

Bogotá D.C
Agosto 2024
ISSN: 2590 - 4663
Publicación Trimestral



Educación



Icfes

SABER AL DETALLE

EDICIÓN 15



**¿Qué es el Funcionamiento
Diferencial de los Ítems y cómo
se detecta en las pruebas que
aplica el Icfes?**



Presidente de la República
Gustavo Francisco Petro Urrego

Ministro de Educación Nacional
Jose Daniel Rojas Medellín

Elaboración del documento
John Alexander Calderón Rodríguez
Andrés Ricardo Rodríguez Nagles
Mishell Marcela Ramos de la Hoz
Nila Fernanda Amaya Melo

Revisado por
Hanwen Zhang
Nelson Andrés Rodríguez Rivera
Juan Carlos Rubriche Cárdenas

Diseño y diagramación
Andrea del Pilar López Pulido

Bogotá D.C., Agosto 2024

Todos los derechos de autor reservados ©.

Directora General
Elizabeth Blandón Bermúdez

Secretaria General
Blanca Irene Echavarría Lotero

Director de Evaluación
Rafael Eduardo Benjumea Hoyos

Subdirector de Estadísticas
Cristian Fabian Montaña Rincón

Subdirectora de Análisis y Divulgación
Alejandra Neira Aroca

Directora de Producción y Operaciones
Luz Patricia Loaiza Cruz

Subdirector de Producción de Instrumentos
Gustavo Andrés Monsalve Londoño

Director de Tecnología e Información
Luis Rodrigo Cadavid Durán

Subdirector de información
Diego Mauricio Salas Ramírez

Jefe Oficina Asesora de Comunicaciones y Mercadeo
Byron Andrés Vélez Valdés

Jefa Oficina Asesora de Gestión de Proyectos de Investigación
Jennyfer Paola Guio Veloza



Términos y condiciones de uso para las publicaciones y obras que son propiedad del Icfes

El Instituto Colombiano para la Evaluación de la Educación (Icfes) pone a disposición de la comunidad educativa, y del público en general, de forma gratuita y libre de cualquier cargo, un conjunto de publicaciones disponibles en su portal web www.icfes.gov.co. Estos materiales y documentos están normados por la presente política y se encuentran protegidos por derechos de propiedad intelectual y derechos de autor a favor del Icfes. Si tiene conocimiento de alguna utilización contraria a lo establecido en estas condiciones de uso, por favor infórmenos al correo prensaicfes@icfes.gov.co.

Queda prohibido el uso o publicación total o parcial de este material con fines de lucro. Únicamente está autorizado su uso para fines académicos e investigativos. Ninguna persona natural o jurídica, nacional o internacional, podrá vender, distribuir, alquilar reproducir, transformar¹, promocionar o realizar acción alguna con la cual se lucre directa o indirectamente con este material. Esta publicación cuenta con el registro ISBN (International Standard Book Number, o Número Normalizado Internacional para Libros), que facilita la identificación, no sólo de cada título, sino, también de la autoría, la edición, el editor y el país dónde se edita.

En todo caso cuando se haga uso parcial o total de los contenidos de esta publicación, el usuario deberá consignar o hacer referencia a los créditos institucionales del Icfes, respetando los derechos de cita.

En otras palabras, se podrá hacer uso de esta publicación si dicho uso se contempla en los fines aquí previstos. Es posible, entonces, transcribir pasajes del texto si se cita siempre la fuente de autor. Por supuesto, estas citas no deberían ser excesivas ni frecuentes para que, así, no se considere una reproducción simulada y sustancial que redunde en perjuicio del Icfes.

Asimismo, los logotipos institucionales son marcas registradas y de propiedad exclusiva del Icfes. Por tanto, cuando su uso pueda causar confusión, los terceros no podrán usar las marcas de propiedad del Icfes con signos idénticos o similares respecto a cualquier producto o servicio prestado por esta entidad. En todo caso queda prohibido su uso sin previa autorización expresa por parte del Icfes. La infracción de estos derechos se perseguirá civil y penalmente (en caso de que sea necesario) de acuerdo con las leyes nacionales y tratados internacionales aplicables.

El Icfes realizará cambios o revisiones periódicas a los presentes términos de uso y los actualizará en esta publicación.

¹ La transformación es la modificación de la obra a través de la creación de adaptaciones, traducciones, compilaciones, actualizaciones, revisiones y, en general, cualquier modificación que se pueda realizar, haciendo que la nueva obra resultante se constituya en una obra derivada protegida por el derecho de autor, con la única diferencia, respecto de las obras originales, que aquellas requieren, para su realización, de la autorización expresa del autor o propietario, para adaptar, traducir, compilar, etc. En este caso, el Icfes prohíbe la transformación de esta publicación. Términos y condiciones de uso para las publicaciones y obras que son propiedad del Icfes.

¿Qué es el Funcionamiento Diferencial de los Ítems y cómo se detecta en las pruebas que aplica el Icfes?

¿Qué es DIF?

En el análisis del comportamiento de los ítems que componen una prueba, se entiende que un ítem presenta Funcionamiento Diferencial del Ítem (DIF, por sus siglas en inglés) cuando evaluados con el mismo nivel de habilidad tienen una probabilidad distinta de responderlo correctamente por el hecho de pertenecer a un grupo determinado, por ejemplo, ser evaluados en diferentes periodos de aplicación de un examen o pertenecer a una subpoblación determinada por alguna característica como el género o la zona de residencia (rural o urbana), entre otros.

El análisis de DIF en una prueba es importante ya que este comportamiento de los ítems podría tener implicaciones en la medición. Por una parte, el análisis de DIF permite identificar violaciones a los supuestos de invarianza en la prueba, lo que implica que los ítems no estarían midiendo a las personas de la misma forma. Otra razón está asociada con la validez¹ de los instrumentos en relación con la comparabilidad de los resultados en el tiempo y entre grupos de evaluados.

Por lo tanto, la presencia de ítems con DIF en una prueba puede afectar su imparcialidad, ya que se puede desfavorecer a algunos individuos debido a razones ajenas al objeto de la medición. Este aspecto es especialmente relevante cuando se toman decisiones a partir de los resultados del examen como el ingreso a la educación superior. Además, impacta

sobre la confiabilidad de la prueba ya que afecta la precisión de la medición al introducir una fuente de variación indeseada; asimismo, la validez se ve afectada ya que las interpretaciones que se realicen sobre los resultados pueden estar sesgadas bajo las diferencias detectadas.

Algunas posibles causas del DIF se asocian con el uso de un lenguaje complicado para algunas personas, las escalas u opciones de respuesta que se utilizan y el uso de situaciones o elementos en los ítems que son muy específicas de un contexto cultural (Hambleton, 2006). En este sentido, los análisis cuantitativos de DIF brindan información precisa respecto a las diferencias presentadas por el ítem que puede ser complementada con análisis cualitativos para profundizar sobre las causas de estas diferencias.

Por lo anterior, y teniendo en cuenta que la presencia de DIF impacta la calidad de las mediciones, el Icfes implementa procesos estrictos de análisis de DIF en las pruebas que aplica con el fin de garantizar su confiabilidad y validez para generar mediciones más precisas y justas para los evaluados. En este documento se presentan algunas precisiones conceptuales, se hace un breve repaso histórico y se profundiza en una de las metodologías implementadas por el Icfes para el análisis de DIF en las pruebas.

¹ De acuerdo con los estándares de la AERA, la APA y NCME (2018), la validez se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de la prueba para los usos propuestos.

¿Cuáles son las diferencias entre DIF, sesgo e impacto?

Es posible que, en algunos casos, un ítem muestre un comportamiento distinto para dos grupos de evaluados cuando se presentan otras variables irrelevantes para el constructo que interfieran en la medición, esto se denomina **sesgo**. Por ejemplo, el sesgo se puede presentar cuando se incluyen elementos particulares del entorno como los ascensores o el metro (medio de transporte), en el contexto de una pregunta de una prueba que desfavorezcan la evaluación de un grupo en particular.

En este caso, poblaciones de algunas zonas rurales pueden desconocer estos elementos porque no están en su entorno, por lo que se introducirían competencias o conocimientos adicionales a los que se pretende evaluar para responder la pregunta. Esta situación pondría en desventaja a las personas de dichas zonas frente a otras poblaciones de zonas urbanas, quienes pueden estar más familiarizados con los elementos en cuestión.

En este sentido, el análisis de **DIF** brinda información cuantitativa para detectar estadísticamente las posibles diferencias entre dos grupos al responder un ítem debido a la presencia de sesgo en la evaluación; sin embargo, no detecta las razones de fondo sobre estas diferencias por lo que se requieren estudios específicos de sesgo que permitan indagar sobre las causas. No obstante, se debe considerar que no toda diferencia observada en el desempeño de dos grupos de evaluados se debe a la presencia de sesgo en la medición. También se pueden presentar diferencias relevantes para el constructo que se desea medir, lo cual se denomina **impacto**. Por ejemplo, si se evidencian diferencias sobre un ítem que aborda conocimientos en plantas para estudiantes de poblaciones ubicadas en zonas urbanas y rurales, estas son irrelevantes al constructo en caso de que la prueba sea de lenguaje (sesgo), pero, si la prueba evalúa conocimientos en ciencias naturales, estas diferencias sí están relacionadas con el constructo a evaluar, y, por tanto, estamos bajo la presencia de impacto.

Por lo anterior, es importante distinguir si las diferencias detectadas se deben a elementos relevantes en la medición del constructo de interés (impacto) o si por el contrario suceden por variables externas a la medición (sesgo), cuyo caso sería indeseado para obtener una medición justa.

¿Cómo surgieron los estudios asociados con el DIF?

A principios del siglo XX, en el contexto de la educación y la evaluación, se identificaron diferencias en el desempeño de subpoblaciones debido a sus características culturales (Stern, 1914 y Binet, 1916, citados en Herrera et al., 2001); sin embargo, es hasta la década de los 60 en donde empieza a cobrar mayor relevancia el estudio de dichas diferencias. En Estados Unidos se presentaron diversos movimientos para la defensa de los derechos civiles y legales de algunos grupos minoritarios o algunas subpoblaciones que se veían afectadas por el uso de instrumentos que desfavorecían a estos grupos debido a la presencia de terminologías específicas desconocidas por algunos sectores sociales, por la presencia de lenguaje discriminatorio, entre otros (Moreira-Mora, 2008).

En la década de los 80, se generaron investigaciones para la detección de ítems con sesgo que pudieran generar comportamientos diferenciados entre subgrupos. De esta manera, se consolidó un marco estadístico que permitió robustecer el análisis de DIF, lo cual fue incorporado en los procesos de evaluación que se realizaban en la época, especialmente en el ámbito educativo y laboral. A partir de entonces, estos procedimientos se han integrado para contribuir a las evidencias de validez de los instrumentos de medición con el objetivo de lograr una evaluación más justa (Moreira-Mora, 2008).

En este sentido, los procedimientos para la detección de DIF se han ampliado en los últimos años teniendo en cuenta las implicaciones sociales, educativas y laborales que pueden tener los instrumentos de medición para las personas evaluadas. El uso de estos procedimientos se ha difundido para contribuir con el

desarrollo de instrumentos que cumplan con criterios que garanticen la calidad en la medición.

¿Cómo se hace un análisis de DIF?

Para la detección de DIF se han desarrollado varias metodologías estadísticas enfocadas principalmente al análisis de pruebas conformadas por ítems dicotómicos², las cuales se pueden clasificar como paramétricas o no paramétricas dependiendo de si estas hacen uso o no de los parámetros de los ítems. Algunas de las más comunes emplean el estadístico de Mantel Haenszel (MH) (Mantel & Haenszel, 1959), la regresión logística (LR) (Swaminathan and Rogers, 1990), la prueba del cociente de verosimilitudes entre modelos anidados (LRT, por sus siglas en inglés) (Bing & Babu, 2019), el SIBTEST (Shealy & Stout, 1993) y el Índice no compensatorio de DIF (NCDIF, por sus siglas en inglés) (Raju et al., 1995), entre otros. Para seleccionar una o varias metodologías que permitan realizar un análisis de DIF, se consideran algunas características como el modelo de calificación usado en la prueba, el tamaño de la población evaluada y el diseño de armado de la prueba que determina si existen grupos de estudiantes sin respuestas a un conjunto de ítems.

Teniendo en cuenta lo anterior, en el Icfes una de las metodologías empleadas es el Índice NCDIF para realizar el análisis de DIF dado que a) en las pruebas del instituto se implementan principalmente modelos de Teoría Respuesta al Ítem (TRI) para la generación de puntajes, como en el examen Saber 11 o los módulos genéricos del examen Saber PRO y TyT, b) existe un gran número de evaluados y c) el diseño de armado es BIBs³ en las pruebas. Este índice pertenece a las metodologías paramétricas en las que se usan los parámetros de los ítems para realizar el análisis de DIF.

Las metodologías paramétricas de análisis de DIF se basan en la comparación de las curvas características de los ítems (CCI), las cuales relacionan la probabilidad de contestar correctamente una pregunta y están determinadas por los parámetros de los ítems, ya que sobre estas se puede cuantificar directamente la diferencia de las probabilidades de acierto y por ende la presencia de DIF en el ítem.

Como se mencionó en el Boletín 1 de Saber al detalle⁴, la TRI asume que existe una relación entre la habilidad de un evaluado y la probabilidad de responder correctamente un ítem, y dicha relación está dada por la CCI que está determinada por el modelo logístico de tres parámetros: discriminación, dificultad y pseudo-azar. Desde el punto de vista paramétrico, un ítem presenta DIF si hay diferencias significativas entre las curvas características del ítem en los grupos de interés para los cuales se está realizando el análisis. En los estudios de DIF, al conjunto de evaluados que representa el foco de interés se llama grupo focal, mientras que el grupo respecto al cual se compara el grupo focal se denomina grupo de referencia.

¿Qué tipos de DIF existen?

El DIF se puede clasificar en dos tipos: uniforme y no uniforme. Los tipos de DIF varían según las diferencias que se presentan en los parámetros de dificultad y discriminación del ítem entre los grupos. Respecto al DIF uniforme, este se da cuando la variación entre los grupos corresponde únicamente al parámetro de dificultad y no al de discriminación.

En este caso, la magnitud de la diferencia en la probabilidad de acertar el ítem $P(\theta)$ ⁵ entre los grupos se mantiene igual o similar en todos los niveles de habilidad.

2 Los ítems dicotómicos corresponden a aquellos que solamente tienen dos posibles puntuaciones. Por ejemplo, en los ítems de selección múltiple con única respuesta se puntúan de manera dicotómica con 1 para la opción correcta y 0 para las opciones incorrectas.

3 Para ampliar información sobre el armado, puede consultar la edición 2 de Saber al Detalle, disponible en: <https://www.icfes.gov.co/documents/39286/2231027/Boletin+2+-+que+diseño+del+armado+se+emplea+en+el+icfes+para+medir+las+pruebas+saber+.pdf>

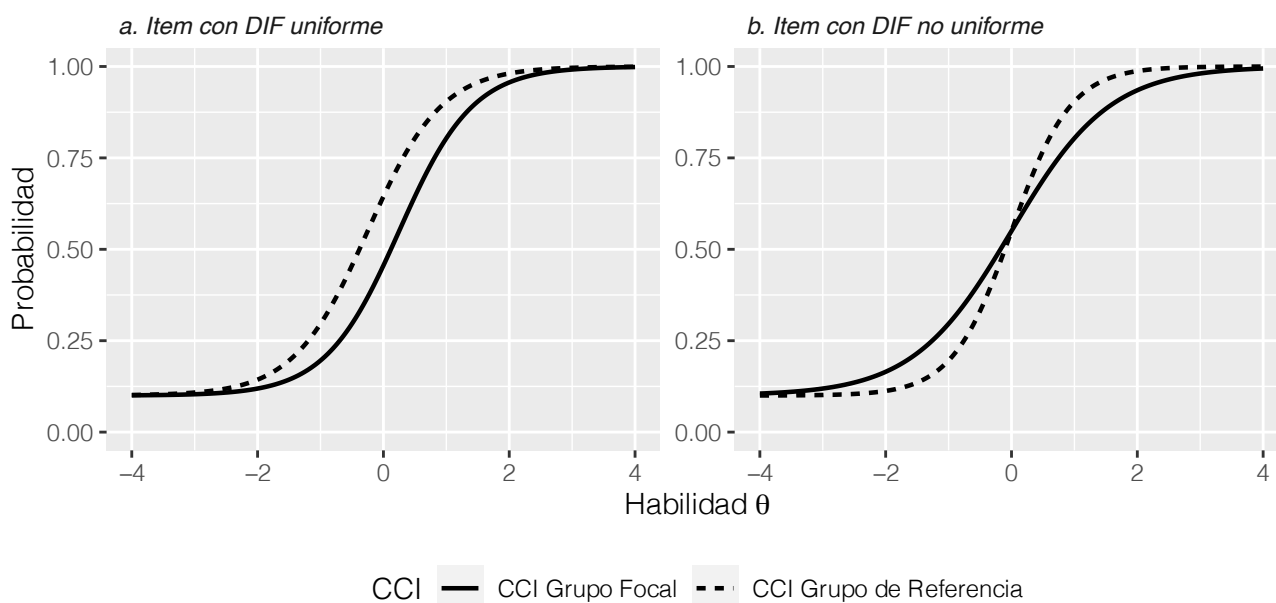
4 Para consultar este boletín, puede ingresar al siguiente enlace: <https://www.icfes.gov.co/documents/39286/2231027/Boletin+1+-+cómo+se+generan+los+puntajes+en+las+pruebas+.pdf>

5. (θ) se refiere a la habilidad de un evaluado.

Para entender este tipo de DIF, se presenta la Figura 1a. En esta gráfica se presenta la CCI para dos poblaciones diferentes, es decir, expresa la relación entre la probabilidad de responder correctamente el ítem acorde con los diferentes niveles de habilidad de cada grupo. En esta figura se puede observar

que, aunque el ítem tiene la misma discriminación y pseudo-azar para ambos grupos (discriminación = 1, pseudo-azar = 0.1), el ítem es más fácil para las personas del grupo de referencia (dificultad = -0.25) que para las personas del grupo focal (dificultad = 0.25).

Figura 1 Ejemplos de tipos de DIF



Nota: La línea continua corresponde a la curva característica del ítem para el grupo focal. En el panel a, se presenta DIF uniforme porque en la CCI del grupo de referencia solo cambia el parámetro de dificultad, lo cual genera un desplazamiento horizontal de la curva respecto a la CCI del grupo focal. En el panel b, se presenta DIF no uniforme, ya que en la CCI del grupo de referencia solo cambia la discriminación, generando intersección entre las curvas, lo cual beneficia al grupo focal en partes bajas de la habilidad y al grupo de referencia en partes altas de habilidad.

Un elemento importante que caracteriza el DIF uniforme a nivel gráfico es que, como se puede observar en la Figura 1a, por el cambio del parámetro de dificultad la CCI de un grupo queda desplazada respecto al otro; es decir, la probabilidad de acertar el ítem en el grupo de referencia es mayor que en el grupo focal, lo que implica que el ítem favorece sistemáticamente a un grupo a lo largo de todas las habilidades.

Por otra parte, el DIF no uniforme se presenta cuando la variación entre los grupos se da en la discriminación del ítem, pero no en su dificultad. Este tipo de DIF se

caracteriza porque la diferencia en la probabilidad de acertar el ítem entre los grupos varía a lo largo de los niveles de habilidad, favoreciendo o desfavoreciendo al grupo en ciertos rangos de la habilidad. En la Figura 1b se puede observar que, si bien los parámetros de dificultad y pseudo-azar del ítem para los dos grupos es la misma (dificultad = 0, pseudo-azar = 0.1), en los niveles bajos de habilidad las personas del grupo focal (discriminación = 0.75) tienen mayor probabilidad de acertar el ítem que las personas del grupo de referencia (discriminación = 1.25); mientras que en los niveles altos de habilidad se presenta la situación contraria.

Respecto al cambio del parámetro de pseudo-azar en los modelos 3PL, este usualmente genera cambios en el parámetro de discriminación y dificultad lo cual comúnmente produce la presencia de DIF no uniforme; sin embargo, ciertas configuraciones de parámetros generan un favorecimiento a un grupo a lo largo de todas las habilidades con diferentes magnitudes.

¿Cómo se detecta la presencia de DIF mediante el índice NCDIF?

El Índice no compensatorio de DIF (NCDIF, por sus siglas en inglés) es una medida estadística de la magnitud del DIF en un ítem, la cual se basa en cuantificar la diferencia entre las curvas características del ítem de ambos grupos ponderada por la población de interés (Raju et al., 1995). Como su nombre lo indica, este índice no compensa las diferencias entre ambos grupos cuando la diferencia cambia de signo ya que toma el valor esperado del cuadrado.

El NCDIF para un ítem j se define como:

$$NCDIF_j = E \left[\left(P_{jF}(\theta) - P_{jR}(\theta) \right)^2 \right]$$

En la expresión anterior,

- E es la función del valor esperado,
- P_{jF} es la probabilidad de acertar el ítem del grupo focal,
- P_{jR} es la probabilidad de acertar el ítem del grupo de referencia, y
- θ es la habilidad del evaluado.

En las pruebas Saber se emplea este índice para identificar si un ítem presenta DIF entre aplicaciones o entre subpoblaciones. Para esto se obtienen las calibraciones⁶ de los ítems, se transforman a una escala común, se calcula el NCDIF para cuantificar las diferencias del comportamiento del ítem y por último se prueba la hipótesis nula (H_0) de que el ítem no presenta DIF y se clasifica el NCDIF de acuerdo con el

tamaño del efecto para evaluar cuán significativas son las diferencias detectadas (los detalles del procedimiento se describen en la siguiente sección).

Para ilustrar este último punto, se presenta la Figura 2, en la cual se observa el valor del NCDIF para tres ítems al realizar el análisis de DIF entre dos aplicaciones, una aplicación pasada (histórica) y una más reciente.

En la Figura 2a se presenta un ítem sin comportamiento diferencial. En esta se observa que las curvas características del ítem son similares y la diferencia en la probabilidad de responder correctamente el ítem para un mismo punto de habilidad es muy reducida por lo que el valor del NCDIF es muy pequeño (0.00029) y no se puede afirmar que sean diferentes desde un punto de vista estadístico.

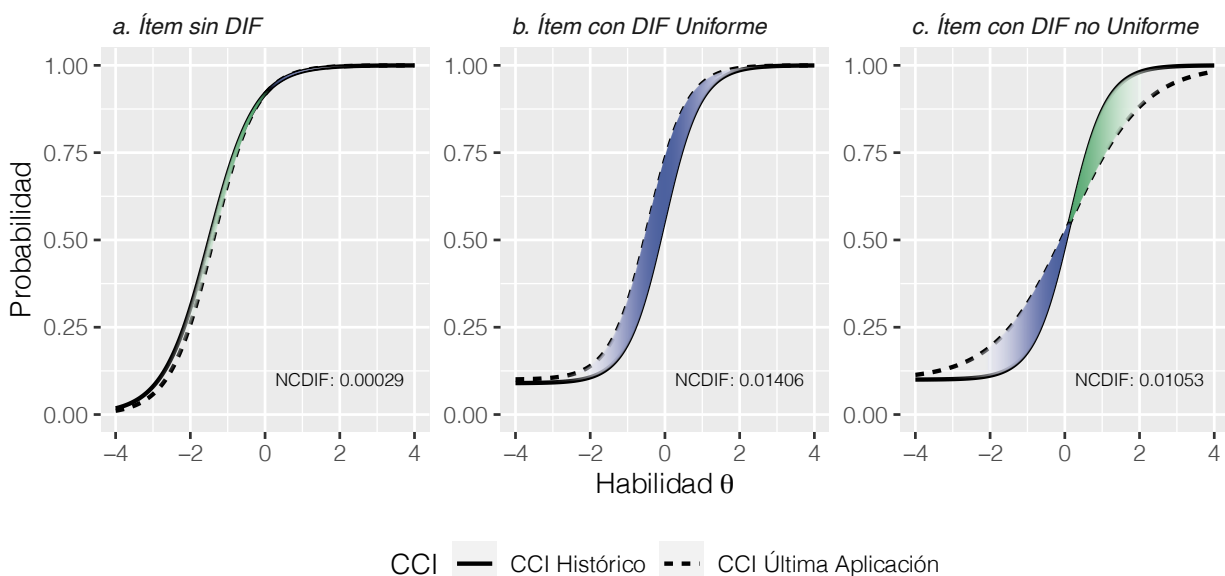
Por su parte, en la Figura 2b se presenta un ítem con un funcionamiento diferencial uniforme, con tamaño del efecto grande, y en consecuencia existe una diferencia importante entre las curvas características y un valor grande para el índice NCDIF (0.01406). Para este ítem, ignorar la presencia de DIF le da una ventaja espuria a los evaluados que presentaron el ítem en la última aplicación, lo cual tendría como consecuencia que para estos evaluados se obtendrían estimaciones de la habilidad más altas (sobrestimadas) de las reales.

De forma similar, en la Figura 2c se presenta un ítem con funcionamiento diferencial no uniforme con un valor grande en el índice NCDIF (0.01053). En este caso se evidencia que las dos curvas se cruzan y las diferencias entre estas no son constantes a lo largo de la habilidad, característico del DIF no uniforme. En la figura se observa como el ítem favorece a la población de la última aplicación en habilidades bajas mientras que en habilidades altas se desfavorece.

⁶ La obtención de las calibraciones es un proceso que se realiza para estimar los parámetros del modelo de calificación para cada ítem que conforman las pruebas, con el fin de verificar que los ítems presenten un funcionamiento psicométrico óptimo. Para más información respecto a este proceso, se recomienda consultar el Boletín 1 de Saber al detalle:

<https://www.icfes.gov.co/documents/39286/2231027/Boletin+1++como+se+generan+los+puntajes+en+las+pruebas.pdf/7ac-f8c50-e00d-70a9-f1ef-6b1a323ff520?version=1.1&t=1678150116881>

Figura 2. Análisis de DIF con el índice NCDIF

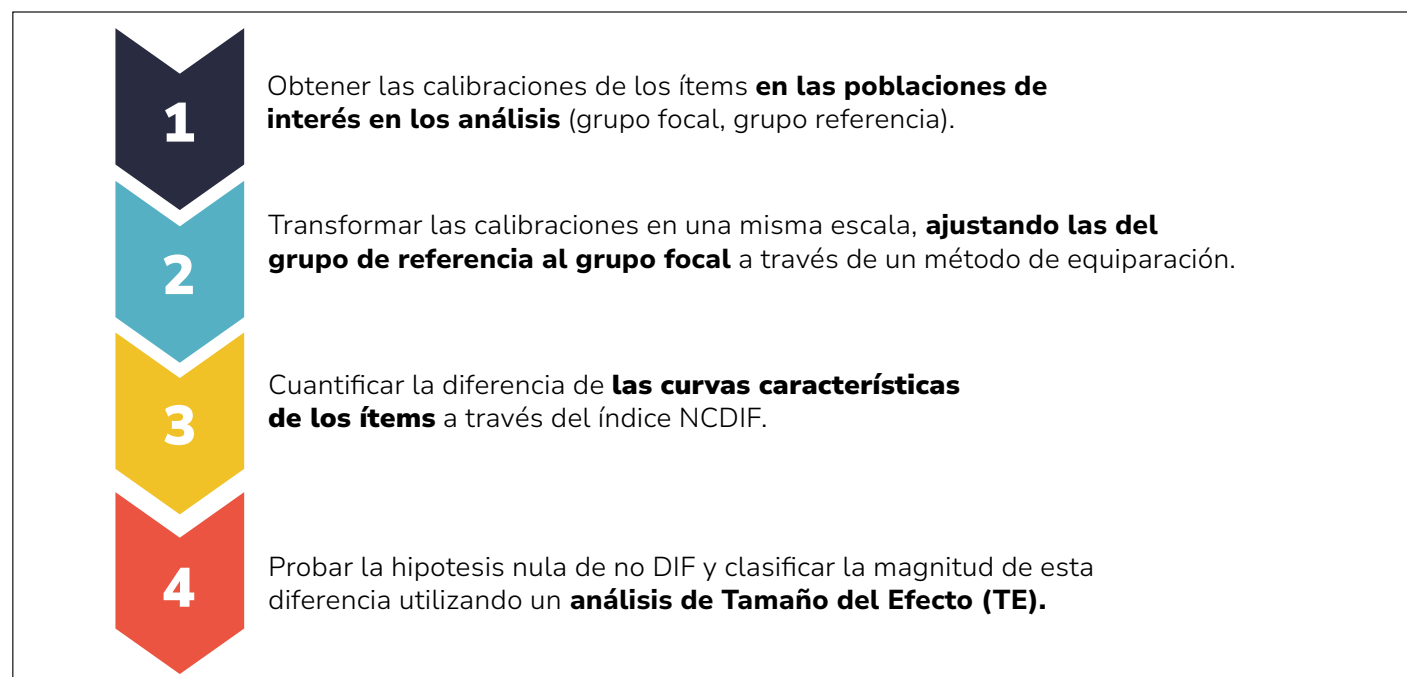


¿Cómo se aplica esta metodología en el Icfes?

En el Icfes se implementa el uso del NCDIF descrito en la sección anterior para la detección y análisis de

DIF. Este procedimiento se emplea en los exámenes para la detección de ítems con DIF entre poblaciones, por ejemplo, entre aplicaciones del examen y entre subpoblaciones de interés (género, zona, entre otros). Los pasos del procedimiento se describen en la Figura 3.

Figura 3. Procedimiento utilizado en el Icfes para el análisis de DIF utilizando el NcdIF.



Inicialmente, se obtienen las calibraciones de los ítems empleando el modelo de TRI utilizado en la calificación de la prueba para las poblaciones de interés en el análisis, esto es, obtener las calibraciones de los ítems de forma independiente (libre) ajustando el modelo sobre los evaluados del grupo focal y del grupo de referencia. En el caso del Icfes, para realizar el análisis de DIF entre aplicaciones se suele tomar las calibraciones históricas como grupo focal y las calibraciones de la aplicación más reciente como grupo de referencia.

Una vez obtenidas las calibraciones del grupo focal y referencia de forma independiente, estas ajustan a una misma escala a través de un método de equiparación. Específicamente, se transforman las calibraciones del grupo de referencia a la escala de las calibraciones del grupo focal a través del método de equiparación de Stocking-Lord (Stocking & Lord, 1983). En este, se realiza una transformación lineal que permite construir una escala común minimizando las diferencias entre las curvas características de los ítems (Kolen y Brennan, 2014).

Cuando las calibraciones se encuentren en una misma escala, se cuantifica la diferencia de las curvas características de los ítems a través del índice NCDIF, el cual, como se mencionó antes, consiste en la estimación del valor esperado de las diferencias cuadráticas entre las curvas características ponderado por la distribución del grupo focal.

Para probar la hipótesis de DIF se realiza una estimación cuantitativa de la distribución del estadístico NCDIF mediante una metodología replicación de parámetros (IPR, por sus siglas en inglés). En esta inicialmente se generan 1.000 conjuntos de posibles valores de los parámetros de los ítems para el grupo focal a partir de la distribución teórica de estos. Después, se calcula el índice NCDIF de cada conjunto de parámetros generados contra los parámetros originales del grupo focal. Sobre el conjunto de valores del NCDIF obtenidos se calculan puntos de corte al 95 % y 99 % superior, los cuales

finalmente son comparados con el valor del NCDIF calculado entre el grupo focal y de referencia para probar la hipótesis de presencia de DIF. Teniendo en cuenta la alta sensibilidad de la prueba de hipótesis por replicación de parámetros debida a los grandes tamaños poblacionales, se realiza la clasificación de la magnitud de esta diferencia a través de un análisis de Tamaño del Efecto. En este se clasifica la magnitud de DIF dada por el estadístico NCDIF en despreciable (A), moderado (B) o grande (C), siguiendo los puntos de corte establecidos por Wright y Oshima (2015) y guardando relación con la clasificación propuesta por el ETS para el DIF (Holland & Thayer, 1985, 1988).

¿Qué implicaciones tiene el análisis de DIF en las pruebas que aplica el Instituto?

Una vez se realiza el análisis de DIF se socializan los resultados con los especialistas en cada una de las pruebas. Con los ítems identificados con DIF se profundiza y analiza mediante un análisis cualitativo las posibles causas y consecuencias detrás de este funcionamiento con el fin de retroalimentar el proceso de construcción de ítems y minimizar este comportamiento en las pruebas.

Adicionalmente, a los ítems que presentan DIF entre aplicaciones se les realiza un proceso de recalibración estadística utilizando la población de la última aplicación para la actualización de sus parámetros. A pesar de que la cantidad de ítems que se identifican con DIF es mínima en cada aplicación de los exámenes, el anterior procedimiento garantiza que los ítems se mantengan en la evaluación de forma adecuada.

Por último, con el procedimiento implementado en el Icfes para el análisis de DIF y las acciones que se toman frente a estos ítems, se proporcionan evidencias de validez de las pruebas que aplica el Instituto que contribuyen a generar una medición de calidad y justa para los evaluados.

Referencias

- American Educational Research Association -AERA, American Psychological Association - APA, & National Council on Measurement in Education –NCME (2018). *Estándares para pruebas educativas y psicológicas*. American Educational Research Association.
- Bing L., Babu, G. (2019). A Graduate Course on Statistical Inference. Springer. p. 331. ISBN 978-1-4939-9759-6.
- Hambleton, R. K. (2006). Good Practices for Identifying Differential Item Functioning. *Medical Care*, 44(11), S182–S188. <http://www.jstor.org/stable/41219518>
- Herrera, A., Sánchez, R. y Gómez, J. (2001). Funcionamiento diferencial de los ítems: Una revisión conceptual y metodológica. *Acta colombiana de psicología*, 5(1) 41-61.
- Holland, P. W. and Thayer, D. T. (1985). An alternative definition of the ETS delta scale of item difficulty. Research Report RR-85-43. Princeton, NJ: Educational Testing Service.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.
- Kolen, M. J. y Brennan, R. L. (2014). *Statistics for social and behavioral sciences. Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer Science + Business Media.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Moreira-Mora, T. (2008). El funcionamiento diferencial del ítem: Un asunto de validez y equidad. *Avances en medición*, 6 (1), 5-14.
- Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19, 353–368. doi:10.1177/014662169501900405
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika*, 58(2), 159–194.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- Swaminathan, H. and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370. doi: 10.1111/j.1745-3984.1990.tb00754.x
- Wright, K. & Oshima, T. (2015). An Effect Size Measure for Raju's Differential Functioning for Items and Tests. *Educational and Psychological Measurement*. 75(2): 338–358. DOI 10.1177/0013164414532944.



Instituto Colombiano para la Evaluación de la Educación, Icfes
Oficinas Calle 26 #69 - 76. Torre 2, pisos 16, 17, 18
Edificio Elemento, Bogotá. Colombia

 **@ICFEScol**

 **ICFES**

 **icfescol**

 **ICFES**