

DOCUMENTOS DE **TRABAJO**

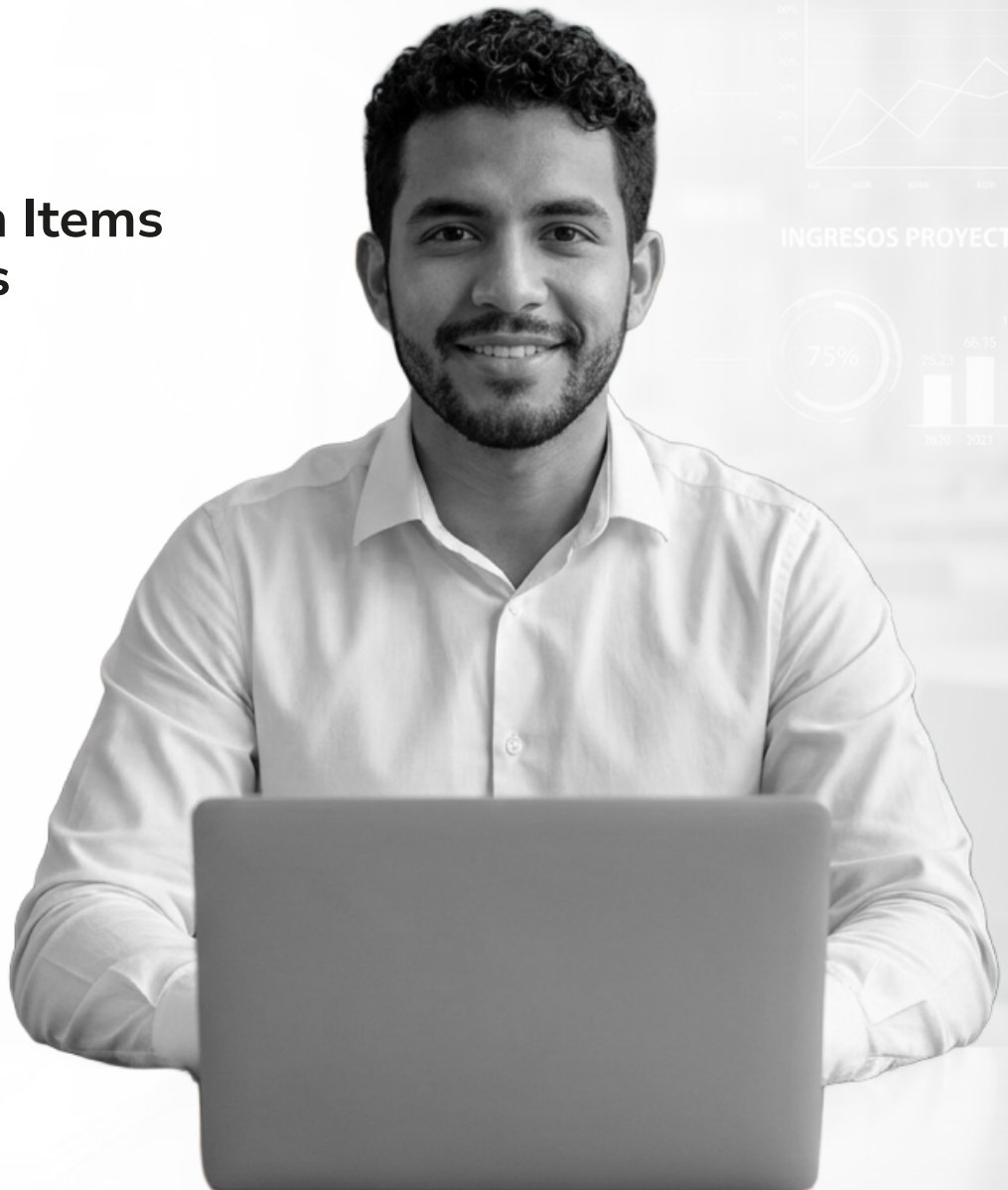
Saber Investigar

Instituto Colombiano para la Evaluación de la Educación

N° 26

**Análisis de DIF en Items
politómicos en las
pruebas Saber**

**Alexander Calderon
Andres Rodríguez
Sebastian Argüello**



TASA DE GRADUACIÓN



INGRESOS PROYECTADOS



Septiembre 2026



Serie Documentos de Trabajo Saber Investigar

Edición No. 26

Septiembre, 2026

Edición digital

ISSN: 2954-6583

Instituto Colombiano para la Evaluación de la Educación

Oficina de Gestión de Proyectos de Investigación

Calle 26 N. 69-76, Edificio Elemento, Torre II, piso 18, Bogotá,
D. C.

Teléfono: (601) 4841410

proyectosinvestigacion@icfes.gov.co

<https://www.icfes.gov.co>

Directora general

Elizabeth Blandón Bermúdez

Jefe de Oficina de Gestión de Proyectos de Investigación

Paola Guio Veloza

Autores

Alexander Calderon

Andres Rodríguez

Sebastian Argüello

Advertencia:

El contenido de este documento es el resultado de investigaciones y obras protegidas por la legislación nacional e internacional. No se autoriza su reproducción, utilización ni explotación a NINGÚN tercero. Solo se autoriza su uso para fines exclusivamente académicos. Esta información no podrá ser alterada, modificada o enmendada.

Citar este documento en estilo APA así:

Calderón, A., Rodríguez, A., & Argüello, S. (2026). Análisis de DIF en ítems politómicos en las pruebas Saber. (Documentos de Trabajo Saber Investigar No. 26). Instituto Colombiano para la Evaluación de la Educación (Icfes).

<https://www.icfes.gov.co/web/guest/saber-investigar>

DOCUMENTOS
DE TRABAJO

**Saber
Investigar**

La serie de Documentos
de Trabajo Saber
Investigar del Icfes tiene
como propósito hacer
divulgación de los
resultados de
investigaciones sobre
evaluación y análisis de la
calidad de la educación.

Análisis de DIF en ítems politómicos en las pruebas Saber¹

Alexander Calderón²

Andrés Rodríguez²

Sebastián Argüello²

Resumen

El análisis de Funcionamiento Diferencial del Ítem (DIF) es un procedimiento fundamental en la validación psicométrica de un instrumento, al aportar evidencia sobre la comparabilidad de las mediciones entre distintos grupos poblacionales. En el contexto de la estimación de puntajes mediante modelos de Teoría de Respuesta al Ítem (TRI), resulta pertinente evaluar el DIF mediante metodologías paramétricas basadas en los parámetros de los ítems. El objetivo de este documento es presentar el uso del Índice No Compensatorio de DIF (NCDIF) para el análisis de DIF en ítems politómicos ajustados mediante el Modelo de Respuesta Graduada (GRM), así como describir su implementación, incluyendo la equiparación, la prueba de hipótesis y la estimación del tamaño del efecto. Además, se propone un ajuste a los puntos de corte del tamaño del efecto para mejorar la clasificación del estadístico en muestras de gran tamaño. Finalmente, la metodología se aplica al análisis de DIF de la prueba no cognitiva de Acciones y Actitudes Ciudadanas (AyA) del examen Saber 3°, 5°, 7° y 9° entre 2022 y 2023, administrada por el Icfes. Los resultados sugieren que la metodología y los ajustes propuestos mejoran la precisión del análisis de DIF en ítems politómicos para evaluaciones a gran escala, aunque estos hallazgos deben validarse mediante estudios adicionales.

Palabras claves: DIF, NCDIF, TRI, ítems politómicos, GRM, Acciones y Actitudes Ciudadanas.

¹Las ideas, opiniones, tesis y argumentos expresados son de propiedad exclusiva de los autores y no representan el punto de vista del Icfes.

²Subdirección de Estadísticas, Instituto Colombiano para la Evaluación de la Educación - Icfes, Colombia

DIF Analysis in Polytomous Items in Saber Assessments¹

Alexander Calderón²

Andrés Rodríguez²

Sebastián Argüello²

Resumen

Differential Item Functioning (DIF) analysis is a fundamental procedure in the psychometric validation of an instrument, as it provides evidence regarding the comparability of measurements across different population groups. In the context of score estimation using Item Response Theory (IRT) models, it is appropriate to evaluate DIF through parametric methodologies based on item parameters. The objective of this paper is to present the use of the Non-Compensatory Differential Item Functioning Index (NCDIF) for DIF analysis in polytomous items fitted with the Graded Response Model (GRM), as well as to describe its implementation, including score equating, hypothesis testing, and effect size estimation. In addition, an adjustment to the effect size cutoffs is proposed to improve the classification of the statistic in large samples. Finally, the methodology is applied to the DIF analysis of the non-cognitive Actions and Citizenship Attitudes (AyA) assessment from the Saber 3, 5, 7, and 9 examinations administered by Icfes between 2022 and 2023. The results suggest that the proposed methodology and adjustments improve the accuracy of DIF analysis for polytomous items in large-scale assessments, although these findings should be validated through further studies.

Keywords: DIF, NCDIF, IRT, polytomous items, GRM, Actions and Attitudes (AyA)

¹The ideas, opinions, theses, and arguments expressed are the sole responsibility of the authors and do not represent the views of Icfes.

²Subdirección de Estadísticas, Instituto Colombiano para la Evaluación de la Educación - Icfes, Colombia

Contenido

1. Introducción	4
2. Revisión de literatura y estado del arte	5
3. Metodología	7
3.1. Modelos TRI para ítems politómicos	7
3.2. Funcionamiento Diferencial del Ítem (DIF)	9
3.3. El índice No Compensatorio de DIF (NCDIF)	10
3.4. Tipos de DIF	11
3.5. Equiparación y purificación	12
3.6. Prueba de hipótesis y propuesta del tamaño del efecto	14
4. Datos	16
4.1. El examen Saber 3°, 5°, 7° y 9°	16
4.2. Acciones y Actitudes Ciudadanas (AyA) en Saber 5°, 7° y 9°	17
5. Resultados	18
6. Conclusiones	22

1 Introducción

Las evaluaciones educativas estandarizadas son herramientas importantes para medir y monitorear la calidad educativa y constituyen un insumo relevante para la toma de decisiones en política pública. En estas evaluaciones es usual aplicar a los estudiantes instrumentos cognitivos para medir su desempeño en distintas áreas del conocimiento, como lenguaje o matemáticas. Adicionalmente, se aplican cuestionarios de carácter no cognitivo o contextual, orientados a la medición de habilidades socioemocionales, factores asociados, acciones y actitudes ciudadanas, características sociodemográficas, entre otras.

Para la obtención de puntajes del rasgo latente a partir de los instrumentos aplicados es habitual ajustar modelos de la Teoría de Respuesta al Ítem (TRI), que permiten relacionar la probabilidad de respuesta correcta o de una categoría específica, con los parámetros que caracterizan los ítems. En el caso de instrumentos cognitivos, en donde predominan ítems de respuesta correcta se ajustan modelos dicotómicos, mientras que para instrumentos no cognitivos donde predominan ítems con múltiples categorías de tipo nominal u ordinal se ajustan modelos politómicos.

El uso de los modelos de la TRI facilita la construcción de escalas y la implementación de procedimientos de equiparación y anclaje entre aplicaciones. Uno de los supuestos de estos modelos es la invarianza de los ítems en las poblaciones, esto es, que los parámetros que caracterizan los ítems se mantienen estables en las poblaciones, lo cual garantiza la comparabilidad de los puntajes reportados. En la práctica, para garantizar la comparabilidad de los puntajes entre diferentes aplicaciones, se realizan análisis de Funcionamiento Diferencial del Ítem (DIF) entre las poblaciones de las diferentes aplicaciones.

El DIF se refiere al favorecimiento de una población sobre otra a pesar de que sus individuos tengan el mismo nivel de rasgo latente, esto desde la perspectiva de la TRI indica diferencias en la probabilidad de responder correctamente para ítems dicotómicos, o de seleccionar una categoría específica (o mayor o igual, en modelos de respuesta graduada) en ítems politómicos, entre grupos de examinados, aún cuando estos tienen el mismo nivel en el rasgo latente medido. Esta característica pone en riesgo la equidad de las pruebas, ya que podría introducir sesgos que comprometan la validez de las interpretaciones basadas en los puntajes obtenidos.

Cuando se identifican ítems con DIF entre diferentes aplicaciones, es posible tomar de-

cisiones orientadas a evitar la introducción de sesgos en la medición, garantizando que las comparaciones entre los grupos sean justas. Este proceso permite revisar, recalibrar o incluso excluir ítems que puedan afectar la comparabilidad de los resultados. En consecuencia, se fortalece la calidad psicométrica de las evaluaciones, especialmente en contextos de alto impacto, como las evaluaciones educativas, donde la interpretación justa de los puntajes resulta determinante.

En este documento se describe la metodología empleada para el análisis de DIF en ítems politómicos utilizando el Índice No Compensatorio de DIF (NCDIF), el cual adopta un enfoque basado en los parámetros de los ítems para medir el DIF. Además, desde la Subdirección de Estadísticas del Icfes se propone un ajuste a la medida de tamaño del efecto utilizada, con el fin de mitigar la sensibilidad del estadístico ante grandes tamaños de muestra. Finalmente, esta metodología se aplica a la prueba de Acciones y Actitudes Ciudadanas (AyA) del examen Saber 3°, 5°, 7° y 9°, aplicado por el Icfes, que evalúa las competencias del pensamiento ciudadano en los estudiantes.

2 Revisión de literatura y estado del arte

El análisis de DIF es un paso fundamental dentro del proceso de validación en evaluación y uso de pruebas, en la medida en que permite verificar que los ítems se comportan de manera invariante entre distintas poblaciones, garantizando así que las inferencias derivadas de los puntajes sean justas y comparables (American Educational Research Association, American Psychological Association, y National Council on Measurement in Education, 2014). En el marco de los estándares para la evaluación educativa y psicológica, la ausencia de DIF constituye una fuente de evidencia de validez relacionada con la estructura interna del instrumento. En este sentido, un ítem presenta DIF cuando, condicionado a un mismo nivel de la habilidad latente, la probabilidad de respuesta difiere entre grupos, es decir, cuando se tiene evidencia estadística de que algún o algunos presuntos factores ajenos influyen diferencialmente en el desempeño de los examinados (Ackerman, 1992).

En el ámbito de las evaluaciones educativas a gran escala, los análisis de DIF han formado parte sistemática del control de calidad psicométrico, desde las primeras propuestas metodológicas (Angoff y Ford, 1973; Scheuneman, 1979) hasta las aplicaciones contemporáneas en estudios internacionales como PISA, TIMSS y PIRLS (Penfield y Camilli, 2007). En estos contextos, la comparabilidad de resultados entre grupos, países y ciclos de aplicación depende en gran medida del funcionamiento equivalente de los ítems entre subpoblaciones.

La presencia de DIF no identificado puede distorsionar tanto las comparaciones entre grupos como las estimaciones de tendencias en el tiempo, comprometiendo las conclusiones sobre la calidad educativa (Hambleton, 2006; Sireci y Rios, 2013).

En este marco, los análisis de DIF suelen orientarse en dos direcciones complementarias: por un lado, la evaluación de la invarianza entre subpoblaciones dentro de una misma aplicación (por ejemplo, según género, zona o región), y por otro, el análisis de la estabilidad de las propiedades métricas de los ítems a lo largo de distintas aplicaciones del instrumento. Cuando el DIF se manifiesta de manera sistemática entre aplicaciones temporales, se habla de *Item Parameter Drift*, considerado un caso particular de DIF en contextos longitudinales (Goldstein, 1983; Holland y Wainer, 1993). Este fenómeno puede originarse en cambios en las poblaciones evaluadas, variaciones curriculares, modificaciones en los procedimientos de aplicación o exposición previa a los ítems (Engels, Lüdtke, y Robitzsch, 2026; Wells, Subkoviak, y Serlin, 2002), por lo que su detección resulta clave para preservar la validez de las comparaciones de tendencia (Kolen y Brennan, 2014).

Los procedimientos de detección de DIF se han desarrollado ampliamente para ítems dicotómicos y suelen clasificarse en dos enfoques: los métodos paramétricos, basados en los parámetros de los ítems, y los métodos no paramétricos, basados en la comparación de probabilidades observadas de respuesta entre grupos (Magis, Béland, Tuerlinckx, y De Boeck, 2010). Entre estos últimos se destacan el estadístico de Mantel-Haenszel (Holland y Thayer, 1988; Mantel y Haenszel, 1959), el método de estandarización (standardization) (Dorans y Kulick, 1986) y el procedimiento SIBTEST (Shealy y Stout, 1993). La regresión logística, por su parte, ocupa una posición intermedia al permitir la detección de DIF uniforme y no uniforme mediante la comparación de modelos anidados (Swaminathan y Rogers, 1990). En el enfoque paramétrico, basado en la TRI, se destacan la prueba de razón de verosimilitud (Thissen, Steinberg, y Wainer, 1988), la prueba chi-cuadrado de Lord (Lord, 1980) y el método de Raju, fundamentado en el área entre las curvas características de los ítems (Raju, 1988, 1990).

Aunque el desarrollo metodológico del DIF se ha centrado principalmente en ítems dicotómicos, los avances para ítems politómicos han sido más recientes (Penfield y Camilli, 2007). Este tipo de ítems es frecuente en escalas tipo Likert, evaluaciones con crédito parcial y pruebas de desempeño con rúbricas, y suele modelarse mediante el Modelo de Respuesta Graduada (GRM) (Samejima, 1969) y el Modelo de Crédito Parcial Generalizado (Muraki, 1992). En coherencia con estos modelos, se han propuesto extensiones de los métodos de

DIF: Cohen, Kim, y Baker (1993) desarrollaron versiones politómicas de la prueba de Lord y de las medidas de área de Raju, French y Miller (1996) extendieron la regresión logística, y Chang, Mazzeo, y Roussos (1996) propusieron Poly-SIBTEST para ítems con respuestas ordenadas.

Dentro de los métodos paramétricos, el NCDIF se ha consolidado como una herramienta utilizada para la identificación de DIF, al cuantificar el área promedio ponderada entre las curvas características del ítem para el grupo focal y el de referencia, asignando mayor peso a las regiones de la escala donde se concentra el grupo focal (Oshima y Morris, 2008). Este índice generaliza la medida de área sin signo de Raju y puede formularse bajo modelos de TRI unidimensionales (Raju, van der Linden, y Fleer, 1995), su extensión al caso politómico, junto con el método de Replicación de Parámetros del Ítem (IPR) para evaluar la significancia estadística, fue desarrollada por Raju y cols. (2009).

Finalmente, para el cómputo del NCDIF es necesario que los parámetros de los ítems se encuentren en una escala común, lo cual se logra mediante procedimientos de equiparación como los métodos de Stocking-Lord (Stocking y Lord, 1983) o de Haebara (Haebara, 1980). Este procedimiento es fundamental, ya que errores en la equiparación pueden generar detecciones erróneas de ítems con DIF. Asimismo, en contextos de gran escala, donde los tamaños muestrales son elevados, la alta sensibilidad de las pruebas estadísticas puede llevar a identificar diferencias despreciables, por lo que se recomienda complementar los contrastes de hipótesis con medidas de tamaño del efecto (Hambleton, 2006; Sireci y Rios, 2013; Wright y Oshima, 2015).

3 Metodología

3.1. Modelos TRI para ítems politómicos

Para ítems politómicos existen diferentes modelos de TRI cuya selección depende de la naturaleza de ordenamiento de las categorías de respuesta. En el caso de categorías no ordenadas, el Modelo de Respuesta Nominal (NRM) es uno de los más utilizados, dado que permite analizar ítems cuyas opciones no siguen un orden jerárquico definido. Por su parte, cuando las categorías son ordenadas, el Modelo de Crédito Parcial (PCM), asume que las categorías reflejan niveles crecientes del rasgo latente evaluado. Finalmente, el GRM, propuesto por Samejima y ampliamente reconocido en la literatura, se centra en la probabilidad de alcanzar o superar un nivel de respuesta, lo que lo hace ideal para escalas de tipo Likert y favorece su amplia aplicación en investigaciones educativas y psicológicas (Samejima, 1969).

En el GRM la probabilidad de que la respuesta U_{in} de un evaluado n con habilidad θ_n al ítem i sea la categoría k , con $k = 1, 2, \dots, K$, se expresa como la diferencia de las probabilidades acumuladas, es decir,

$$P_i(U_{in} = k|\theta_n) = P(U_{in} \geq k-1|\theta_n) - P(U_{in} \geq k|\theta_n) \quad (1)$$

donde para $k = 1, \dots, K-1$ el logit de las probabilidades acumuladas puede expresarse como expresión lineal de la habilidad, esto es

$$P(U_{in} \geq k|\theta_n) = P(U_{in} \geq k|\theta_n, \alpha_i, \Delta_{ik}) = \frac{1}{1 + \exp(-(\alpha_i \theta_n - \Delta_{ik}))} \quad (2)$$

las cuales coinciden en forma funcional con las curvas características de ítem de un modelo de dos parámetros. Cabe notar que $P(U_{in} \geq 0|\theta) = 1$ y $P(U_{in} \geq K|\theta) = 0$.

En este modelo, el parámetro α_i recibe el nombre de pendiente del ítem i , el cual es común a todas las categorías, y los parámetros Δ_{ik} son los interceptos del ítem i asociados a la categoría k . Esta parametrización se denomina en la literatura como parametrización en términos de interceptos y pendiente, la cual es usual en los softwares psicométricos por la facilidad de su implementación en cálculos numéricos.

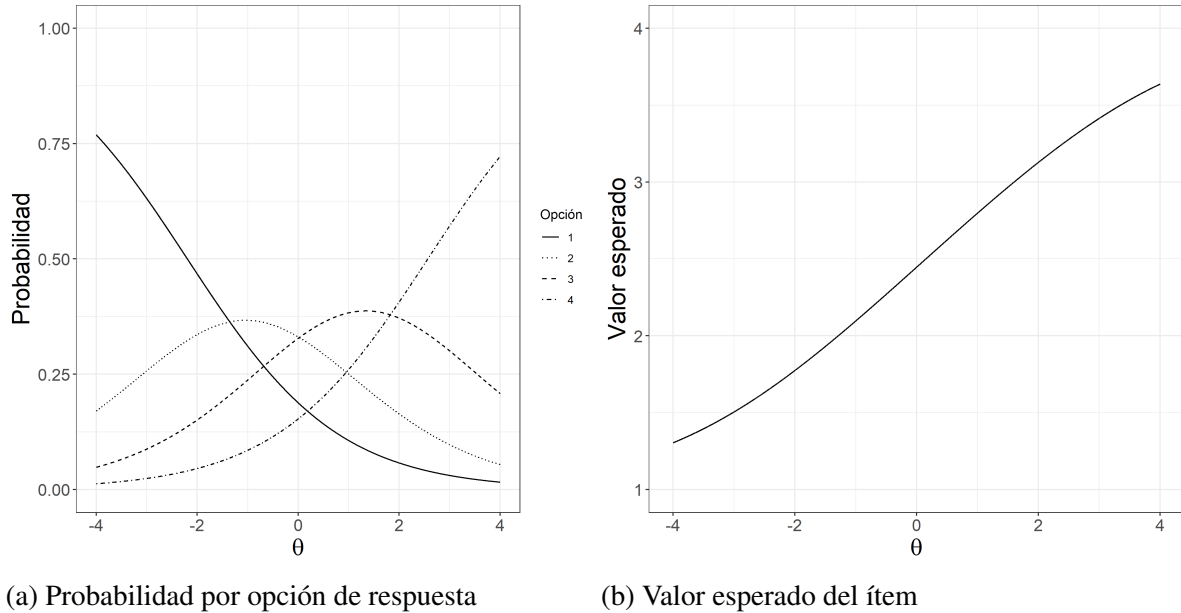
Por otra parte, existe una parametrización alternativa del modelo ampliamente utilizada en la literatura debido a la facilidad de su interpretación, conocida como la parametrización en términos de discriminación, dificultad media y umbrales. Bajo esta formulación, las probabilidades acumuladas del ítem i asociadas a la categoría k están dadas por la expresión

$$P(U_{in} \geq k|\theta) = P(U_{in} \geq k|\theta_n, a_i, b_{ik}) = \frac{1}{1 + \exp(-a_i(\theta_n - b_{ik}))} \quad (3)$$

donde a_i es el parámetro de discriminación del ítem i , b_{ik} son los parámetros de los umbrales del ítem i asociados a la categoría k , y $\bar{b}_i = \frac{1}{K-1} \sum_{k=1}^{K-1} b_{ik}$ es la dificultad media del ítem i .

Considerando que el GRM es un modelo de TRI, comparte los supuestos de esta teoría. En particular, asume la unidimensionalidad, la independencia local de los ítems y la invarianza de los parámetros. Este último resulta particularmente relevante para garantizar una escala común que asegure la comparabilidad de las mediciones entre diferentes aplicaciones.

Figura 1: Probabilidad de opciones y curva del valor esperado para un ítem con cuatro categorías.



En la Figura 1 se muestran las curvas de opciones de respuesta y del valor esperado del ítem para un ítem de 4 categorías con parámetros de discriminación de 0.67 y de umbrales -2.20, 0.11 y 2.57, con una dificultad media de 0.16. Las curvas en el panel (a) muestran el comportamiento de las probabilidades asociadas a cada opción de respuesta de acuerdo con el modelo, las cuales en su totalidad tienen valores entre 0 y 1, y su suma es de 1 para todos los valores de θ . Por otra parte, la figura en el panel (b) muestra el valor esperado del ítem, también llamado valor verdadero, $S(\theta)$ (ver ecuación (5)), el cual indica el valor de la categoría predicha por el modelo para un evaluado con habilidad θ y oscila entre 1 y el número de categorías, en este caso 4.

3.2. Funcionamiento Diferencial del Ítem (DIF)

Cuando un ítem se administra en diferentes poblaciones o aplicaciones, su comportamiento puede variar entre grupos. En la literatura, el estudio de estas diferencias se aborda a través del análisis de DIF. En este contexto, se considera que un ítem presenta DIF cuando, para un mismo nivel de habilidad, la probabilidad de respuesta difiere en función de la pertenencia a un grupo determinado, como distintas aplicaciones o características poblacionales (por ejemplo, género o zona de residencia), entre otros (Holland y Wainer, 1993).

Si bien las metodologías para el análisis de DIF en ítems politómicos han sido menos

desarrolladas que en el caso dicotómico, diversos métodos ampliamente utilizados, como el estadístico de Mantel-Haenszel (MH) o la Regresión Logística (LR), han sido extendidos para su aplicación en este tipo de ítems. Considerando las particularidades de las evaluaciones a gran escala, tales como grandes tamaños de muestra, diseños de armado incompleto y el uso de modelos de TRI para la estimación de puntajes, en este documento se estudia el desempeño del NCDIF, un estadístico de enfoque paramétrico empleado ampliamente en la detección de DIF en ítems dicotómicos y politómicos bajo estas condiciones.

Es importante señalar que, aunque estas metodologías permiten identificar la presencia de DIF en los ítems, la interpretación de sus causas requiere análisis complementarios de carácter cualitativo. Estos son realizados por el equipo constructor de la prueba, con el fin de identificar, comprender y, en lo posible, prevenir los factores que originan este comportamiento.

3.3. El índice No Compensatorio de DIF (NCDIF)

El NCDIF es un estadístico que compara la probabilidad de que evaluados con la misma habilidad respondan correctamente al ítem entre los grupos de interés (Raju y cols., 1995). El NCDIF se define como el valor esperado del cuadrado de la diferencia entre los puntajes estimados del ítem para ambos grupos, es decir,

$$NCDIF_i = E_F \left[(S_{iF}(\theta) - S_{iR}(\theta))^2 \right] \quad (4)$$

donde $S_{iG}(\theta)$ es el puntaje verdadero por el modelo del ítem i para un evaluado perteneciente al grupo G , una asociada al grupo focal (F) y una al grupo de referencia (R) y la esperanza se toma bajo la distribución del grupo focal. Este puntaje verdadero, también llamado puntaje esperado del ítem, para ítems politómicos se puede expresar como

$$S_i(\theta) = \sum_{k=1}^K X_k \cdot P_i(k|\theta) \quad (5)$$

donde X_k es el puntaje asociado a la categoría k y $P_i(k|\theta) = P(U_{in} = k|\theta, \alpha_i, \vec{\Delta}_i)$ es la probabilidad de que la respuesta del evaluado n con habilidad θ al ítem i sea la categoría k . En la práctica, se acostumbra que el puntaje de la categoría coincide con el valor de la categoría, es decir, $X_k = k$.

Sin embargo, cada examinado puede tener una curva de puntaje verdadero S_{iG} asociada a cada grupo G . Si $S_{iF} = S_{iR}$, entonces el puntaje verdadero no depende del grupo al que pertenece el evaluado, pero si son diferentes estamos bajo la presencia de DIF.

A partir de las ecuaciones (4) y (5) podemos expresar el NCDIF para un ítem politómico i en términos de las probabilidades asociadas a las categorías k de este ítem, es decir,

$$NCDIF_i = E_F \left[\left(\sum_{k=1}^K k \cdot (P_{iF}(k|\theta) - P_{iR}(k|\theta)) \right)^2 \right] \quad (6)$$

Es importante mencionar que para el cálculo del NCDIF se asume que las escalas de ambos grupos son iguales, no obstante, en la práctica, la escala de cada grupo se establece con base en la distribución de habilidades de la población analizada. Por esta razón, antes de estimar el valor del estadístico, se debe garantizar la comparabilidad de las escalas de ambos grupos, lo cual requiere la aplicación de un procedimiento de equiparación.

3.4. Tipos de DIF

El funcionamiento diferencial en ítems politómicos puede manifestarse mediante distintos patrones que reflejan la forma en que las probabilidades de respuesta varían entre los grupos evaluados a lo largo del rasgo latente. La identificación de estos patrones permite comprender mejor la naturaleza de las diferencias observadas y facilita la interpretación de los resultados del análisis de DIF. Siguiendo la clasificación tradicional (Gómez-Benito, Hidalgo, y Guilera, 2010), el DIF puede clasificarse en uniforme, no uniforme y mixto, este último corresponde a una combinación de los dos primeros.

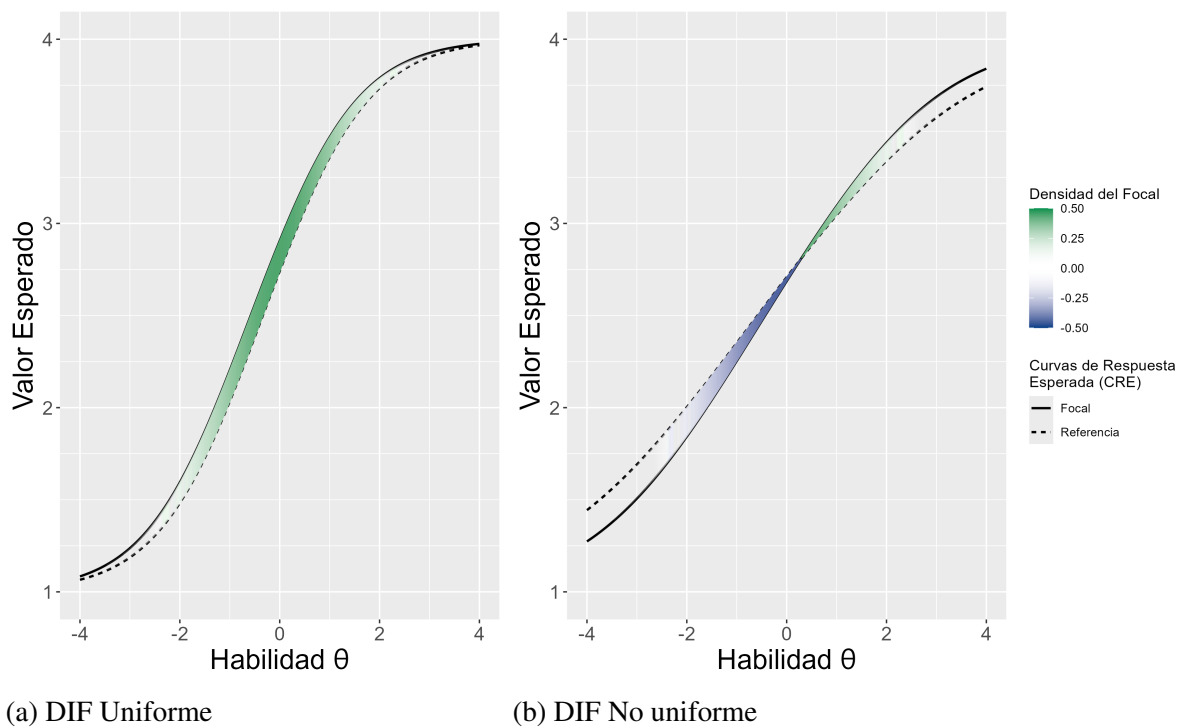
El DIF uniforme se presenta cuando un grupo tiene una probabilidad sistemáticamente mayor o menor de responder en las categorías superiores del ítem a lo largo de toda la escala del rasgo evaluado. En este caso, la ventaja o desventaja se mantiene relativamente constante para todos los niveles de habilidad. En el modelo GRM, este patrón suele asociarse principalmente con diferencias en los parámetros de dificultad (umbrales), mientras que la discriminación permanece similar entre los grupos. Gráficamente, las funciones características del ítem tienden a mantenerse separadas sin presentar cruces importantes.

El DIF no uniforme ocurre cuando las diferencias entre grupos dependen del nivel del rasgo latente. En estos casos, un grupo puede verse favorecido en ciertos niveles de habilidad, mientras que el otro presenta una ventaja en niveles diferentes. Este patrón suele reflejarse en

cruces entre las funciones características del ítem. En el modelo GRM, puede estar asociado a diferencias en la discriminación, aunque en ítems politómicos también puede originarse por diferencias en los umbrales, incluso cuando los cambios en discriminación son reducidos.

En ítems con cuatro categorías de respuesta, estos patrones se reflejan en diferencias en las probabilidades de seleccionar las distintas categorías entre los grupos comparados. Su identificación complementa la información proporcionada por los índices de magnitud del efecto y facilita la interpretación de los resultados del análisis de DIF. En la Figura 2 se presentan dos Curvas de Respuesta Esperada (CRE) correspondientes a ítems politómicos de cuatro categorías con DIF uniforme y no uniforme.

Figura 2: Curvas esperadas de ítems politómicos bajo DIF uniforme y no uniforme.



3.5. Equiparación y purificación

Con el objetivo de garantizar la comparabilidad entre los puntajes estimados de ambos grupos, en la literatura se proponen diferentes metodologías de equiparación orientadas a establecer una escala común. Entre estas metodologías, se destacan aquellas que permiten construir dicha escala mediante la aplicación de transformaciones lineales tanto a las estimaciones de la habilidad de los evaluados como a los parámetros de los ítems (Kolen y Brennan,

2014).

Estas transformaciones son factibles porque, en la TRI el origen y la unidad de la escala de la habilidad pueden fijarse de forma arbitraria y, al aplicar transformaciones lineales tanto a las habilidades como a los parámetros de ítem estimados, las probabilidades asociadas a las diferentes categorías permanecen invariantes. Esta propiedad es conocida en la literatura como indeterminación de la escala o problema de identificabilidad.

Para resolver esta arbitrariedad, al estimar un modelo se fija la escala de la habilidad mediante un supuesto de normalidad, asumiendo que la distribución de la habilidad de la población utilizada en la estimación presenta una media 0 y desviación estándar igual a 1. No obstante, cuando se busca evaluar diferencias entre las curvas características de ítem asociadas a dos o más grupos de evaluados obtenidas a partir de estimaciones independientes, no es posible garantizar que se encuentren en una misma escala, por lo que no son directamente comparables.

En consecuencia, para construir una escala común entre ambos grupos se realiza un proceso de equiparación en el cual se transforman los parámetros del grupo de referencia a la escala del grupo focal. Uno de los procedimientos más utilizados es la metodología de equiparación de Stocking-Lord, en la cual se definen unas constantes de escala kA y kB que permiten la construcción de dicha escala común (Stocking y Lord, 1983).

El procedimiento de Stocking-Lord establece la transformación de los parámetros en la parametrización interceptos - pendiente del grupo de referencia a la escala del grupo focal de la siguiente forma:

$$\alpha' = \alpha/kA; \quad \Delta'_k = \Delta_k + kB \cdot \alpha/kA$$

Mientras que para la parametrización discriminación - dificultad - umbrales la transformación está dada por:

$$a' = a/kA; \quad b'_k = kA \cdot b_k + kB; \quad \bar{b}' = kA \cdot \bar{b} + kB$$

El procedimiento de Stocking-Lord permite establecer una escala común entre ambos grupos, sin embargo, al igual que otros métodos de equiparación, es sensible a la presencia de DIF entre los ítems utilizados en el proceso. Dado que, en general, inicialmente se desconoce que ítems presentan DIF y por tanto, no es posible emplear únicamente estos ítems

en el proceso de equiparación, en los análisis de DIF se recurre a un procedimiento iterativo conocido como purificación (Candell y Drasgow, 1988).

La purificación de ítems inicia empleando la totalidad de los ítems en el proceso de equiparación. Posteriormente, se realiza un primer análisis de DIF bajo estas condiciones. Luego, los ítems identificados con un tamaño de efecto grande de DIF se eliminan del conjunto de anclaje, y se lleva a cabo una nueva equiparación utilizando este conjunto reducido de ítems.

Este procedimiento se repite de manera iterativa, alternando entre la equiparación y la evaluación del DIF, hasta que el conjunto de ítems utilizado para equiparar coincida con aquel que presenta un tamaño de efecto no significativo de DIF. El proceso de purificación finaliza cuando dicho conjunto permanece igual en dos iteraciones consecutivas, en particular, puede completarse en tan solo dos iteraciones si ningún ítem presenta DIF.

3.6. Prueba de hipótesis y propuesta del tamaño del efecto

Para evaluar la hipótesis de DIF en un ítem, teniendo en cuenta que el NCDIF no posee una distribución teórica conocida, se emplea la metodología de Replicación de Parámetros del Ítem (IPR) con el fin de poner a prueba el supuesto de que los puntajes estimados del ítem son iguales en ambos grupos.

Inicialmente, bajo la hipótesis nula, se simulan duplas de parámetros para los grupos focal y de referencia a partir de los parámetros estimados del grupo focal y de sus correspondientes matrices de covarianza. Posteriormente, para cada dupla se calcula el valor del NCDIF. A partir de estos valores se construye la distribución empírica del estadístico, la cual permite evaluar la significancia del DIF (Oshima, Raju, y Nanda, 2006).

Sin embargo, es importante mencionar que las pruebas de replicación de parámetros son sensibles a los tamaños poblacionales de los grupos de interés. En particular, para tamaños poblacionales grandes, la hipótesis nula tiende a rechazarse en un número considerable de ítems, incluso cuando las diferencias entre las curvas características del ítem entre ambos grupos son pequeñas.

Como alternativa a la prueba de hipótesis basada en la IPR, en la literatura se han propuesto diferentes clasificaciones del tamaño del efecto de DIF. En particular, para ítems dicotómicos se han establecido puntos de corte del NCDIF que permiten clasificar la magnitud del DIF como negligible, moderada o grande. No obstante, aunque existen umbrales gene-

rales que permiten identificar la presencia de un DIF con magnitud considerable, estos no garantizan clasificar con precisión la magnitud del DIF. Por lo anterior, a continuación se desarrolla una propuesta específica que integra distintos enfoques planteados en la literatura.

Inicialmente, podemos expresar la diferencia entre los puntajes esperados en términos de las probabilidades acumuladas:

$$\sum_{k=1}^K k \cdot (P_{iF}(k|\theta_n) - P_{iR}(k|\theta_n)) = \sum_{k=1}^{K-1} (P(U_{inF} \geq k|\theta_n) - P(U_{inR} \geq k|\theta_n)) \quad (7)$$

Reemplazando en la ecuación (6) la expresión anterior podemos acotar el valor del NCDIF en términos de las diferencias cuadradas de las probabilidades acumuladas asociadas a cada categoría, más concretamente:

$$NCDIF_i = E \left[\left(\sum_{k=1}^{K-1} (P(U_{inF} \geq k|\theta_n) - P(U_{inR} \geq k|\theta_n)) \right)^2 \right] \quad (8a)$$

$$\leq (K-1) \sum_{k=0}^{K-1} E \left[(P(U_{inF} \geq k|\theta_n) - P(U_{inR} \geq k|\theta_n))^2 \right] \quad (8b)$$

donde la desigualdad anterior es consecuencia de la siguiente propiedad de los números reales

$$\left(\sum_{k=1}^K a_k \right)^2 \leq \left(\sum_{k=1}^K |a_k| \right)^2 \leq K \sum_{k=1}^K a_k^2 \quad (9)$$

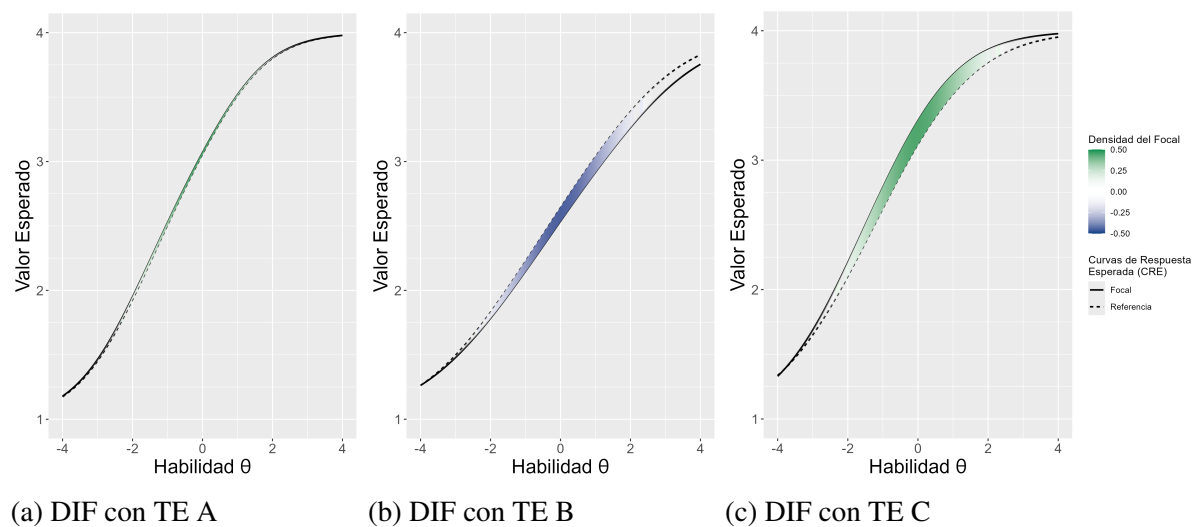
Así, como cada uno de los términos presentes en la sumatoria en el lado derecho de la ecuación (8b) es análogo al cálculo del NCDIF en el caso del modelo logístico de dos parámetros, 2PL, de Birnbaum (Birnbaum, Lord, y Novick, 1968), en tanto la expresión $P(U_{inG} \geq k|\theta_n)$ presenta la misma forma funcional que la de un ítem en un modelo 2PL (ver ecuación (3)), es posible aplicar la regla general de Tamaño del Efecto (TE) propuesta por Wright y Oshima (2015) (Wright y Oshima, 2015) a cada uno de estos términos.

En consecuencia, se obtiene la siguiente regla de TE para el valor del NCDIF en el caso de ítems de respuesta graduada:

$$TE(NCDIF_i) = \begin{cases} A & \text{si } NCDIF_j \leq 0,003 \cdot (K-1)^2 \\ B & \text{si } 0,003 \cdot (K-1)^2 < NCDIF_j \leq 0,008 \cdot (K-1)^2 \\ C & \text{si } NCDIF_j > 0,008 \cdot (K-1)^2 \end{cases} \quad (10)$$

donde K es el número de categorías y $(K - 1)^2$ aparece al tener en cuenta el factor común $(K - 1)$, y la cantidad de términos presentes en la sumatoria, también $(K - 1)$. Es importante mencionar que en el caso en el que existan solo dos categorías, es decir, en caso de tener un ítem dicotómico, la clasificación del TE propuesta en la ecuación (10) coincide con la clasificación propuesta en Wright y Oshima (2015) (Wright y Oshima, 2015). En la Figura 3 se presentan 3 Curvas de Respuesta Esperada (CRE) del ítem empleando los puntos de cortes propuestos en $TE(NCDIF_i)$.

Figura 3: Ejemplos de CRE de acuerdo con la clasificación del Tamaño del Efecto propuesta.



4 Datos

4.1. El examen Saber 3°, 5°, 7° y 9°

El examen Saber 3°, 5°, 7°, 9° es una evaluación nacional que se aplica periódicamente a estudiantes de educación básica de todo el país, con el fin de conocer el desarrollo de sus competencias en diferentes áreas. El Icfes y el Ministerio de Educación Nacional (MEN) restructuraron el examen, actualizando los marcos conceptuales y generando un nuevo ciclo de la prueba con línea base en 2022.

El examen evalúa en 3° las pruebas de Lectura, Matemáticas y el cuestionario Socioeconómico. Para los grados 5°, 7° y 9° a unos estudiantes les evalúa las pruebas de Lectura, Matemáticas, Escritura, Ciencias Naturales y Educación Ambiental y a otros las pruebas de Lectura, Matemáticas, Acciones y Actitudes junto con Pensamiento Ciudadano, en ambos casos se aplica el cuestionario Socioeconómico.

Ante la dificultad, complejidad y magnitud logística de aplicar a todos los estudiantes matriculados de estos grados, más de 3.100.000 estudiantes en más de 14.000 colegios en todo el país, el examen se aplica a una muestra aleatoria de estudiantes con el fin de obtener estimaciones nacionales por algunas variables de interés como sector (público y privado), zona (urbana y rural), región. El tamaño de la muestra es de alrededor de 180.000 estudiantes, con 39.000 estudiantes en 3° y 5°, y 53.000 en 7° y 9°.

4.2. Acciones y Actitudes Ciudadanas (AyA) en Saber 5°, 7° y 9°

La prueba de AyA hace parte del componente de competencias ciudadanas junto con la prueba de pensamiento ciudadano. Cabe resaltar que el Icfes (2022) define a las competencias ciudadanas como

“...aquellas capacidades cognitivas, emocionales y comunicativas que, integradas entre sí y relacionadas con conocimientos y actitudes, posibilitan que el ciudadano actúe de manera constructiva en la sociedad”. (p. 64).

En particular, la prueba de AyA es un instrumento tipo encuesta compuesto por ítems con escalas Likert de dos tipos: para las Acciones, una escala de Frecuencia con cuatro opciones de respuesta: “Nunca”, “Casi nunca”, “Casi siempre” y “Siempre”; y para las Actitudes, una escala de Acuerdo con cuatro opciones de respuesta: “Muy en desacuerdo”, “Algo en desacuerdo”, “Algo de acuerdo” y “Muy de acuerdo”. Tanto las Acciones como las Actitudes incluyen ítems de polaridad directa e inversa, según la relación positiva o negativa de las categorías de respuesta con el atributo evaluado.

Las especificaciones de esta prueba se dan en conformidad con el “Modelo Centrado en la Evidencia” (Mislevy y Haertel, 2006), por lo tanto, en la Tabla 1 se encuentran las afirmaciones evaluadas. Para 5° se evalúan las afirmaciones 1 a la 4, mientras que para los grados 7° y 9° se evalúan las 6 afirmaciones descritas en dicha tabla.

Tabla 1: Afirmaciones evaluadas por la prueba de Acciones y Actitudes para su aplicación 2023 en el estudio Saber 3°, 5°, 7° y 9°

Tipo	Afirmación	Índice
Acciones	1	Cuida las comunidades a las que pertenece a través de sus acciones cotidianas.
	2	Participa en la construcción colectiva de los asuntos públicos en las comunidades a las que pertenece.
	3	Valora la importancia de entender, analizar críticamente y reevaluar su postura hacia los asuntos de interés ciudadano.
Actitudes	4	Valora los derechos humanos.
	5	Valora las responsabilidades inherentes a la pertenencia a comunidades y la importancia de las instituciones.
	6	Valora la posibilidad de incidir en la comunidad como motivación de las acciones ciudadanas.

Teniendo en cuenta los lineamientos del Icfes (2020), para el armado, o definición de formas finales a aplicar en cuanto a la selección de ítems y su orden de presentación, se utilizaron diferentes diseños para las 4 respectivas formas utilizadas en las aplicaciones de 2022 y 2023. Para el primer año fue escogido un diseño de bloques incompletos balanceados con bloques fijos para todas las formas. En contraste, para el 2023 se utilizó un diseño de bloques completos. Lo anterior hace referencia a que mientras para 2022 no hubo evaluados que respondieran a la totalidad de los ítems, para 2023 si se cumple tal condición, y en específico, toda la población examinada presentó diferentes permutaciones del mismo *pool* de ítems.

5 Resultados

La metodología de análisis de DIF para ítems politómicos descrita previamente se aplicó a la prueba de AyA entre las aplicaciones de 2022 y 2023. Estos análisis se realizan con el fin de garantizar la comparabilidad de la escala en el tiempo, mediante la identificación de ítems con DIF y, cuando corresponda, la recalibración de sus parámetros en la aplicación de 2023. En este análisis, el grupo focal corresponde a la población evaluada en la aplicación de 2022, mientras que el grupo de referencia corresponde a la población de 2023.

En la Tabla 2 se presenta para cada grado, la cantidad de estudiantes que presentaron la prueba, el número de ítems que conformaron la prueba en su versión diagramada y la cantidad de ítems para los cuales se hace el análisis. La diferencia entre la cantidad de ítems de

la prueba y los de análisis se debe a ítems que tuvieron algún ajuste en su contenido y no son comparables entre las dos aplicaciones. En esta se observa una cantidad similar de estudiantes para 5° y 7° entre aplicaciones 2022 y 2023, pero un aumento de alrededor de 5.000 estudiantes en 9°. Respecto a la cantidad de ítems, se observan ajustes sobre todo en 7° con 9 ítems no comparables entre aplicaciones.

Tabla 2: Características de la prueba AyA

Grado	Población		No. de ítems	
	2022 (Foc)	2023 (Ref)	Prueba	Análisis
5°	15.597	15.411	50	44
7°	20.977	20.321	60	51
9°	13.238	18.231	60	55

En la Tabla 3 se presentan los resultados de la equiparación y purificación en los análisis de DIF, las últimas constantes de equiparación empleadas y la clasificación de la magnitud de DIF de los ítems: A - Despreciable, B - Moderado, C - Grande. En esta se observa que en 1 o 2 iteraciones se termina la purificación. Respecto a las constantes de equiparación k_A , estas son valores muy cercanos a 1 lo cual indica una dispersión similar entre aplicaciones, respecto a las constantes k_B , éstas son negativas y muestran una mayor dificultad de las pruebas en 2023, en particular para 9°.

Respecto a los ítems con DIF, se observa que la mayoría presenta un tamaño del efecto despreciable (categoría A), lo que indica un funcionamiento equivalente entre las poblaciones en un alto porcentaje de los ítems. Este resultado sugiere que el instrumento, en términos generales, no presenta variaciones significativas entre las dos aplicaciones y conserva sus propiedades métricas, lo que respalda la comparabilidad de las mediciones. En menor proporción se identifican ítems con un tamaño del efecto moderado (categoría B). Finalmente, pocos ítems presentan un tamaño del efecto grande (categoría C): 3 en 5°, 4 en 7° y 8 en 9°. En conjunto, estos resultados indican que el DIF identificado afecta a una proporción reducida de los ítems y que, en la mayoría de los casos, las diferencias observadas entre las poblaciones son de magnitud despreciable.

Tabla 3: Resumen de los resultados del análisis de DIF por prueba.

Grado	Iteraciones Purificación	Constantes de Equiparación		No. de ítems TE		
		kA	kB	A	B	C
5°	2	1.03	-0.03	30	11	3
7°	1	1.00	-0.16	30	17	4
9°	2	1.00	-0.23	34	13	8

En la Tabla 4 se presenta información de los ítems que presentaron un tamaño del efecto de DIF grande. Se observa que varios de los ítems pertenecen principalmente a la Afirmación 1, Cuida las comunidades a las que pertenece a través de sus acciones cotidianas; y a la Afirmación 4, Valora los derechos humanos. Además, para 9° se observa una mayor cantidad de ítems con DIF en Actitudes.

Al analizar los ítems se observa que la mayoría presentan evidencia de DIF uniforme, lo que indica que las diferencias entre aplicaciones se manifiestan principalmente como desplazamientos sistemáticos en el funcionamiento de los ítems. En contraste, un subconjunto reducido presenta DIF no uniforme asociado a cambios en el parámetro de discriminación. Aunque estos últimos son menos frecuentes, resultan de especial interés, ya que en ítems politómicos tipo Likert el nivel de discriminación ha sido asociado con el denominado Estilo de Respuesta Extremo (*Extreme Response Style*) (Hibbing, Cawvey, Deol, Bloeser, y Mondak, 2019). Una mayor discriminación implica una transición más marcada entre las categorías de respuesta, favoreciendo una mayor probabilidad de seleccionar las categorías extremas (“Muy de acuerdo”/“Muy en desacuerdo”, “Siempre”/“Nunca”). En este sentido, los cambios observados podrían reflejar variaciones en la forma en que los evaluados utilizan la escala de respuesta, más allá de cambios en el constructo medido.

Es notable que estos cambios en la discriminación se concentran principalmente en las competencias de actitudes y en ítems con esquemas de polaridad inversa. Al revisar los resultados descriptivos también se identifica una variación en la proporción de escogencia de la categoría de mayor valor, tanto en las escalas de acuerdo (“Muy de acuerdo”) como en las escalas de frecuencia (“Siempre”). En la mayoría de los casos, estos resultados sugieren que las diferencias observadas podrían estar asociadas con una disminución del estilo de respuesta extremo, especialmente en los niveles cercanos al último umbral del ítem. Este patrón constituye un hallazgo de interés para futuras revisiones de contenido y análisis cualitativos,

Tabla 4: Ítems con un tamaño del efecto de DIF grande (C).

Grado	Ítem	Competencia	Afirmación	Esquema	NCDIF	Tipo DIF	Comportamiento 2023
5°	Ítem_05	Acciones	1	Frecuencia directa	0.08171	Uniforme	Menor Frecuencia
	Ítem_10	Acciones	2	Frecuencia directa	0.15663	Uniforme	Menor Frecuencia
	Ítem_40	Actitudes	4	Acuerdo directa	0.12717	Uniforme	Menor Acuerdo
7°	Ítem_02	Acciones	1	Frecuencia directa	0.10149	Uniforme	Mayor Frecuencia
	Ítem_18	Acciones	1	Frecuencia inversa	0.14183	No Uniforme	Reducción categoría superior
	Ítem_42	Actitudes	5	Acuerdo inversa	0.09868	No Uniforme	Reducción categoría superior
	Ítem_56	Actitudes	6	Acuerdo inversa	0.16597	No Uniforme	Reducción categoría superior
9°	Ítem_02	Acciones	1	Frecuencia directa	0.07270	Uniforme	Mayor Frecuencia
	Ítem_12	Acciones	2	Frecuencia directa	0.07680	Uniforme	Mayor Frecuencia
	Ítem_15	Acciones	2	Frecuencia directa	0.22420	Uniforme	Menor Frecuencia
	Ítem_48	Actitudes	4	Acuerdo inversa	0.08033	Uniforme	Menor Desacuerdo
	Ítem_49	Actitudes	4	Acuerdo directa	0.85052	Uniforme	Menor Acuerdo
	Ítem_52	Actitudes	4	Acuerdo inversa	0.16206	No Uniforme	Reducción categoría superior
	Ítem_53	Actitudes	4	Acuerdo inversa	0.14216	No Uniforme	Reducción categoría superior
	Ítem_58	Actitudes	5	Acuerdo Inversa	0.15781	No Uniforme	Reducción categoría superior

con el fin de establecer si dichas diferencias obedecen a cambios en la población evaluada o a características propias de los ítems.

6 Conclusiones

Los resultados obtenidos muestran que la metodología basada en el índice NCDIF es una alternativa adecuada para la detección y clasificación del tamaño del DIF en ítems politómicos ordinales. Su implementación es consistente con el uso de modelos de TRI para la estimación de los rasgos latentes y proporciona un marco metodológico para la identificación y evaluación del DIF mediante procedimientos de prueba de hipótesis y medidas de tamaño del efecto.

Adicionalmente, la naturaleza paramétrica de la metodología, junto con los procedimientos de equiparación, prueba de hipótesis y estimación del tamaño del efecto, facilita su incorporación en procesos estandarizados de análisis psicométrico, como los desarrollados en las pruebas Saber. Estas características amplían las posibilidades de aplicación del procedimiento, permitiendo realizar estudios de funcionamiento diferencial tanto entre subpoblaciones como entre diferentes aplicaciones de una misma prueba, favoreciendo el seguimiento del comportamiento de los ítems a lo largo del tiempo.

Respecto a los resultados obtenidos para la prueba de Acciones y Actitudes entre las aplicaciones de 2022 y 2023, se identificó un número reducido de ítems con evidencia de funcionamiento diferencial de tamaño grande. En términos generales, estos resultados respaldan la comparabilidad de los puntajes de la prueba entre aplicaciones, al tiempo que evidencian algunos cambios en el comportamiento de determinados ítems, posiblemente asociados a la naturaleza de los constructos evaluados y a las variaciones propias de la población entre aplicaciones. La identificación de estos casos proporciona información valiosa para orientar la revisión cualitativa de los ítems y apoyar las decisiones relacionadas con el mantenimiento y actualización del banco de preguntas.

Finalmente, uno de los principales aportes de este trabajo es la propuesta de puntos de corte para clasificar el tamaño del efecto del índice NCDIF en ítems politómicos, desarrollada por la Subdirección de Estadística a partir de las características de los datos y las necesidades específicas de las pruebas analizadas por el Icfes. Los resultados muestran que estos valores constituyen una aproximación adecuada para evaluar las diferencias entre las funciones de puntaje esperado de los grupos comparados. No obstante, es conveniente continuar evaluando su desempeño mediante estudios de simulación que consideren diferentes tamaños muestrales, distribuciones de habilidad y configuraciones de parámetros, con el propósito de consolidar su validez y ampliar su aplicabilidad en distintos contextos de evaluación.

Agradecimientos

Los autores agradecen a la Subdirección de Estadísticas (SE) del Icfes por facilitar el tiempo destinado a la investigación y el acceso a los datos de Saber 3°, 5°, 7° y 9°. Este proyecto se desarrolló en el marco de las actividades de la SE, que, en cumplimiento de su misión institucional, desarrolla proyectos de investigación sobre métodos psicométricos y estadísticos. Los autores agradecen especialmente a Cristian Fabián Montaña por sus valiosos comentarios en la revisión de este documento y la supervisión de este trabajo.

Referencias

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67–91.
- American Educational Research Association, American Psychological Association, y National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Angoff, W. H., y Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 10, 95–106.
- Birnbaum, A., Lord, F. M., y Novick, M. R. (1968). Statistical theories of mental test scores. *Some latent trait models and their use in inferring an examinee's ability*. Addison-Wesley, Reading, MA.
- Candell, G. L., y Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, 12(3), 253–260.
- Chang, H. H., Mazzeo, J., y Roussos, L. (1996). SIBTEST for detecting differential item functioning in polytomously scored items. *Journal of Educational Measurement*, 33, 333–353.
- Cohen, A. S., Kim, S. H., y Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, 17, 335–350.
- Dorans, N. J., y Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355–368.
- Engels, O., Lüdtke, O., y Robitzsch, A. (2026). Estimating trends with differential item functioning: A comparison of five IRT-based approaches. *Educational and Psychological Measurement*.
- French, A. W., y Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, 33, 315–332.
- Goldstein, H. (1983). Measuring changes in educational attainment over time: problems and possibilities. *Journal of Educational Measurement*, 20, 369–377.
- Gómez-Benito, J., Hidalgo, M. D., y Guilera, G. (2010). El sesgo de los instrumentos de medición. tests justos. *Papeles del psicólogo*, 31(1), 75–84.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 144–149.
- Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, 44, S182–S188.
- Hibbing, M. V., Cawvey, M., Deol, R., Bloeser, A. J., y Mondak, J. J. (2019, 03). The relationship between personality and response patterns on public opinion surveys: The big five, extreme response style, and acquiescence response style. *International Journal of Public Opinion Research*, 31(1), 161–177. Descargado de <https://doi.org/10.1093/ijpor/edx005>
- Holland, P. W., y Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer y H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Holland, P. W., y Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Icfes. (2020). *Diseño de armado para pruebas estandarizadas: usos y metodologías*. Autor.
- Icfes. (2022). *Informe nacional de resultados. saber 3°, 5°, 7° y 9°. aplicación 2022*. Autor.
- Kolen, M. J., y Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York, NY: Springer.

- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Magis, D., Béland, S., Tuerlinckx, F., y De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, 42(3), 847–862.
- Mantel, N., y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Mislevy, R. J., y Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational measurement: issues and practice*, 25(4), 6–20.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Oshima, T. C., y Morris, S. B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, 27(3), 43–50.
- Oshima, T. C., Raju, N. S., y Nanda, A. O. (2006). A new method for assessing the statistical significance in the Differential Functioning of Items and Tests (DFIT) framework. *Journal of Educational Measurement*, 43(1), 1–17.
- Penfield, R. D., y Camilli, G. (2007). Differential item functioning and item bias. En C. R. Rao y S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. psychometrics* (pp. 125–167). Amsterdam: Elsevier.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495–502.
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14, 197–207.
- Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., y Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement*, 33, 133–147.
- Raju, N. S., van der Linden, W. J., y Fleer, P. F. (1995). IRT-Based Internal Measures of Differential Functioning of Items and Tests. *Applied Psychological Measurement*, 19(4), 353–368.
- Samejima, F. (1969, marzo). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(1), 1–97.
- Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, 16, 143–152.
- Shealy, R., y Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences. *Psychometrika*, 58, 159–194.
- Sireci, S. G., y Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, 19(2–3), 170–187.
- Stocking, M. L., y Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- Swaminathan, H., y Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.
- Thissen, D., Steinberg, L., y Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. En H. Wainer y H. Braun (Eds.), *Test validity* (pp. 147–170). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wells, C. S., Subkoviak, M. J., y Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26, 77–87.
- Wright, K. D., y Oshima, T. C. (2015). An effect size measure for Raju's differential functioning for items and tests. *Educational and Psychological Measurement*, 75(2), 338–358.