ESTUDIO DE FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS DE LA PRUEBA EXIM

Olga Rosalba Rodríguez Jiménez, Brayam Alexander Pineda Miranda, Ángela Tatiana Sierra Muñoz

Universidad Nacional de Colombia
Bogotá, Colombia

Resumen

El funcionamiento diferencial de los ítems afecta la validez de las pruebas, en tanto la probabilidad de respuesta depende de variables diferentes a tener el atributo que se mide. En este sentido la psicometría ha estudiado el asunto y ha sugerido una serie de procedimientos para hallar DIF en pruebas estandarizadas. A partir de lo anterior, el trabajo presenta el estudio de DIF realizado para cada uno de los componentes del examen de Ciencias básicas de ingeniería -EXIM- que aplica anualmente ACOFI. El trabajo se realizó para la aplicación del año 2020 y un total de 850 estudiantes, siendo las mujeres consideradas el grupo focal. Haciendo uso del programa JASP 0.14.1 y la librería difR en RStudio 1.4.17.17 se calculó el DIF con el método de Regresión logística. Los resultados evidencian presencia de DIF tanto uniforme como no uniforme en tres de los cuatro componentes, siendo el de matemáticas el que evidencia mayor número de ítems y el de Química el que no presenta ítems con DIF. A partir del análisis cualitativo de cada pregunta se evidencian algunas diferencias en el manejo de conceptos evaluados o en la influencia cultural debida al rol de género que facilita la aproximación práctica a la resolución de las situaciones planteadas en la prueba.

Palabras clave: DIF; EXIM, género

Abstract

The Differential Item Functioning (DIF) affects the validity of the tests, since the probability of response depends on variables other than having the attribute being measured. In this sense, psychometrics has studied the issue and has suggested a series of procedures to find DIF in standardized tests. In

this way, the paper presents the DIF study carried out for each of the components of the Basic Engineering Sciences Test -EXIM- applied annually by ACOFI. The work was carried out for the 2020 application and a total of 850 students, being women considered the focal group. Using the JASP 0.14.1 program and the difR library in RStudio 1.4.17.17, the DIF was calculated using Logistic Regression methods. The results show the presence of both uniform and non-uniform DIF in three of the four components, with the mathematics component showing the highest number of items and the chemistry component showing no items with DIF. The qualitative analysis of each question reveals some differences in the handling of the concepts evaluated or in the cultural influence due to the gender role that facilitates the practical approach to the resolution of the situations presented in the test.

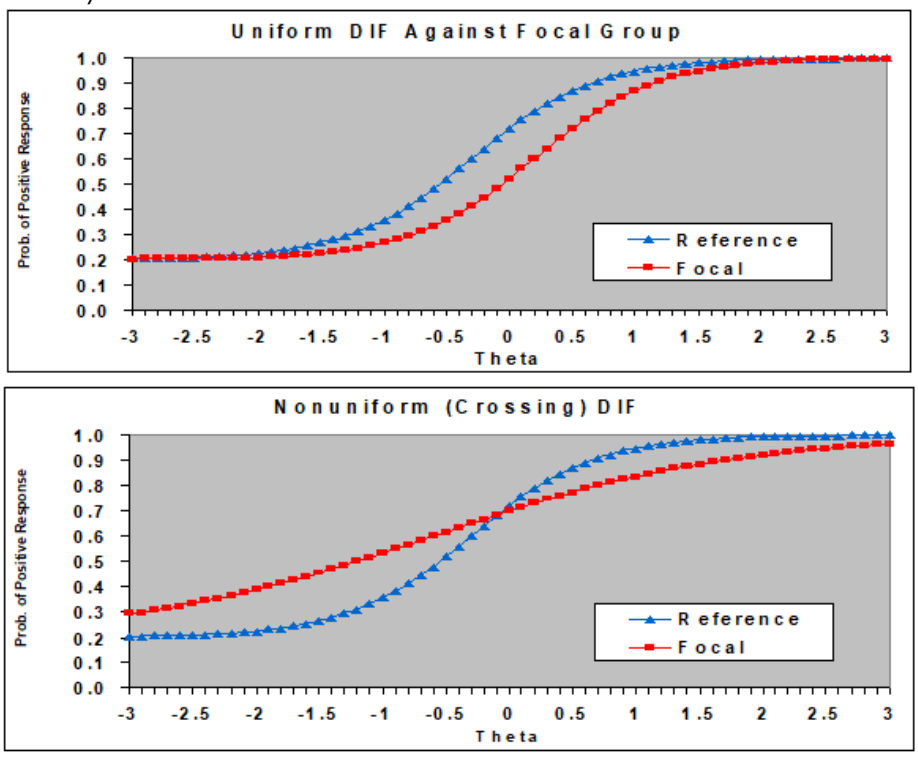
***Keywords**: DIF; EXIM; gender*

1. Introducción

En pruebas estandarizadas es necesario valorar sus calidades psicométricas, una de las cuales hace referencia a la validez, la cual hace referencia a calidad de las inferencias que se realizan a partir de los puntajes de una prueba (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2018) y es justamente esta propiedad la que se afecta cuando los puntajes en la prueba dependen de aspectos diferentes al atributo que se mide, en este caso se puede hablar de funcionamiento diferencial del ítem -DIF-.

El DIF se refiere a un procedimiento estadístico empleado para identificar si un ítem se comporta a nivel psicométrico de una forma distinta en función de un atributo diferente al que mide la prueba (Gómez Benito & Hidalgo, 1997), tales como la etnia, el nivel socioeconómico o el sexo. En este contexto se debe tener dos grupos, uno focal y el de comparación, el cual se denomina de referencia, aunque no existe un criterio específico para asignar a los grupos, se espera que el grupo focal sea sobre el cual se espera un DIF desventajoso (Reveco-Quiroz, 2021) Este concepto se diferencia de impacto y sesgo en tanto el primero se refiere a una diferencia real en el atributo (Moreira-Mora, 2008) y el sesgo a la explicación que se daría a esta diferencia. Existen dos tipos de DIF a saber: Uniforme y no uniforme, tal como se ilustra en la figura 1.

Figura 1. DIF UNiforme y no uniforme



Tomado de Abad (2010)

Como se aprecia en la figura 1 en el DIF uniforme la probabilidad de respuesta al ítem se mantiene constante a lo largo del atributo, en contraste con el DIF no uniforme en el cual esta probabilidad puede variar a lo largo de este continuo; como se aprecia para el grupo focal en contraste con el grupo de referencia, primero se tiene una probabilidad de respuesta más alta para las personas de bajo nivel de atributo y luego esta probabilidad cambia para los de mayor atributo.

Se han propuesto distintos métodos para determinar el DIF de los ítems, se clasifican en métodos empíricos y teóricos, acorde con Moreira-Mora (2008) en el primer grupo se encuentra el método del Delta Gráfico, el chi-cuadrado de Scheuneman, el Mantel-Haenszel y el método de regresión logística. Y en el segundo los métodos basados en la Teoría de Respuesta al Ítem (IRT).

Acorde con Gómez & Hidalgo (1997) la regresión logística estima la probabilidad de acierto al ítem para cada grupo. La variable dependiente es el ítem y el nivel de habilidad y el grupo junto con su interacción corresponde a la variable independiente. "El modelo general es

$$p(x=1)=e^{z}/(1+e^{z})$$
$$z=\tau 0+\tau 1\ H+\tau 2\ G+\tau 3(HG)$$

Donde x la respuesta al item, H el nivel de habilidad del sujeto, G el grupo de pertenencia (R o F) y HX G el producto de dos variables individuales, H y G. El parámetro $\tau 2$ corresponde a las diferencias de grupo en la ejecución del item, mientras que el parámetro $\tau 3$, corresponde a la interacción entre grupo y nivel de habilidad. Un item muestra DIF uniforme si $\tau 2$, es distinto de cero y $\tau 3$ es cero, y DIF no uniforme si $\tau 3$ es distinto de cero, sea o no $\tau 2$, igual a cero. La hipótesis

nula a comprobar es si ambos parámetros no son significativamente distintos de cero" (Gómez & Hidalgo, 1997, p. 13). El uso de este método indica la magnitud y dirección del DIF y fue el método empleado en el presente estudio.

2. Metodología

Para el análisis de datos se usaron JASP 0.14.1 y la librería difR en RStudio 1.4.17.17 (Magis et al., 2020). La muestra constó de 264 mujeres y 500 hombres para el área de biología; 291 mujeres y 537 hombres para física; 281 mujeres y 569 hombres en matemáticas; y por último, 272 mujeres y 536 hombres para el área de química.

El análisis se llevó a cabo en tres etapas, en la primera se compararon las medias de las puntuaciones obtenidas por hombres y mujeres en la prueba, así como en cada una de las competencias evaluadas por medio del estadístico U Mann-Whitney para muestras independientes con el fin de identificar si se presentaban diferencias significativas entre los puntajes.

Posteriormente, se evaluó el funcionamiento diferencial de los ítems por medio de la regresión logística con la prueba de razón de verosimilitud y se aplicó el procedimiento de purificación de los ítems incluido en el paquete estadístico (Bandalos, 2018; Magis et al., 2020).

Una vez identificados los ítems que presentaron DIF, se realizaron entrevistas individuales con docentes de cada una de las áreas de conocimiento evaluadas para obtener su criterio experto sobre una posible explicación de las diferencias encontradas.

3. Resultados

En la tabla 1 se presentan la media y desviación estándar de hombres y mujeres en cada una de las pruebas y en las competencias. Es posible observar que, salvo por química, las puntuaciones medias de hombres son siempre mayores a las de las mujeres.

Tabla 2. Estadísticos descriptivos de hombres y mujeres por prueba y competencia

Competencia	Estadísticos	Matemáticas		Física		Química		Biología	
		Hombres	*Mujeres*	*Hombres*	*Mujeres*	*Hombres*	*Mujeres*	*Hombres*	*Mujeres*
	Media	47,385	45,038	40,632	37,479	39,926	39,369	55,208	52,966
	DE	8,275	6,931	7,567	6,644	7,565	8,093	9,306	9,147
Capacidad de abstracción, análisis y síntesis	*Media*	48,067	46,601	39,176	35,928	40,739	41,515	55,652	53,545
	DE	11,174	9,825	10,530	9,902	10,848	9,930	12,088	11,838
Capacidad de aplicar los conocimientos en la práctica	*Media*	44,981	42,335	43,808	40,255	39,939	40,570	54,764	52,386
	DE	8,976	7,863	10,498	9,540	11,094	10,665	9,883	10,248
Capacidad para identificar, plantear y resolver problemas	*Media*	49,602	46,999	38,515	35,788	35,782	35,786	–	–
	DE	9,828	9,022	8,825	7,494	9,268	9,882		

DE: Desviación estándar. – No se evalúa.

La tabla 2 presenta los *p valores* obtenidos en las pruebas U de Mann-Whitney para comparación de grupos, así como el tamaño del efecto calculado por medio de correlación biserial. Con un nivel de significancia de 0.05, se encontraron diferencias significativas entre hombres y mujeres en todas las áreas de conocimiento y sus competencias, excepto en química. Cabe resaltar que aunque el tamaño del efecto obtenido en todas las pruebas fue bajo, este puede variar debido a la disparidad muestral (Ledesma, Macbeth, y De Kohan, 2010)

Respecto al cálculo de ítems con funcionamiento diferencial, las figuras 2 a la 5, muestran con color rojo las preguntas presentan este comportamiento, cada fila representa una competencia diferente. En el eje de las ordenadas se presenta el valor del estadístico en la prueba y en las abscisas cada uno de los ítems, la línea horizontal que cruza cada gráfico representa el umbral de detección, siendo este de 5.995 para la regresión. Por último, cabe aclarar que el indicativo del ítem que aparece en el eje x, es el que toma dentro de la competencia y no dentro de la prueba en general.

Tabla 2. P valores y tamaño del efecto de la comparación de medias entre hombres y mujeres

Competencia	Matemáticas		Física		Química		Biología	
	p valor	*rbis*	*p valor*	*rbis*	*p valor*	*rbis*	*p valor*	*rbis*
	<0,001	*-0,190*	*<0,001*	*-0,229*	*0,345*	*0,060*	*0,002*	*-0,142*
Capacidad de abstracción, análisis y síntesis	0,062	-0,085	<0,001	-0,170	0,323	0,059	0,021	-0,115
Capacidad de aplicar los conocimientos en la práctica	<0,010	-0,184	<0,001	-0,197	0,440	0,047	0,002	-0,098
Capacidad para identificar, plantear y resolver problemas	<0,010	-0,171	<0,001	-0,173	0,995	-0,007	–	–

Como es posible observar en las figuras 2 a la 5, mediante la prueba de regresión logística, se encontraron siete (7) preguntas con funcionamiento diferencial para la prueba de matemáticas, cuatro (4) para Física, uno (1) Química y cuatro (4) para Biología. Estos, sin embargo, presentan distribuciones tanto uniformes como no uniformes.

Figura 2. ítems con DIF en Matemáticas

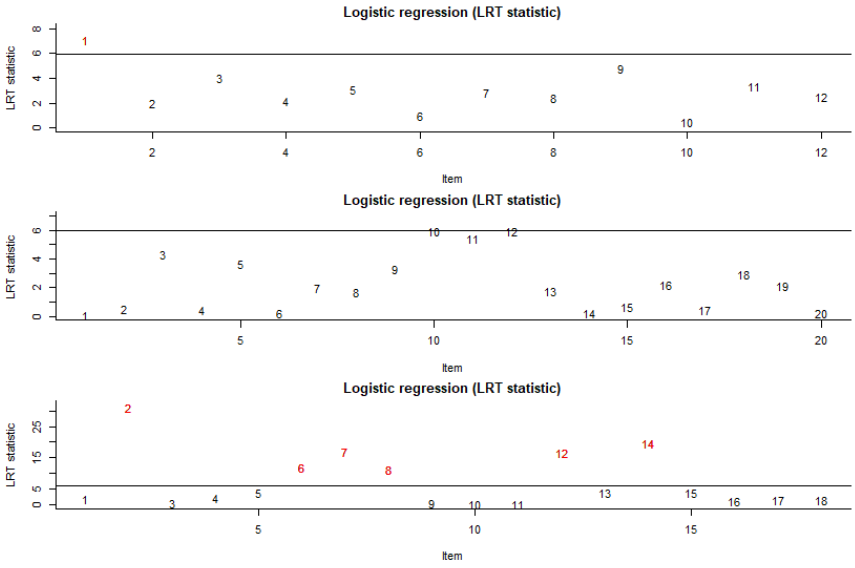


Figura 3. ítems con DIF en Física

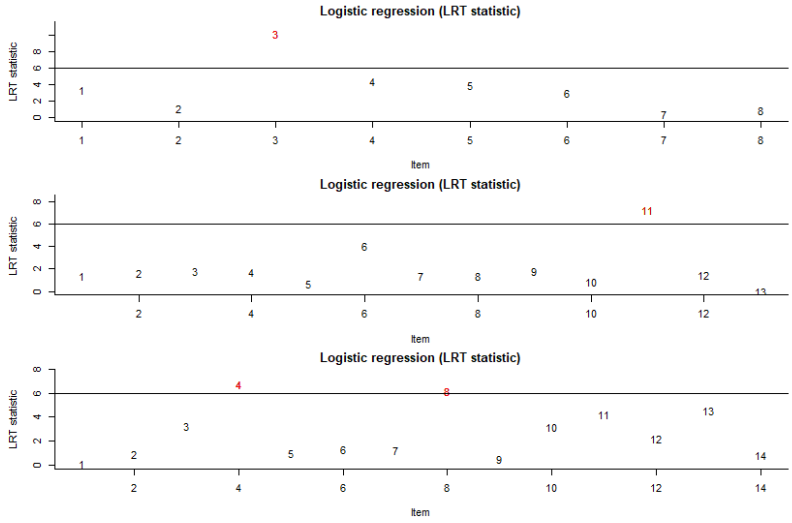


Figura 4. ítems con DIF en Química

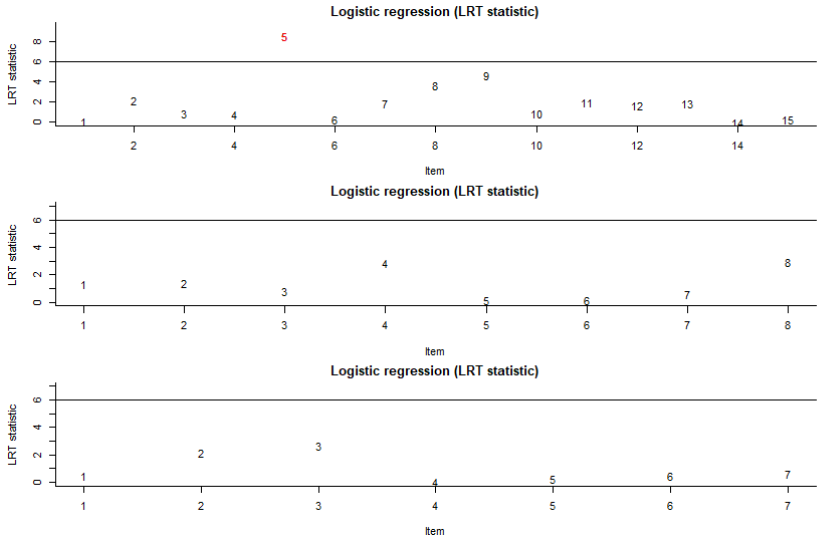


Figura 5. ítems con DIF en Biología

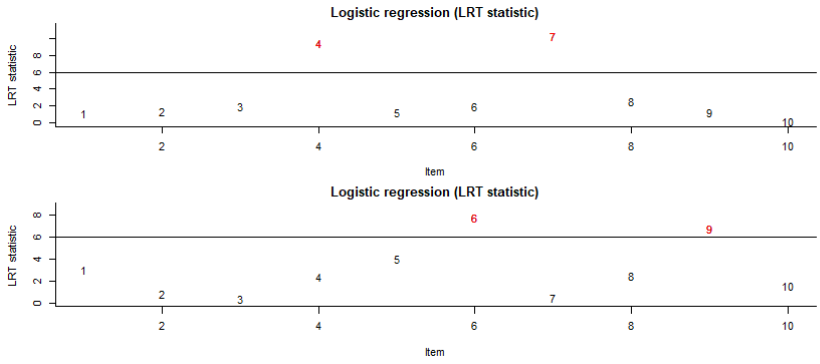


Tabla 3. Resumen de las preguntas con DIF

Matemáticas		
Capacidad de abstracción, análisis y síntesis		
1	No uniforme	Focal
Capacidad para identificar, plantear y resolver problemas		
2	No uniforme	Focal
6	Uniforme	Referencia
7	No uniforme	Focal
8	No uniforme	Focal
12	Uniforme	Referencia
14	Uniforme	Referencia
Física		
Capacidad de abstracción, análisis y síntesis		
3	Uniforme	Referencia
Capacidad de aplicar los conocimientos en la práctica		
11	No uniforme	Focal
Capacidad para identificar, plantear y resolver problemas		
4	No uniforme	Focal
8	Uniforme	Focal
Química		
Capacidad de abstracción, análisis y síntesis		
5	No uniforme	Focal
Biología		
Capacidad de abstracción, análisis y síntesis		
4	No uniforme	Focal
7	No uniforme	Focal
Capacidad de aplicar los conocimientos en la práctica		
6	No uniforme	Referencia
9	No uniforme	Referencia

La tabla 3 resume el tipo de funcionamiento diferencial encontrado por ítem y el grupo que se ve beneficiado en caso del DIF uniforme y aquel cuyo valor termina siendo superior a mayor nivel de rasgo en el caso del DIF no uniforme, esta información se basa en el cálculo a partir del modelo de regresión logística, tomando a los hombres como grupo de referencia y a las mujeres como grupo focal.

En la Tabla 3 se observa que la única competencia en la cual se encontró al menos un ítem con DIF fue Capacidad de abstracción, análisis y síntesis, hallándose que en todos los casos, salvo en el de Física sus ítems presentaron funcionamiento no uniforme y con sesgo hacia las mujeres con mayor atributo. Así también, en el área de Matemáticas en donde se encuentran más ítems con estas características.

En total fueron 16 preguntas con funcionamiento diferencial, se encontró que el tamaño del efecto era moderado para uno de ellos según el estadístico Jodoin & Gierl, esto fue para la pregunta número 2 de la tercera competencia de Matemáticas, los demás ítems tuvieron un valor mínimo en el indicador.

Con base en estos resultados, se reunió un grupo de docentes expertos en cada una de las áreas de interés y se les presentaron los ítems identificados, así como el flujo de respuestas para hombres y mujeres. Se encuentra en la revisión que aquellos ítems que estaban relacionados con entender algún concepto, estaban relacionados con presentar funcionamiento no uniforme el cual, en el nivel de rasgo más alto, favoreció a las mujeres. Así mismo, aquellas preguntas que favorecen a los hombres y eran atravesadas por posibles factores socioculturales, presentaban DIF uniforme. Por último, aquellos ítems en los cuales los evaluados debían analizar con detenimiento lo que se preguntaba favorecen a los hombres.

4. Discusión y conclusiones

El estudio calculó el DIF en cada componente de la prueba EXIM, encontrando 16 ítems con DIF. La retroalimentación cualitativa realizada por los docentes de cada área coincidió en que es necesaria la claridad de conceptos básicos en ingeniería para poderlos aplicar eficazmente en la resolución de los ítems presentados. Sin embargo, son de destacar dos aspectos claves en el análisis del comportamiento diferencial de los mismos: el manejo de conceptos teóricos y la influencia de los roles de género en la capacidad de transferencia del conocimiento obtenido a diversos contextos, tales como habilidades cognitivas diferenciadas para la lectura de gráficas.

La evidencia sugiere que los estilos de aprendizaje (López-Aguado, 2011) y las capacidades intelectuales y académicas, como las evaluadas por las pruebas de inteligencia Wechsler (Díaz y Lynn, 2016) o la prueba PISA (González de San Román y de La Rica, 2016), pueden sustentar empíricamente dichas diferencias. En otros atributos como los estilos de aprendizaje, el estudio de López-Aguado (2011) evaluó, mediante el *Cuestionario de Estilos de Aprendizaje* CHAEA, a una muestra de 805 estudiantes universitarios en España. Se encontró que el estilo de aprendizaje más frecuente en los estudiantes fue el reflexivo, caracterizado por personas "receptivas, analíticas y exhaustivas, observadoras, pacientes, detallistas, investigadoras y

asimiladoras" (Mayaute, 2011, p.74). No obstante, se encontraron diferencias significativas ($p<0,05$) entre ambos géneros. Si bien los hombres presentaron mayormente un estilo activo y pragmático, caracterizados respectivamente, por ser personas que prefieren las experiencias nuevas y planificar y solucionar los problemas de forma práctica (Mayaute, 2011) y las mujeres presentaron un estilo reflexivo; en la facultad de ingeniería, tanto hombres como mujeres presentaron niveles muy cercanos en el estilo pragmático. Esto es congruente con el análisis de DIF por cada competencia, donde las mujeres tuvieron un mejor desempeño en el de "capacidad de abstracción, análisis y síntesis" y ambos grupos tuvieron un desempeño equiparable en "capacidad de aplicar los conocimientos en la práctica" y "capacidad para identificar, plantear y resolver problemas".

5. Referencias

- Abad, P. (2010). Funcionamiento Diferencial de los Ítems. Apuntes de clase. Máster en Ciencias del Comportamiento y la Salud. Universidad Autónoma de Madrid.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). Estándares para pruebas educativas y psicológicas (M. Lieve, Trans.). Washington, DC: American Educational Research Association. (Original work published 2014).
- Bandalos, D. L. (2018). Measurement theory and applications for the social sciences. Guilford Publications.
- Díaz, R. R. y Lynn, R. (2016). Sex differences on the WAIS-IV in Chile. *Mankind Quarterly,* Vol *57, N°* 1, p. 52.
- Gómez Benito, J. & Hidalgo Montesinos, M. D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: una revisión metodológica. Anuario de Psicología, Vol 74, p.p. 3-32.
- González de San Román, A. y de La Rica, S. (2016). Gender Gaps in PISA Test Scores: The Impact of Social Norms and the Mother's Transmission of Role Attitudes. Estudios De Economía Aplicada, Vol 34, N° 1, p.p. 79-108.
- Ledesma, R., Macbeth, G. y De Kohan, N. C. (2010). Tamaño del efecto: revisión teórica y aplicaciones con el sistema estadístico ViSta. Artículos desde 2007 hasta 2013. Revista latinoamericana de psicología, Vol 40, N° 3, p.p. 425-439.
- López-Aguado, M. (2011). Estilos de aprendizaje. Diferencias por género, curso y titulación. Revista De Estilos De Aprendizaje, Vol 4, 7.
- Magis, D., Beland, S., & Raiche, G. (2020). DifR: Collection of methods to detect dichotomous differential item functioning (dif). https://CRAN.R-project.org/package=difR
- Mayaute, L. M. E. (2011). Análisis psicométrico del Cuestionario de Honey y Alonso de Estilos de Aprendizaje (CHAEA) con los modelos de la Teoría Clásica de los Test y de Rasch. *Persona,* Vol 14, 71-109.
- Moreira-Mora, T.E. (2008). El funcionamiento diferencial del ítem: un asunto de validez y equidad. Avances en medición, Vol 6, p.p. 5-14.
- Reveco-Quiroz, P.F. (2021). Evaluación psicométrica de sesgo en estructuras dimensionales empíricas presentes en pruebas de matemáticas rendidas por estudiantes chilenos, en base a los métodos de Wald y de Mantel-Haenszel para la detección del funcionamiento diferencial del ítem. Tesis de Maestría [no publicada]. Universidad Católica de Chile.

Sobre los autores

- **Olga Rosalba Rodríguez Jiménez**: Psicóloga. Magister en Educación y en Métodos en Ciencias del Comportamiento. Doctora en Psicología y Educación. Profesora asociada de la Universidad Nacional de Colombia. orrodriguezj@unal.edu.co
- **Brayam Alexander Pineda Miranda**: Psicólogo, Especialista en analítica estratégica de datos, Fundación Universitaria Konrad Lorenz. bapinedam@unal.edu.co
- **Ángela Tatiana Sierra Muñoz**: Estudiante de Psicología. Universidad Nacional de Colombia. ansierram@unal.edu.co

---