIF grande (C). Esta categorización (A, B y C) se establece a partir de la magnitud del índice ΔMH y de su significancia estadística de la siguiente forma:

- Un ítem se clasifica en la categoría A, cuando el valor de ΔMH no es estadísticamente diferente de cero o cuando su magnitud es menor que una unidad delta, es decir, $|\Delta MH| < 1$, lo que indica ausencia de DIF relevante para una combinación específica de grupo de referencia y grupo focal.
- Un ítem se clasifica en la categoría C, cuando la magnitud del ΔMH excede 1.5 unidades delta, es decir, $|\Delta MH| > 1.5$, lo que sugiere la presencia de DIF de magnitud elevada.
- Todos los demás ítems se clasifican en la categoría B, correspondiente a DIF de magnitud moderada.
- En las categorías A y C, la significancia estadística se evalúa a un nivel de $\alpha = 0.05$ para cada ítem individual.



Es importante mencionar que los análisis se realizaron empleando el lenguaje de programación *R*, bajo el entorno *Rstudio*. Además, se consideró el paquete *mirt* (Chalmers, 2012) y *diffR* (Magis et al., 2010).

Consideraciones éticas

Como consideraciones éticas se resguardó la información de las personas examinadas, así como cualquier dato sensible. Además, se consideraron a todas las personas participantes que realizaron la pruebas en los años 2021, 2022 y 2023. Por otra parte, el proyecto está regulado por la normativa del Comité Ético Científico de la Universidad de Costa Rica.

RESULTADOS Y DISCUSIÓN

Como primer resultado se presentan en la tabla 3 los estadísticos MH, sus respectivos valores de significancia y el tamaño del efecto de los ítems para los años 2021 al 2023 según sexo. Es importante recordar que, en todos los análisis, el grupo focal estuvo conformado por los hombres, mientras que las mujeres constituyeron el grupo de referencia.

Tabla 3. Estadísticos MH para los ítems de las pruebas 2021 a 2023, según sexo

Ítem	Año								
	2021			2022			2023		
	<i>MH</i>	<i>p</i>	ΔMH	<i>MH</i>	<i>p</i>	ΔMH	<i>MH</i>	<i>p</i>	ΔMH
IT1	5.396	.020	.602	.386	.535	.165	.000	.988	-.016
IT2	.307	.580	-.137	.222	.637	.119	1.078	.299	-.310
IT3	---	---	---	.822	.365	.224	1.747	.186	.356
IT4	3.772	.052	-.446	.000	.988	.016	.437	.509	-.188
IT5	1.652	.199	.280	15.085	.000	1.050	3.197	.074	-.425
IT6	7.016	.008	-.574	1.795	.180	.323	7.483	.006	-.704
IT7	2.134	.144	-.312	.084	.772	.089	---	---	---
IT8	.001	.982	-.005	.000	.987	.008	.760	.383	.234
IT9	.137	.711	-.086	2.102	.147	.353	.507	.477	-.166
IT10	.715	.398	-.193	---	---	---	.000	.988	.010
IT11	.302	.583	-.126	.236	.627	.125	8.967	.003	-.780
IT12	3.601	.058	-.455	6.394	.011	.604	.002	.966	-.024
IT13	---	---	---	.393	.531	.168	.151	.698	.101
IT14	1.457	.227	-.261	.139	.709	.097	.495	.482	-.170



Ítem	Año								
	2021			2022			2023		
	<i>MH</i>	<i>p</i>	<i>ΔMH</i>	<i>MH</i>	<i>p</i>	<i>ΔMH</i>	<i>MH</i>	<i>p</i>	<i>ΔMH</i>
IT15	2.920	.087	-.354	3.218	.073	.442	.794	.373	.245
IT16	.047	.828	.056	2.862	.091	.445	.124	.725	-.088
IT17	1.560	.212	.325	.018	.892	.043	---	---	---
IT18	.569	.451	-.191	.176	.675	.113	.156	.692	-.103
IT19	1.031	.310	.249	.003	.954	.029	3.079	.079	-.478
IT20	2.935	.087	-.368	---	---	---	.025	.874	.059
IT21	.004	.948	.024	.000	.989	.016	3.018	.082	.408
IT22	3.217	.073	-.438	.687	.407	.211	.017	.897	-.040
IT23	---	---	---	.735	.391	.212	.575	.448	-.192
IT24	.855	.355	-.203	.000	.985	.007	.002	.963	.023
IT25	1.463	.226	-.258	.427	.513	.174	2.113	.146	.354
IT26	5.650	.017	-.546	3.298	.069	.463	.165	.685	-.110
IT27	34.553	.000	1.292	3.145	.076	.427	---	---	---
IT28	.000	.996	-.009	.025	.874	.048	12.620	.000	.896
IT29	1.742	.187	-.278	.000	.988	.008	.160	.689	-.111
IT30	5.108	.024	-.489	---	---	---	.634	.426	-.203
IT31	.396	.529	.145	.535	.465	.197	.030	.863	.049
IT32	11.466	.001	.730	.092	.762	.089	.411	.521	-.167
IT33	---	---	---	6.907	.009	.596	12.807	.000	.894
IT34	5.500	.019	.530	.375	.540	.155	.017	.897	.047
IT35	9.125	.003	.696	.323	.570	.137	.205	.651	-.146
IT36	6.769	.009	.599	---	---	---	.233	.629	.123
IT37	.229	.632	-.113	.001	.970	.003	---	---	---
IT38	.089	.765	.091	4.750	.029	.520	16.340	.000	1.070
IT39	.565	.452	.199	2.623	.000	1.149	.036	.850	.066
IT40	7.450	.006	.654	1.964	.161	.389	3.180	.075	-.460

* Mantel-Haenszel (MH); Delta de Mantel-Haenszel (ΔMH); *p* (p-value)

Para la aplicación del año 2021, la mayoría de los ítems no mostró evidencia estadísticamente significativa de funcionamiento diferencial entre mujeres y hombres. Sin embargo, se identificaron



algunos ítems con valores de MH significativos ($p < .05$), tales como IT1, IT6, IT26, IT27, IT30, IT32, IT34, IT35, IT36 y IT40.

En el año 2022, nuevamente se observó que la mayor parte de los ítems presentó un funcionamiento equivalente entre los grupos comparados. Un conjunto un poco más reducido de ítems, entre ellos IT5, IT12, IT33, IT38 y IT39, mostró evidencia estadísticamente significativa de DIF. De forma similar, en la aplicación correspondiente al año 2023, se detectaron algunos ítems con valores significativos del estadístico MH, como IT6, IT11, IT28, IT32, IT33 y IT38, mientras que el resto de los ítems no presentó diferencias relevantes entre mujeres y hombres.

En total, de acuerdo con el método MH para el año 2021 un 27.7% del total de los ítems presenta algún tipo de DIF, para el 2022, solamente un 13.8% y para el 2023 un 16.6%. Sin embargo, según la escala establecida por el *ETS*, solamente el IT27 para el año 2021, el IT5 e IT39 para el 2022 y el IT38 para el año 2023 presentan un DIF tipo B (esto de acuerdo con la clasificación establecida), todos los demás ítems antes mencionados presentan un DIF tipo A, es decir, un DIF despreciable, ya sea porque la magnitud del índice ΔMH es inferior a una unidad delta en valor absoluto o porque no se alcanza significancia estadística. Por lo tanto, aunque en términos porcentuales se identifica la presencia de DIF en una proporción no despreciable de ítems en los distintos años analizados, la magnitud de dicho DIF es, en la gran mayoría de los casos, lo suficientemente pequeña como para no comprometer la equidad ni la interpretación de los puntajes de la prueba.

Con el fin de complementar la evidencia cuantitativa presentada, se incorpora el análisis gráfico de las curvas características de los ítems que presentan DIF moderado, diferenciadas por sexo. Estas curvas obtenidas a partir del modelo TRI de dos parámetros, describen la relación entre el nivel de habilidad latente (θ) y la probabilidad de responder correctamente un ítem, permitiendo examinar visualmente posibles diferencias en el funcionamiento del ítem a lo largo del continuo de habilidad.



Figura 1. Curva característica del ítem 27 del año 2021, según sexo

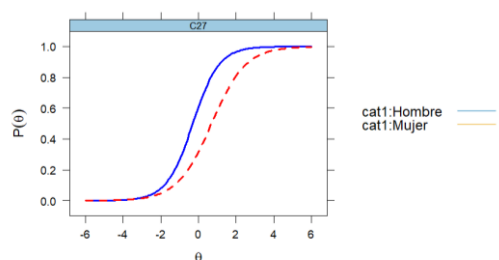


Figura 2. Curva característica del ítem 5 del año 2022, según sexo

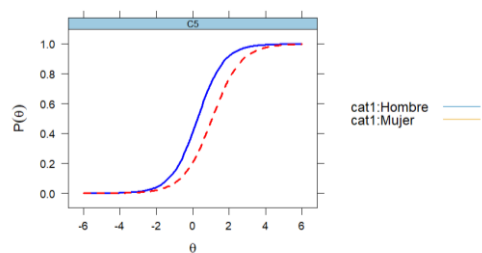


Figura 3. Curva característica del ítem 39 del año 2022, según sexo

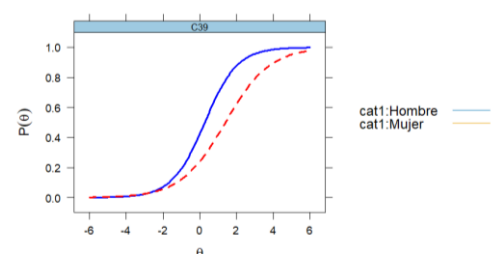
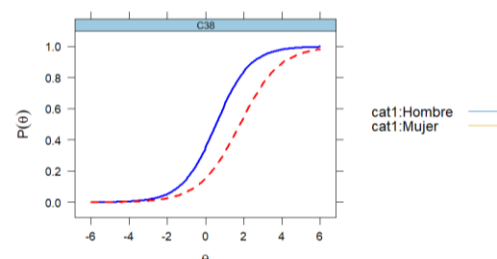


Figura 4. Curva característica del ítem 38 del año 2023, según sexo



En conjunto, las curvas características de los ítems presentadas en las figuras 1,2,3 y 4 muestran diferencias entre hombres y mujeres que son coherentes con la clasificación de un DIF moderado identificada mediante el índice ΔMH . En todos los casos, estas curvas presentan pendientes similares y no evidencian cruces relevantes a lo largo del continuo de habilidad, lo que sugiere un DIF uniforme. Sin embargo, estas diferencias se mantienen acotadas y no se amplían en niveles extremos de habilidad, lo que indica que, si bien algunos ítems favorecen ligeramente al grupo de los hombres, la magnitud del efecto es limitada.

Como segundo resultado se tienen los indicadores de Mantel-Haenszel para la variable tipo de colegio, esto por cada año 2021, 2022, 2023, en los que se consideró como grupo focal a los estudiantes provenientes de colegios privados y como grupo de referencia a aquellos de colegios públicos, los indicadores se presentan en la tabla 5.

Tabla 5. Estadísticos MH para los ítems de las pruebas 2021 a 2023, según tipo de colegio público o privado

Ítem	Año								
	2021			2022			2023		
	<i>MH</i>	<i>p</i>	ΔMH	<i>MH</i>	<i>p</i>	ΔMH	<i>MH</i>	<i>p</i>	ΔMH
IT1	1.112	.292	-.376	.049	.824	.084	.250	.617	.207
IT2	.059	.808	-.089	.511	.475	.205	.012	.914	-.062
IT3	---	---	---	.010	.919	.044	.014	.906	-.054
IT4	2.153	.142	.404	.908	.341	.287	.240	.624	-.170
IT5	2.334	.127	.404	.145	.703	.130	5.692	.017	.636
IT6	1.882	.170	-.372	.000	.991	.012	.892	.345	.286
IT7	.010	.920	.038	.536	.464	.227	---	---	---
IT8	.049	.824	.070	.965	.326	.281	.139	.710	.125
IT9	2.226	.136	-.388	1.961	.161	.390	2.399	.121	-.388
IT10	.866	.352	.245	---	---	---	14.417	.000	1.124
IT11	1.740	.187	.370	1.351	.245	.341	.419	.517	.196
IT12	1.079	.299	.300	.015	.901	.051	.084	.772	-.100
IT13	---	---	---	.738	.390	.256	1.273	.259	-.312
IT14	.952	.329	-.252	1.467	.226	.332	.872	.351	-.247
IT15	.595	.440	-.195	.001	.978	.023	6.103	.013	.736
IT16	2.116	.146	-.380	1.686	.194	.387	4.955	.026	-.560
IT17	.195	.658	-.132	.137	.711	.114	---	---	---
IT18	.524	.469	.202	2.011	.156	.402	2.480	.115	.420
IT19	.402	.526	-.184	9.223	.002	.941	5.558	.018	.698
IT20	3.077	.079	-.453	---	---	---	1.468	.226	.360
IT21	2.228	.136	.435	.019	.889	.058	3.269	.071	-.482
IT22	5.390	.020	.683	.706	.401	.253	2.055	.152	.364
IT23	---	---	---	.941	.332	.265	5.222	.022	-.627
IT24	.045	.832	.065	.244	.621	.145	1.795	.180	-.369
IT25	.002	.961	-.024	1.104	.293	.301	.924	.337	-.265
IT26	.000	.987	.018	1.677	.195	.371	.000	.994	.017
IT27	.291	.589	.151	4.342	.037	.582	---	---	---
IT28	1.806	.179	.344	3.500	.061	.510	.088	.766	.097
IT29	1.223	.269	-.281	.405	.524	.179	.225	.635	-.143
IT30	.804	.370	-.235	---	---	---	.552	.458	-.204
IT31	1.912	.167	-.369	.904	.342	.313	.021	.885	.048
IT32	.056	.813	-.074	7.512	.006	.834	.355	.551	-.179
IT33	---	---	---	.050	.823	.072	4.025	.045	-.547
IT34	.254	.614	.145	.033	.856	.067	.456	.500	-.201
IT35	.099	.753	-.095	.029	.865	.057	.173	.678	-.149
IT36	1.020	.313	.267	---	---	---	.260	.610	.145
IT37	2.337	.126	.389	.027	.869	.057	---	---	---
IT38	3.516	.061	-.548	2.832	.092	.464	.877	.349	.286
IT39	.000	.987	-.011	.512	.474	.218	.326	.568	.178



Ítem	Año								
	2021			2022			2023		
	<i>MH</i>	<i>p</i>	ΔMH	<i>MH</i>	<i>p</i>	ΔMH	<i>MH</i>	<i>p</i>	ΔMH
IT40	.014	.906	.045	5.207	.022	.668	.251	.617	-.151

* Mantel-Haenszel (MH); Delta de Mantel-Haenszel (ΔMH); *p* (p-value)

En la aplicación correspondiente al año 2021, se identificó un número reducido de ítems con valores significativos del estadístico MH ($p < .05$), específicamente el ítem IT22, representando únicamente un 2.7% del total, mientras que el resto de los ítems mostró un funcionamiento equivalente entre estudiantes de colegios públicos y privados. No obstante, el valor del índice ΔMH asociado a este ítem se mantuvo por debajo de 1 unidad delta en valor absoluto, lo que sugiere que la magnitud del DIF es baja.

Para el año 2022, nuevamente se observó que la mayoría de los ítems no evidenció diferencias estadísticamente significativas entre los grupos. Sin embargo, algunos ítems presentaron valores significativos del estadístico MH, entre ellos IT19, IT27, IT32 e IT40. A pesar de alcanzar significancia estadística, los valores de ΔMH correspondientes a estos ítems no superaron el umbral establecido por el *ETS* para ser considerados de magnitud moderada o grande, lo que indica la presencia de DIF despreciable.

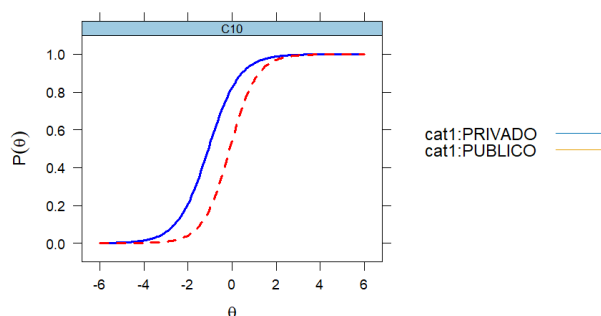
Finalmente, en la aplicación del año 2023, se detectaron ciertos ítems con valores estadísticamente significativos del estadístico MH, tales como IT5, IT10, IT15, IT16, IT19, IT23 e IT33, representando un 19.4% del total de reactivos de la prueba. Sin embargo, al igual que en los años anteriores, aunque estos ítems mostraron diferencias estadísticamente significativas entre estudiantes de colegios públicos y privados, la mayoría de los valores del índice ΔMH se ubicó por debajo de una unidad delta en valor absoluto, lo que, según la escala establecida por el *ETS*, el impacto práctico del DIF es limitado.

En términos globales, de acuerdo con el método MH, la proporción de ítems que presentó algún tipo de DIF asociado al tipo de colegio fue relativamente baja en los tres años analizados. Además, según la clasificación establecida por el *ETS*, la gran mayoría de los ítems se clasificó en la categoría A, correspondiente a DIF despreciable. No se identificaron ítems con DIF tipo C y únicamente el ítem IT10 de la aplicación 2023 alcanzó un valor compatible con DIF tipo B.



De igual forma, en la figura 5 se presenta la curva característica del ítem 10 del año 2023, diferenciada por tipo de colegio, en donde se logra apreciar una diferencia uniforme a lo largo de la curva que favorece al grupo de examinados provenientes de colegios privados. Sin embargo, esta diferencia se mantiene acotada y no se amplifica en los extremos de la habilidad, indicando así que el tamaño del efecto es limitado y coincide con la clasificación de DIF tipo B.

Figura 5. Curva característica del ítem 10 del año 2023, según tipo de colegio



Si bien el contraste estadístico asociado al método de Mantel–Haenszel permite identificar diferencias significativas entre grupos, el *p-valor* no proporciona información sobre la magnitud ni la relevancia práctica de estas diferencias. Dado que el análisis de DIF se orienta a detectar sesgos sustantivos que puedan afectar la equidad de la medición, por ello es necesario complementar la evidencia estadística con índices que miden el tamaño del efecto, particularmente la transformación ΔMH , esta aproximación permite distinguir entre diferencias meramente estadísticas y DIF con implicaciones prácticas para la validez y justicia de la prueba.

La existencia de DIF indica que sujetos que tienen igual grado de habilidad presentan diferente probabilidad de responder correctamente a un ítem, lo que se relaciona con el sesgo, es decir, favorecer a un grupo sobre otro en la evaluación. Como se pudo apreciar en los indicadores de DIF según sexo, para el año 2021 (IT27), 2022 (IT5 y IT39) y 2023 (IT38) las personas que tienen mayor probabilidad de responder correctamente son los hombres, por lo que se puede inferir que favorece más a este grupo. Esto podría ser debido a la manera de representar los datos del encabezado de los ítems, ya que estos incluyen porcentajes y se debe realizar cierta transformación y extracción de datos de diferentes tipos de representación.



Además, algunos ítems que presentan DIF pueden darse porque sean más típicos o comunes para los hombres que para las mujeres, esto quizás se deba al tipo de entrenamiento llevado a cabo por la población. Igualmente, puede deberse como lo indica Rodríguez Jiménez et al., (2021) al manejo de conceptos teóricos y la influencia de los roles de género en la capacidad cognitiva de transferir conocimiento en diferentes contextos. Por otro lado, los ítems detectados con DIF hacen referencia a establecer relaciones numéricas y análisis de datos y de acuerdo con Finch & Hernández Finch (2013) estos ítems afectan de manera diferente a los distintos grupos, pero la presencia de DIF puede ser multicausal.

Cabe mencionar que la presencia de DIF tipo B no implica necesariamente que el ítem sea inválido o que deba eliminarse de la prueba. En contextos de evaluación educativa, el DIF puede reflejar diferencias sistemáticas en las estrategias de resolución, la familiaridad con ciertos formatos de representación o el entrenamiento previo de los grupos como se mencionó anteriormente, sin que esto suponga una medición de un constructo ajeno al razonamiento cuantitativo.

Con respecto al análisis considerando tipo de colegio, solo se obtuvo un ítem para el año 2023 (IT10) favoreciendo a los colegios privados. Esto, en parte, puede ser debido a la formación. Los colegios privados en Costa Rica, en su gran mayoría, contratan docentes que tienen mejor formación matemática que los colegios públicos. En cambio, para los colegios públicos el ente empleador es el Ministerio de Educación Pública, y no existe regulación en cuanto a la formación matemática del profesorado; donde dos de las características que más diferencian a los sectores público-privado son, según Fernández y del Valle (2013) el porcentaje de profesores que poseen un título de licenciatura o superior (37% en el sector público y 50% en el privado).

Además, muchos de los colegios privados dan un mayor énfasis y privilegia la enseñanza de la matemática y, por ende, el desarrollo de habilidades de razonamiento cuantitativo. Esto coincide con el estudio realizado por Fernández y del Valle (2013) al indicar que el estudiantado de colegios privados obtenía mejores resultados en la prueba de matemática de PISA que los de colegios públicos.

Se debe destacar que estos resultados deben interpretarse con cautela. El método MH permite detectar diferencias en la probabilidad de acierto condicionadas por la puntuación total, pero asume



equivalencia en la medición del rasgo latente y se limita principalmente a identificar DIF uniforme. En contraste, los modelos IRT multigrupo permiten evaluar simultáneamente la invariancia de parámetros del ítem, detectar DIF no uniforme y modelar explícitamente el rasgo latente, proporcionando una aproximación más completa al funcionamiento diferencial.

CONCLUSIONES

Primeramente, es indispensable realizar análisis de DIF en el desarrollo de pruebas educativas, tanto en aquellas que son de diagnóstico como las que se emplean con propósitos de admisión. Esto porque permite evaluar la imparcialidad de las pruebas y por lo tanto comprobar las afirmaciones sobre su validez. Al comparar las puntuaciones totales de distintos grupos se puede explorar el rendimiento de estos; sin embargo, no es suficiente para determinar la imparcialidad.

Además, es importante indicar que desarrollar pruebas justas es un valor al que todos los educadores deberían aspirar para garantizar que estas no solo sean precisas para la retroalimentación del estudiantado, sino también para informar las modificaciones en la enseñanza y crear política educativa. Igualmente, las pruebas justas son necesarias para promover y retener a los grupos subrepresentados en los campos de la ciencia, tecnología, ingeniería y matemáticas, y para el desarrollo de habilidades tan indispensables como el razonamiento cuantitativo.

Finalmente, es indispensable atender de la mejor manera posible la exigencia de equidad, la literatura especializada actual recomienda que, para desarrollar pruebas de rendimiento en gran escala, la atención a la diversidad se considere desde el diseño de prueba. Realizar más estudios considerando entrevistas cognitivas o protocolos verbales a grupos de examinados de tal manera que se logre detectar cuáles son las frases o palabras que influyen en las probabilidades de acertar o no un ítem.

REFERENCIAS BIBLIOGRÁFICAS

AERA, APA, NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, Estados Unidos: American Educational Research Association.

Albano, T., French, BF y Vo, TT (2024). Traditional vs Intersectional DIF Analysis: Considerations and a Comparison Using State Testing Data. *Applied Measurement in Education*, 37 (1), 57–7.

<https://doi.org/10.1080/08957347.2024.2311935>



- Arroyo Resino, D. (2018). Funcionamiento diferencial del ítem por sexo en alumnos de educación secundaria. En G. de la Cruz Flores (Coord.), *Experiencias educativas en el aula de Infantil, Primaria y Secundaria*. (pp. 66-76). Eindhoven, NL: Adaya Press. <https://doi.org/1.58909/ad18177496>
- Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Dorans, N.J. & Holland, P.W. (1992). DIF detection and description: Mantel-Haenszel and standardization 1,2. *ETS Research Report Series*, 1992: i-40. <https://doi.org/10.1002/j.2333-8504.1992.tb01440.x>
- Fernández, A & Del Valle R. (2013). Desigualdades en Costa Rica: la brecha entre estudiantes de colegios públicos y privados. Análisis de los resultados de la evaluación Internacional PISA. *Revista CEPAL*, 111, <https://www.un-ilibrary.org/content/journals/16820908/2013/111/3>
- Ferreres Traver, D., Fidalgo Aliste, A. M & Muñiz, J. (2000). Detección del funcionamiento diferencial de los ítems no uniforme: comparación de los métodos Mantel-Haenszel y regresión logística. *Psicothema*, 12(2), 220-225, <https://www.psicothema.com/pdf/552.pdf>
- Fidalgo, A.M & Scalonb, J.D. (2012). Uso dos Métodos Mantel-Haenszel para a Detecção do Funcionamento Diferencial dos Itens e Software Relacionado. *Psicologia: Reflexão e Crítica*, 25(1), <https://doi.org/1.1590/S0102-79722012000100008>
- Fidalgo, AM. (2011). GMHDIF: A Computer Program for Detecting DIF in Dichotomous and Polytomous Items Using Generalized Mantel-Haenszel Statistics. *Applied Psychological Measurement*, 35(3), 247–249. <https://doi.org/1.1177/0146621610375691>
- Finch, WH, y Hernández Finch, ME (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement*, 73(6), 973–993, <https://journals.sagepub.com/doi/epub/10.1177/0013164413494776>
- 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*



<https://ojs.rmie.mx/index.php/rmie/article/view/507>

- Gómez Guillén, K. (2025). Evidencias de Validez en la Valoración del Constructo Satisfacción del Estudiantado hacia los MOOCs como Recurso Complementario en el Curso de Precálculo de la Universidad de Costa Rica durante el I Ciclo del 2024. *Revista Veritas De Difusão Científica*, 6 (2), 5930–5949. <https://doi.org/1.61616/rvdc.v6i2.918>
- Holland, P.W & Wainer, H. (1993). Differential Item Functioning. Educational Testing Service. Routledge.
- Jodoin, M. G., & 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 (4), 329–349, https://doi.org/10.1207/S15324818AME1404_2
- Magis, D., Bél and, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, 42, 847–862. <https://doi.org/10.3758/BRM.42.3.847>
- Martínez Arias, R., Hernández Lloreda, M.J., Hernández Lloreda, M.V. (2014). Psicometría. Alianza Editorial.
- Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. *CBE life sciences education*, 16(1), rm2. <https://doi.org/1.1187/cbe.16-10-0307>
- Montero-Rojas, E & Moreira-Mora, T.E. (2017). Funcionamiento diferencial del ítem en pruebas de español y matemática en estudiantes reportados con TDAH. *Actualidades Investigativas en Educación*, 17(2), <http://dx.doi.org/1.15517/aie.v17i1.28661>
- Moreira Mora, T. (2008). Funcionamiento diferencial del ítem en pruebas de matemática para la educación media. *Actualidades en Psicología*, 22, 91-113. <https://doi.org/1.15517/apv22i109.16>



- Ordóñez Gutiérrez, G. (2025). Evidencias de Validez de la Prueba de Habilidades Cuantitativas Asociadas a su Estructura Interna. *Ciencia Y Reflexión*, 4(1), 1184–1202. <https://doi.org/1.70747/crv4i1.162>
- Pérez-Gil, J.A., Chacón Moscoso, S., & Moreno Rodríguez, R. (2000). Validez de constructo: el uso del análisis factorial exploratorio- confirmatorio para obtener evidencias de validez. *Psicothema*. 12 (2). <https://www.psicothema.com/pi?pii=601>
- Portell, M & Vives, J. (2024). *Investigación en Psicología y logopedia: Introducción a los diseños experimentales, cuasiexperimentales y ex post facto*. Servei de publicacions de la UAB.
- Rodríguez Jiménez, O.R., Pineda Miranda, B.A., & Sierra Muñoz, T. (2021). Estudio de Funcionamiento diferencial de los ítems de la prueba EXIM. *Encuentro Internacional De Educación En Ingeniería*. <https://doi.org/1.26507/ponencia.1966>
- Sánchez Mendiola, M., García Minjarez, M., Martínez González A., & Buzo Casanova, E. (2020). El examen de ingreso a la Universidad Nacional Autónoma de México: Evidencias de validez de una prueba de Alto impacto y gran escala. *Revista Iberoamericana de Evaluación Educativa*, 13 (2), 107-120, <https://doi.org/1.15366/riee202.13.2.006>
- Silva Morales, F & Santelices Etchegaray, M.V. (2016). Funcionamiento diferencial del ítem en una Evaluación Estandarizada según necesidades Educativas Especiales Transitorias. *Revista Iberoamericana de Evaluación Educativa*, 9 (1), 145-160, <https://doi.org/10.15366/riee2016.9.1.009>
- Stockemer, D. (2019). *Quantitative Methods for the Social Sciences: A Practical Introduction with Examples in SPSS and Stata*. <https://doi.org/1.1007/978-3-319-99118-4>
- Suh, Y. (2016). Effect Size Measures for Differential Item Functioning in a Multidimensional IRT Model. *Journal of Educational Measurement*, 53 (4). <https://doi.org/10.1111/jedm.12123>

