



UNIVERSIDAD NACIONAL DE COLOMBIA

**SESGO CULTURAL EN LOS ÍTEMS DE LAS PRUEBAS DEL
EXAMEN SABER 11° EN COLOMBIA**

Martha Ligia Cuevas Mendoza

Universidad Nacional de Colombia

Facultad de Ciencias Humanas

Departamento de Psicología

Maestría en Psicología

2013

**SESGO CULTURAL EN LOS ÍTEMS DE LAS PRUEBAS DEL
EXAMEN SABER 11° EN COLOMBIA**

Martha Ligia Cuevas Mendoza

Tesis para optar al título de Magíster en PSICOLOGÍA

Dirigido por Aura Nidia Herrera Rojas (Ph.D.)

Línea de Investigación:

Métodos e instrumentos para la investigación en ciencias del comportamiento

Universidad Nacional de Colombia

Facultad de Ciencias Humanas

Departamento de Psicología

Maestría en Psicología

2013

A la profesora Aura Nidia Herrera Rojas por todas sus enseñanzas, su apoyo, comprensión y paciencia antes y durante la Maestría.

A mi familia, en especial a mi hermana Carolina Cuevas por escucharme.

A mis amigos Sandra Camargo, Ana Cristina Santana, Viviana Vargas, Nubia López, Catheryne Lancheros, Ángela Berrío, Erika Arias, Víctor Cervantes, Álvaro Uzaheta, Carolina Lopera y Jenyfer García por todo su apoyo para que culminara este proceso.

AGRADECIMIENTOS

A la Universidad Nacional de Colombia por el apoyo académico y económico otorgado para el desarrollo de la Maestría.

A Jorge Iván Ardila por el apoyo en el análisis cualitativo.

A Sandra Camargo por su colaboración en los grupos focales.

A Ángela Berrío y Erika Arias por el desarrollo de sus tesis, las cuales fueron un insumo importante en este trabajo.

Al Instituto Colombiano para la Evaluación de la Educación (ICFES) por haber facilitado las bases de datos y por el apoyo en la realización de los grupos focales. A Julián Mariño -Director de Evaluación-, Patricia Pedraza -Subdirectora de Diseño de Instrumentos- y Claudia Sáenz -Subdirectora de Producción de Instrumentos- por su apoyo en el contacto de los participantes en los grupos focales y su disposición en el desarrollo de los mismos.

A los participantes en los grupos focales por su interés y contribuciones al presente trabajo:

Cindy Acero	Édgar Antonio López
Camilo Correa	Anyela Paola Malagón
Ismael Corredor	Roberto Palomino
Orlando Díaz	Yamile Pérez
Carolina Esquivel	Magali Pinilla
Gloria Fajardo	Carolina Rojas
David García	Javier Toro
Yesid González	Fernando Torres

A María Dolores Hidalgo, jurado de tesis, por sus enseñanzas en el desarrollo de mi pasantía.

A Carlos pardo y Olga Rodríguez, jurados de tesis, por sus observaciones que ayudaron a mejorar el presente documento.

RESUMEN

El presente estudio se encarga de la identificación de posibles fuentes de sesgo cultural en los ítems de las pruebas que componen el examen SABER 11°, tomando como grupo focal a los estudiantes indígenas. En una primera fase se detectaron los ítems con DIF en siete pruebas aplicadas en el segundo semestre de 2006 y en el primero de 2007, a través de los procedimientos Mantel-Haenszel y Diferencia de la dificultad. Posteriormente, ítems de las pruebas de Lenguaje, Matemáticas, Biología y Sociales que habían sido detectados fueron revisados en grupos focales. Los resultados muestran que los procedimientos usados para la detección de DIF son adecuados para su aplicación a datos reales y bajo las condiciones del examen SABER 11°. Por otra parte, el análisis substantivo de los ítems, sugirió tres potenciales fuentes de sesgo cultural que pueden ser transversales a las pruebas: experiencias más frecuentes en un grupo que en otro, epistemología de las comunidades de origen y problemas de construcción, siendo esta última relativamente nueva frente a lo reportado en la literatura. Adicionalmente, se encontraron otras cuatro posibles fuentes emergentes: tecnicismos, colonialismo religión-Estado, juicios de valor hacia teorías sociales o de mercado y escuela tradicional. Todas estas posibles fuentes sirvieron de base para la elaboración de unas primeras pautas para evitar el sesgo cultural en los ítems.

Palabras clave: DIF, Indígenas, Sesgo, Sesgo Cultural, examen de Estado SABER 11°.

ABSTRACT

The present study is conducted in order to identify the possible sources of cultural bias in the tests items which compose the SABER 11° exam, taking the indigenous people as focal group. In a first phase, I detected the items with DIF in seven tests applied in the second semester of 2006 and the first semester of 2007 throughout Mantel-Haenszel and Difficulty difference procedures. After that, items of Language, Mathematics, Biology and Social sciences tests, which had been detected, were reviewed in focal groups. Results show that the procedures used for the detection of DIF are suitable for their application in real data and under the conditions of SABER 11°. On the other hand, the substantive analysis suggested three potential cultural bias sources which can be transversal to the tests: experiences more often in a group than in another, epistemology of the origin communities and construction problems; this last was a relative new source respect to previous studies. Additionally, I found other possible emergent four sources: technical terms, colonialism-religion-State, value judgments toward social theories and traditional school. All of them were used as basis to do some first guidelines in order to avoid the cultural bias in the items.

Key words: DIF, Indigenous, Bias, Cultural Bias, State exam SABER 11°.

TABLA DE CONTENIDO

RESUMEN	4
ABSTRACT	5
TABLA DE CONTENIDO	6
LISTA DE TABLAS	9
LISTA DE FIGURAS	10
INTRODUCCIÓN.....	11
REVISIÓN BIBLIOGRAFICA.....	19
Funcionamiento diferencial del ítem	20
Mantel-Haenszel (MH)	25
Diferencia de la dificultad.....	26
Sesgo	29
Sesgo Cultural.....	37
Examen de Estado de la Educación Media ICFES - SABER 11°	48
Prueba de Lenguaje.....	52
Prueba de Matemáticas.....	54
Prueba de Biología.....	55
Prueba de Sociales.....	57
Grupos indígenas y educación en Colombia.....	58
MÉTODO	66
Población y Muestra.....	66
Instrumentos.....	68
Procedimiento	69
Fase 1. Identificación de Ítems con DIF.....	69

Fase 2. Análisis substantivo de los ítems con DIF.....	75
RESULTADOS	81
Detección de Ítems con DIF.	81
Verificación de la Unidimensionalidad.....	81
Identificación de Impacto.....	85
Evaluación del Ajuste del modelo	86
Detección de ítems con DIF a través del MH y la diferencia de la dificultad	86
Análisis de sesgo.....	93
Clasificación de los Ítems Detectados con DIF por Componente y Competencia.	94
Sesión de la prueba de Lenguaje	96
Sesión de la prueba de Matemáticas	99
Sesión de la prueba de Biología	103
Sesión de Ciencias Sociales.	106
Posibles fuentes de sesgo transversales y nuevas fuentes de sesgo.	110
DISCUSIÓN Y CONCLUSIONES	117
REFERENCIAS	132
Anexo 1. Distribución de la población indígena según etnias por zonas territoriales del DANE y departamentos.....	141
Anexo 2. Resguardos Indígenas por Zonas Territoriales del DANE y Departamento.....	142
Anexo 3. Organización Lingüística de las Comunidades Colombianas.....	143
Anexo 4. Distribución porcentual por etnias de los estudiantes que presentaron SABER 11° en el segundo semestre de 2006 y que fueron considerados como indígenas para el análisis	144
Anexo 5. Distribución porcentual por etnias de los estudiantes que presentaron SABER 11° en el primer semestre de 2007 y que fueron considerados como indígenas para el análisis	145

Anexo 6. Protocolo para el análisis de sesgo cultural con expertos y constructores de ítems	146
Anexo 7. Acuerdo de confidencialidad y consentimiento informado	156
Anexo 8. Cuestionario sobre los ítems de las pruebas del examen de Estado SABER 11°.....	157
Anexo 9. Definiciones de fuentes de sesgo o factores culturales reportados en la literatura.....	158
Anexo 10. Formato para consignar las respuestas de los participantes en la revisión individual de ítems	162
Anexo 11. Número de veces que cada ítem fue detectado con DIF a través de las muestras por las pruebas estadísticas y las métricas de los procedimientos utilizados.....	163
Anexo 12. Promedio de las dificultades de los ítems con DIF en las muestras en las que fueron detectados por alguna métrica, total y para cada uno de los grupos.	171
Anexo 13. Algunas pautas para evitar sesgo cultural en los ítems.....	174

LISTA DE TABLAS

Tabla 1. Algunos métodos para la identificación de DIF.....	23
Tabla 2. Porcentaje de varianza explicado por los dos primeros componentes en los análisis de componentes principales (ACP)	82
Tabla 3. Número de factores y componentes sugeridos por el análisis paralelo.....	83
Tabla 4. Resultados del análisis de unidimensionalidad a partir del ACP de los residuales.	84
Tabla 5. Porcentaje de muestras clasificadas de acuerdo con su categoría de impacto, promedio de las diferencias de número de respuestas correctas entre los dos grupos y de la d de Cohen.....	85
Tabla 6. Porcentaje de ítems a través de las muestras de acuerdo con el ajuste al modelo evaluado a través del outfit.	86
Tabla 7. Porcentaje promedio de ítems detectados con DIF en las muestras para la II aplicación de 2006	88
Tabla 8. Porcentaje promedio de ítems detectados con DIF en las muestras para la I aplicación de 2007	88
Tabla 9. Porcentaje de ítems detectados con DIF según categoría de desajuste (outfit) para la II aplicación de 2006.....	89
Tabla 10. Porcentaje de ítems detectados con DIF según categoría de ajuste (outfit) para la I aplicación de 2007	89
Tabla 11. Ítems sugeridos para revisión por haber sido detectados con DIF en alguna de las muestras por al menos una de las métricas de los métodos usados.....	90
Tabla 12. Índice Kappa en la detección de ítems a través de las muestras entre pruebas estadísticas y métricas del MH y de la diferencia de la dificultad	93
Tabla 13. Número de Ítems que fueron detectados con DIF clasificados por competencia	94
Tabla 14. Número de Ítems que fueron detectados con DIF clasificados por componente.....	96
Tabla 15. Afirmaciones de los participantes en la sesión de análisis de sesgo de la prueba de Biología.....	104
Tabla 16. Afirmaciones de los participantes en la sesión de análisis de sesgo de la prueba de Sociales.	107

LISTA DE FIGURAS

Figura 1. Tipos de DIF..	23
Figura 2. Procedimiento para la identificación de ítems con DIF basado en el ΔMH	27
Figura 3. Proceso de identificación de sesgo en un ítem.....	33
Figura 4. Estructura del Examen de Estado aplicado en el primer Semestre de 2007..	52
Figura 5. Participación de indígenas respecto al total departamental.....	61
Figura 6. Porcentaje de nivel educativo por etnia..	62
Figura 7. CCI del ítem 20 de la prueba de Lenguaje de la segunda aplicación de 2007 y la muestra 15 que favorece al grupo focal.	91
Figura 8. CCI del ítem 2 de la prueba de Matemáticas de la segunda aplicación de 2007 y la muestra 14 que favorece al grupo de referencia.	92
Figura 8. Ítem con DIF en la prueba de Lenguaje que favorece a los no indígenas.....	97
Figura 9. Ítem con DIF de la prueba de Matemáticas que favorece a los no indígenas.	101
Figura 10. Ítem con DIF de la prueba de Biología y que favorece a los indígenas.....	106
Figura 11. Ítem con DIF en la prueba de Sociales que favorece a los no indígenas.	110
Figura 12. Ítem con DIF de la prueba de Sociales y que favorece a los no indígenas.	114

INTRODUCCIÓN

En 1995 la Misión para la Modernización de la Universidad Pública en Colombia en su informe final propone 15 estrategias con el fin de aumentar la eficiencia de la universidad pública y la calidad de la formación profesional, mencionando en la segunda estrategia “asegurar la equidad social como compromiso de la universidad pública, reafirmando su carácter de universidad para todos, pero focalizando esfuerzos sobre los sectores más vulnerables del país” (pág. 11). De acuerdo con ese informe, la ejecución de dicha estrategia requiere que las instituciones de educación superior (IES) canalicen esfuerzos, entre otros, para organizar nuevos criterios e instrumentos con el fin de que la selección de los mejores no sólo sea realizada a partir de una medición exclusiva de conocimientos, sino que esté orientada a identificar las aptitudes y las habilidades básicas, que están distribuidas de forma parecida entre los diferentes estratos.

Así mismo, el Educational Testing Service – ETS - (2009), afirma respecto a sus pruebas que:

“La revisión de la equidad es un paso esencial para contar con evaluaciones justas y válidas porque dicha revisión ayuda a asegurar que las pruebas del ETS muestran respeto por diversos grupos de personas, son sensibles a las necesidades y sentimientos de los evaluados, evitan imágenes y contenido que son insultantes o humillantes y están libres de barreras innecesarias para el éxito de todos los evaluados. La revisión de la equidad ayuda a asegurar que las personas que toman las pruebas del ETS no son distraídas innecesariamente al estar enfadadas, irritadas o frustradas por ellas.” (p. 2).

Estas apreciaciones no sólo son aplicables a las pruebas del ETS, sino que también se pueden extender a cualquier evaluación a amplia escala, en cualquier campo, bien sea educativo, clínico, organizacional, entre otros, y en cualquier país, incluyendo Colombia.

En este contexto, el Examen de Estado de la Educación Media ICFES - SABER 11°, desarrollado y aplicado por el Instituto Colombiano para la Evaluación de la Educación (ICFES) y reglamentado por el decreto 869 de 2010, tiene entre sus objetivos¹ “proporcionar a las instituciones educativas información pertinente sobre las competencias de los aspirantes a ingresar a programas de educación superior” (pág. 1, Decreto 869 de 2010) y es uno de los criterios más usados en las IES para otorgar sus cupos, en tanto que las oficinas de admisión son sus principales usuarias (Rocha & Pardo, s.f.). En consecuencia, si se quiere controlar el posible efecto de factores de inequidad en el ingreso a la educación superior, uno de los aspectos que debe considerarse es la identificación de cualquier tratamiento desigual en el examen de Estado SABER 11°, para grupos que de una u otra forma se han denominado o considerado desfavorecidos o minoritarios.

Una de las formas en que se pueden presentar desigualdades en los exámenes de Estado se refiere a lo que en la literatura psicométrica se conoce como sesgo en los ítems. El problema del sesgo en las pruebas ha sido de gran importancia con relación a la validez de las mismas, y mayormente, cuando las consecuencias de sus resultados afectan la vida de quienes están siendo evaluados, debido a que las pruebas suelen ser usadas para selección, certificación profesional o admisión (Ross & Okabe, 2006); siendo este último el caso de SABER 11°.

El análisis del Funcionamiento Diferencial del Ítem o FDI o DIF, por sus siglas en inglés, es el paso inicial para la detección de ítems con sesgo desde una perspectiva típica (Camilli y Sehpard, 1994), y es sugerido por la American Educational Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME) (1999) como un procedimiento estándar en cualquier evaluación a amplia escala. En términos sencillos, un ítem presenta DIF cuando su puntuación varía dependiendo de alguna variable como raza, género, etnia, idioma,

¹ Aunque estos son los objetivos correspondientes con la normatividad vigente, los objetivos anteriores del examen son muy similares, tal y como se comenta en un apartado posterior de este trabajo (Examen de Estado de la Educación Media ICFES - SABER 11°).

grupo cultural, entre otras (Herrera, Gómez e Hidalgo, 2005). Dichas variaciones se manifiestan en que examinados de grupos diferentes, pero que presentan la misma habilidad o cantidad de atributo que se está midiendo, tienen una probabilidad distinta de responder correctamente el ítem.

De acuerdo con Acar (2012), existen muchos procedimientos para identificar DIF, los cuales suelen clasificarse como relacionados con la teoría clásica de los test (TCT), basados en tablas de contingencia u originados en la teoría de respuesta al ítem (TRI). Sin importar la clasificación que se haga de los métodos para la detección de DIF, existe una prolífica investigación sobre su desempeño bajo diferentes condiciones; sin embargo dicha productividad de corte metodológico no ha ido acompañada por igual número de investigaciones sobre las fuentes de DIF (Elosua, 2006), es decir, sesgo como tal.

La idea de que existían ítems que varían en función de la cultura, por ejemplo el pertenecer o no a un grupo indígena, ya había sido manifestada por Binet & Simon en 1910, quienes consideraban que algunos ítems podrían medir efectos del entrenamiento cultural más que capacidad mental, ante lo cual eliminaron ciertas categorías de ítems de su prueba de inteligencia que eran sensibles a dichos efectos (Camilli & Shepard, 1994). Pero fue sólo después de las dos guerras mundiales, que comenzó a llamarse la atención acerca de que personas provenientes de culturas diferentes a aquellas en las que se había elaborado la prueba originalmente, se encontrarían en desventaja (Anastasi, 1967).

De acuerdo con Herrera (2005), a partir de una revisión de la literatura se encuentra que los estudios realizados por Eelles, Havighurst, Herrick & Tyler en 1951 y por Jensen en 1969 fueron el punto inicial para el desarrollo de la investigación sobre DIF. El primer trabajo mostró de forma empírica que una gran cantidad de ítems de pruebas de inteligencia eran sensibles a diferencias culturales, y el segundo tuvo un efecto muy importante sobre el desarrollo de procedimientos para la detección de DIF.

En los últimos treinta años se ha dado un incremento en las publicaciones científicas sobre sesgo y DIF, siendo la publicación de Holland & Thayer (1988) uno de los trabajos que ayudaron a precisar el significado de la palabra “sesgo”, en el cual se plantea cambiar

dicho término por el de DIF (Herrera, 2005), ya que consideraban que “sesgo” en muchos ejemplos de ítems con DIF no describe exactamente la situación. De acuerdo con lo anterior, “sesgo” se utilizaría para referirse a un ‘juicio informado’ (Holland & Wainer, 1993, p. xiv, citados por Herrera, 2005) que adicional a los resultados estadísticos tomara en cuenta el objetivo de la prueba y la información histórica, social, o cultural que explicara el DIF; mientras que “DIF” se utilizaría para referirse al hecho de que algunos ítems pueden mostrar propiedades psicométricas diferentes para grupos distintos (Fidalgo, 1996).

Más recientemente en un estudio bibliométrico, Gómez, Hidalgo, Guilera & Moreno (2005), encontraron que en los años 90’s hubo una proliferación de artículos científicos sobre DIF; lo que no resulta inesperado dado la preocupación creciente por garantizar igualdad de oportunidades y el tratamiento equitativo de las personas y los grupos sociales (Herrera, Gómez, Quintero, Arias, Berrío & Cervantes, 2007) y a que la equidad ha sido la prioridad en la evaluación educativa en las últimas décadas (Huang y Han, 2012).

De otra parte, los estudios sobre DIF y sesgo en los ítems en el contexto educativo se han desarrollado en mayor medida con grupos de género (Benbow, 1997; Chilisa, 2000; Kline, 2004; Pomplun & Omar, 2001; Stricker & Emmerich, 1999; Ryan & Chiu, 2001; Walstad & Robson, 1997; Wang & Lane, 1996). Para el caso del sesgo cultural (incluyendo los estudios con etnias y razas) la investigación ha sido menor; sin embargo se encuentran estudios, revisiones, discusiones y propuestas realizados en países que cuentan con multiplicidad de etnias y culturas debido a la constante migración o a que participan en la aplicación de pruebas educativas internacionales.

Entre los trabajos que han versado sobre el sesgo cultural desde diferentes perspectivas se encuentran los desarrollados por Da Costa & Araújo (s.f.), Gallart & Moore (2008), Banks (2006), Wu & Ercikan (2006), Tellegen & Laros (2004), Freedle (2003), Hunter & Schmidt (2000), Freedle & Kostin (1997), Fan, Willson & Kapes (1996), Helms & Van de Vijver (1995) y Veale & Foreman (1983), entre otros. Reconociendo la inseparable asociación entre lenguaje y antecedentes culturales o étnicos (Snetzler y Qualls, 2000), otros estudios relacionados con DIF y sesgo cultural son aquellos realizados con pruebas

adaptadas a un lenguaje diferente o que son aplicadas en un idioma que no corresponde con la lengua materna de los evaluados. Entre dichos estudios se encuentran los de Ross & Okabe (2006), Elosua (2006), Uiterwijk & Vallen (2005), Elder, McNamara & Congdon (2003), Baker (2001), Snetzler & Qualls (2000), Elosua, López, Egaña, Artamendi & Yenes (2000), Al-Fallay (1999) y Price & Oshima (1997),

Las investigaciones para la identificación de posibles fuentes culturales de DIF como las mencionadas anteriormente en exámenes educativos, deben ser consideradas como ejemplo a seguir en el contexto colombiano dada la multiculturalidad de nuestro país. Dichos estudios se requieren con mayor necesidad en pruebas de altas consecuencias (*high-stakes*, en términos de Kellaghan, Greaney & Murray, 2009) y de aplicación obligatoria como SABER 11°, ya que como se ha mencionado, tiene implicaciones para la admisión de estudiantes en diversas IES tanto oficiales como no oficiales.

En el caso particular de las comunidades indígenas colombianas, el posible sesgo en los ítems puede favorecerlas o desfavorecerlas, dadas sus diferencias culturales en la religión, la composición familiar, el contacto con el diseño de objetos y las pautas que siguen para el trabajo con los mismos (Gómez, 2000), la división del trabajo, el multilingüismo, y su arquitectura (Rubio, 2000), entre otros. Igualmente, dado que se considera que el español es una segunda lengua para algunas de las culturas indígenas colombianas, características como transferencia, exposición o práctica y los factores de socialización, mencionadas por Ross & Okabe (2006), podrían también considerarse como factores potenciales que favorecen o desfavorecen a los grupos indígenas a la hora de responder un ítem.

Por lo anterior, la importancia de estudiar el sesgo cultural en los ítems de las pruebas de SABER 11°, tomando como grupo focal (grupo minoritario) a los evaluados provenientes de grupos indígenas, surge en dos sentidos: a) la identificación de posibles fuentes de sesgo cultural que se sumen a las ya encontradas en otros estudios, y b) la contribución para mejorar procedimientos de construcción de ítems y evitar interferencia de

variables no relacionadas con el objeto de medida, es decir, con las competencias en las áreas fundamentales de la Educación Básica y Media.

Conociendo que los métodos para la detección de ítems con DIF pueden verse afectados por diferentes variables, el grupo de investigación “Métodos e Instrumentos en Ciencias del Comportamiento”, de la Universidad Nacional de Colombia, inició el proyecto de investigación “Identificación de Ítems con Sesgo Cultural en las Pruebas de los Exámenes de Estado en Colombia” en el año 2007, el cual fue financiado por COLCIENCIAS, la Universidad de Barcelona, España y la Universidad Nacional de Colombia. En la primera fase de dicha investigación se produjeron investigaciones sobre el desempeño de tres métodos bajo diferentes niveles de razón de tamaños (número de individuos en el grupo focal por cada individuo en el grupo de referencia), impacto, ajuste del modelo y porcentaje de ítems con DIF. Los tres estudios simulaban datos bajo condiciones similares a las que presenta el examen SABER 11° con relación a la comparación de dos grupos (estudiantes indígenas y no indígenas) y trabajaron con un diseño factorial completamente cruzado.

El primer estudio (Arias, 2008) arrojó como resultado que la razón de tamaños, el ajuste y el impacto tenían efecto sobre las tasas de falsos positivos y la potencia del χ^2 de Mantel-Haenszel (χ^2_{MH}). En contraste, la métrica de este procedimiento, el delta de Mantel-Haenszel (Δ_{MH}), fue robusto en todas las condiciones experimentales en cuanto al control del error tipo I. Específicamente, en razones de tamaño inferiores a 250 (1 estudiante del grupo focal por 250 estudiantes del grupo de referencia), el Δ_{MH} presentó tasas de falsos positivos cercanas a 0 y una adecuada tasa de identificación de DIF uniforme.

La segunda investigación (Berrío, 2008) probó el método diferencia de la dificultad, en la cual se encontró que la prueba estadística de este procedimiento presentaba inflación del error tipo I producida por todas las variables manipuladas y de la interacción entre razón de tamaños, impacto y ajuste del modelo sobre la potencia. Teniendo en cuenta lo anterior, se planteó una métrica análoga al Δ_{MH} , encontrando que presentaba un mayor control del error

tipo I que la prueba estadística, y era más robusta a la influencia de las variables mencionadas anteriormente sobre la potencia.

Finalmente, el tercer trabajo (Santana, 2009) mostró que la potencia de la regresión logística se ve afectada por razones de tamaños pequeñas (menores diferencias entre el tamaño del grupo focal y el de referencia), el impacto y el porcentaje de DIF. De la misma forma, la tasa de error tipo I fue mayor en razones de tamaños menores y cuando hay diferencias en la media de habilidad de los dos grupos (impacto).

Con base en los resultados hallados por los dos primeros estudios mencionados, el presente trabajo se ocupa de la siguiente fase del proyecto de investigación que corresponde a:

a) La aplicación a datos reales de las rutinas usadas en las simulaciones, con el fin de detectar posibles ítems con DIF, en las aplicaciones de SABER 11° del segundo semestre de 2006 y el primero de 2007, a través de los procedimientos Mantel-Haenszel y diferencia de la dificultad, tomando como grupo focal a los evaluados indígenas y como grupo de referencia a los demás evaluados.

b) La identificación de posibles fuentes de ese funcionamiento diferencial (sesgo) para, en caso de hallarlas, formular pautas que guíen la construcción de preguntas que eviten el sesgo cultural.

Se decide usar los procedimientos estudiados en las dos primeras investigaciones, porque al momento de iniciar la segunda fase del proyecto ya se habían concluido. Adicionalmente, los dos procedimientos escogidos cuentan con métricas que son más robustas que sus pruebas estadísticas bajo condiciones similares a las del examen objeto de estudio.

En términos generales, este trabajo es una contribución para mejorar procedimientos y evitar interferencia de variables no relacionadas con el objeto de medida de SABER 11°, es decir, con competencias en las áreas fundamentales de la Educación Básica y Media, a

partir de la identificación de potenciales fuentes de sesgo cultural en sus ítems. Tiene como objetivo general *establecer posibles fuentes de sesgo cultural en los ítems del examen SABER 11° aplicado en el segundo semestre de 2006 y en el primer semestre de 2007, con el fin de minimizarlas a través de propuestas de pautas dirigidas a evitar el sesgo cultural y que sirvan para futuros procesos de construcción de ítems*. Para la consecución de este propósito general se plantean los siguientes objetivos específicos:

1. Aplicar una rutina estadística con los procedimientos MH y diferencia de la dificultad a datos reales teniendo en cuenta su prueba estadística y métrica, con el fin de identificar los ítems con DIF en el examen SABER 11° llevado a cabo en el segundo semestre de 2006 y en el primero de 2007, teniendo como grupo focal a los evaluados indígenas.

2. Establecer potenciales fuentes de sesgo cultural en ítems identificados con DIF, mediante el análisis conjunto con constructores de preguntas del examen SABER 11° y expertos en comunidades indígenas, con el fin de que enriquezcan aquellas reportadas en la literatura.

3. Estructurar pautas que orienten y provean algunas estrategias para la construcción de ítems libres de sesgo cultural en pruebas del examen SABER 11°, para que las condiciones de equidad en la evaluación educativa en Colombia sean más plausibles.

REVISIÓN BIBLIOGRAFICA

Los estándares para las pruebas educativas y psicológicas - Standards for educational and psychological testing – desarrollados por la AERA, la APA y el NCME en 1999, definen validez como “el grado en el cual evidencia y teoría soportan las interpretaciones de las puntuaciones de la prueba implicadas en los usos propuestos para la misma.” (p. 9). En este sentido, cualquier variable o dimensión que no esté relacionada con lo que pretende medir y por la cual los ítems de una prueba se vean afectados, puede interferir con la interpretación de las puntuaciones que se hagan de un instrumento y por ende, está atentando contra la validez del mismo. Este tipo de efecto obedece a un error sistemático y es un problema del instrumento de medición.

En relación con las diferencias sistemáticas entre grupos sociales o culturales existen tres conceptos en la psicometría que, aunque están relacionados, son diferentes: Impacto, DIF y Sesgo. El impacto se refiere a diferencias reales en el desempeño en las pruebas entre grupos distintos (Barbero & Prieto, 1997; Gómez & Navas, 1998) y para establecer si existen dichas diferencias suele usarse como criterio la puntuación total en la prueba o en otros ítems de la misma (Camilli & Shepard, 1994). Según Millsap & Everson (1993), la identificación de la posible fuente de las diferencias entre grupos, lleva a distinguir el DIF del impacto, ya que cuando las diferencias entre los grupos se deben a diferencias reales (válidas) en la magnitud de atributo (habilidad objetivo) nos referimos a impacto (Ackerman, 1992) y cuando se deben a variables asociadas a los grupos se habla de DIF (Herrera, 2005). De acuerdo con ello, el impacto se manifiesta como diferencias en las distribuciones del atributo entre los grupos que se están comparando, las cuales pueden darse como diferencias en la media, en la desviación estándar o en ambas.

Cualquier estudio de sesgo debe incluir al menos dos etapas, la primera consiste en la detección de ítems con DIF y la segunda en un análisis que busca encontrar posibles explicaciones a ese comportamiento diferencial. Como se mencionó anteriormente, existen distintos métodos para la identificación de ítems con DIF, sin embargo algunos de ellos son

más adecuados que otros teniendo en cuenta las condiciones y características de los ítems, la población y la prueba, por lo que suelen ser necesarios estudios experimentales bajo condiciones similares a las reales para encontrar los de mejor desempeño en dichos escenarios. Por otra parte, la segunda etapa implica, como mínimo, un análisis substantivo de los ítems con DIF por parte de un equipo interdisciplinario, cuyo objetivo sea observar si existen características que expliquen ese comportamiento y evaluar si son relevantes o no frente a lo que se pretende medir, porque en el primer caso se hablará de impacto y en el segundo de sesgo.

Funcionamiento diferencial del ítem

Un ítem presenta DIF cuando la probabilidad de responderlo correctamente no depende únicamente del nivel de habilidad de la persona en el rasgo medido por la prueba sino también de otras variables, dicha probabilidad puede ser diferente para personas pertenecientes a grupos distintos, infringiendo de esta forma el supuesto de invarianza de medida (Elosua, 2006). En términos de Camilli & Shepard (1994), la idea es ver si evaluados comparables en la puntuación total en una prueba, pero de diferentes grupos, responden de la misma forma a los ítems individuales; en caso de que esto no ocurra el ítem es identificado como potencialmente sesgado o con DIF.

Según Herrera (2005) cuando se habla de DIF, es necesario contar como mínimo con dos grupos: el de referencia y el focal. El grupo de referencia se define como el grupo general o mayoritario de la población, el cual se supone tiene un rendimiento mayor y no es el grupo estudiado. Por otra parte, el grupo focal se refiere al grupo que es objeto de investigación y que generalmente se supone es el que se encuentra en desventaja, es decir, el desfavorecido y que también puede ser minoritario.

Existen diferentes tipos de DIF (Mellenbergh, 1982; Camilli & Shepard, 1994; Rogers & Swaminathan, 1993; Herrera, 2005): el uniforme, el no uniforme y el mixto. El primero se presenta cuando no hay interacción entre la magnitud de atributo y el grupo al que se pertenece (Rogers & Swaminathan, 1993), lo que sucede si la distribución de la magnitud de atributo es diferente entre los dos grupos y hay correlación entre las dimensiones

medidas por el ítem, o cuando los grupos difieren en la distribución de la dimensión irrelevante (Herrera, 2005), en estos casos, en términos de la TRI, el parámetro de dificultad cambia pero el de discriminación no (Cammilli & Shepard, 1994). En el DIF uniforme, cuando se observan las curvas características de los ítems (CCI) de los grupos en comparación, se observa que éstas son paralelas, porque la diferencia en la probabilidad de responder correctamente el ítem permanece constante a lo largo de los niveles del atributo. En contraste, el DIF no uniforme ocurre cuando hay interacción entre la magnitud de atributo y el grupo (Rogers & Swaminathan, 1993) y se da si “la diferencia entre los dos grupos está en la varianza de la distribución de la dimensión irrelevante o en la correlación entre las dos dimensiones; puede ocurrir que el parámetro de dificultad sea el mismo pero el de discriminación no” (Herrera, 2005 p. 25) cuando se usa un modelo de dos parámetros de la TRI. Al observar las CCI de los dos grupos en comparación, se tiene que éstas se cruzan porque la diferencia en la probabilidad de acertar al ítem no es constante en todos los niveles de la habilidad. Finalmente, se habla de DIF mixto cuando ambos parámetros, dificultad y discriminación, son diferentes para los grupos (Herrera, 2005).

Los conceptos de impacto y los tipos de DIF se representan en la figura 1, la cual muestra el comportamiento de cuatro ítems a través de sus curvas características para dos grupos. En el marco de la TRI, el eje X de todas las subfiguras de la figura 1 representa los diferentes niveles de habilidad o de magnitud del atributo (θ), que van desde $-\infty$ a ∞ , y el eje Y la probabilidad de acertar el ítem dado determinado nivel de habilidad. En la TRI se habla en general de tres parámetros: dificultad (b), discriminación (a) y pseudo azar (c). El primero corresponde al valor de θ en el punto de máxima pendiente de la CCI (Herrera, Gómez & Muñiz, 2007). Así, la dificultad corresponderá a la habilidad necesaria para tener una probabilidad igual a 0,5 de puntuar en el ítem. El parámetro a se define como el nivel de inclinación de la CCI o la pendiente de la recta tangente a la CCI en su punto de mayor inclinación (Herrera et al., 2007). Finalmente, el pseudo azar corresponde a la asíntota de la CCI cuando θ tiende a $-\infty$ o también puede definirse como la mínima probabilidad de acertar (Herrera et al., 2007).

De acuerdo con lo mencionado anteriormente, la figura 1a muestra un ítem sin DIF en presencia de impacto, es decir, existen diferencias en términos de habilidad entre los dos grupos, donde el grupo B tiene una mayor habilidad (θ), dado que su media es superior. La 1b muestra un ítem con DIF uniforme, el cual es más fácil para el grupo A ($b_A = -1$) que para el grupo B ($b_B = 0$), aunque no hay impacto, adicionalmente las dos CCI se observan paralelas. La figura 1c ilustra un ítem con DIF no uniforme, en donde la dificultad es igual para los dos grupos ($b = 0$), pero la discriminación no, por lo que en los niveles bajos de habilidad los individuos pertenecientes al grupo B presentan una mayor probabilidad de responder correctamente el ítem y en los niveles altos los del grupo A; en este caso tampoco se presenta impacto. La 1d muestra un ítem con DIF mixto, puesto que la dificultad es diferente para los grupos ($b_A = -1$, $b_B = 0$) y en los niveles muy bajos de habilidad el ítem es más fácil para el grupo B y en los más altos para el grupo A; en este caso también hay diferencias en la habilidad de los dos grupos teniendo mayor magnitud de atributo el grupo B en promedio.

De acuerdo con Acar (2012) existen muchos métodos para identificar DIF. Algunos de ellos están fundamentados en las tablas de contingencia y otros se originan en la teoría de respuesta al ítem (TRI). Los métodos clasificados como de tablas de contingencia se basan en el estudio de las asociaciones que se presentan entre dos variables categóricas, siendo en este caso el grupo y la respuesta en el ítem, cuando se tiene en cuenta otra variable de carácter numérico o categórico, es decir, la cantidad de atributo medida a partir del puntaje total en la prueba (Herrera, 2005). Entre dichos métodos están las aplicaciones del χ^2 , el método de estandarización, el Mantel-Haenszel (MH), los modelos log-lineales y logit, y la regresión logística (Herrera et al, 2005). Los procedimientos basados en la TRI para la detección del DIF, radican en la comparación de los modelos ajustados de forma separada para los grupos de interés, la idea es que si para un ítem la probabilidad de acertar se puede expresar a través del mismo modelo para los dos grupos, éste no presenta DIF, pero si hay diferencias entre los dos, éstas dan cuenta de DIF (Herrera, 2005). Entre los métodos basados en la TRI se encuentran: la diferencia del parámetro b (dificultad), el χ^2 de Lord, la medida de áreas (Camilli & Shepard, 1994) y el índice de DIF no compensatorio –

NCDIF- (Raju, van der Linden & Fleer, 1995). La tabla 1 muestra la descripción de diferentes métodos para la detección de DIF junto con sus potenciales ventajas y desventajas, la cual se basa principalmente en la revisión de Herrera et al. (2005), Herrera et al. (2007) y Santana (2009); aunque se adicionan algunas referencias respecto al NCDIF.

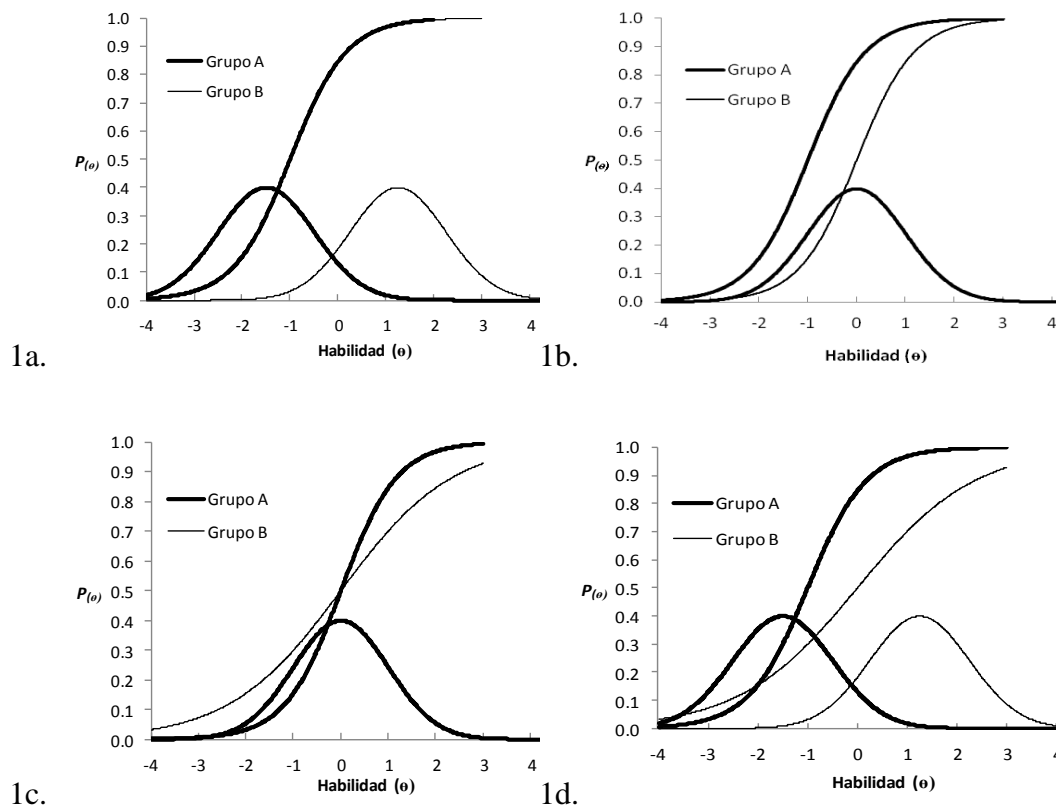


Figura 1. Tipos de DIF. 1a: Ítem sin DIF con presencia de impacto; 1b: Ítem con DIF uniforme y sin presencia de impacto; 1c: Ítem con DIF no uniforme y sin presencia de impacto; 1d: Ítem con DIF mixto en presencia de impacto. Figura adaptada de Herrera (2005).

Tabla 1. Algunos métodos para la identificación de DIF.

Procedimiento (Autor)	Descripción	Algunas ventajas y desventajas
Aplicaciones de X^2 (Scheuneman,1981)	Plantea que se puede descartar la presencia de DIF si la proporción de aciertos es igual para los grupos en cada uno de los estratos (Herrera et al., 2005).	Sencillez y economía. Incapacidad para detectar DIF no uniforme. Inestabilidad cuando hay valores bajos en las celdas, diferencias en los tamaños de muestra en los grupos o presencia de impacto (Herrera et al., 2005).
Método de estandarización. (Kulick & Dorans,	Comparación del comportamiento en un ítem en los grupos controlando por la magnitud de atributo (Herrera et al.,	Cuenta con una métrica y prueba de hipótesis.

Procedimiento (Autor)	Descripción	Algunas ventajas y desventajas
1983)	2005).	
Mantel Haenszel (Holland & Thayer 1986, 1988)	Si un ítem no presenta DIF, la razón entre el número de personas que lo fallan y lo aciertan es igual para los grupos en diferentes niveles de habilidad (Herrera et al., 2005).	Sencillez y economía computacional. Es eficiente al manejar diferentes niveles de habilidad como variable de control. Cuenta con una métrica. Deficiencias al detectar DIF no uniforme. Posible contaminación en el criterio (puntajes observados en la prueba) por cuenta de los ítems con DIF (Herrera et al., 2005).
Modelos log-lineales y modelos logit (Mellenbergh,1982)	Estos modelos buscan ajustar un modelo cuyo fin es predecir la frecuencia esperada de cada celda de la tabla de contingencia como producto de los efectos incluidos en el modelo. (Herrera et al., 2005).	Permite describir las características de las variables que conforman el modelo y sus interacciones. Posible contaminación en el criterio (puntajes observados en la prueba) por cuenta de los ítems con DIF (Herrera et al., 2005).
Regresión logística. (Spray & Carlosn , 1986; Bennet, Rock & Kaplan, 1987; Swaminathan & Togers, 1990).	Caso particular del análisis de regresión múltiple cuando la variable dependiente es dicótoma (Herrera et al., 2005).	Facilidad para ajustarse al análisis de ítems politómicos o a situaciones en las que se tienen más grupos. Capacidad para detectar DIF uniforme y DIF no uniforme. Posible contaminación en el criterio (puntajes observados en la prueba) por cuenta de los ítems con DIF (Herrera et al., 2005).
χ^2 de Lord. Lord (1980)	Compara los vectores de los parámetros estimados cuando se ajustan modelos TRI para dos grupos de forma separada (Herrera et al., 2007).	Procedimiento sencillo de utilizar, cuenta con una prueba estadística y algunos estudios aplicados respaldan su uso. Dificultades de aplicación con modelos de tres parámetros. Exigencias en tamaños de muestra, supone que θ es conocido y es aplicable únicamente con algoritmos de máxima verosimilitud. Desconocimiento del tamaño de muestra necesario para lograr la convergencia a la distribución χ^2 y posible alta tasa de falsos positivos (Herrera et al., 2007).
Medidas de área entre las CCI. Rudner (1977), Rudner, Getson & Knight (1980) Raju (1988)	Evalúa el área entre las dos CCI de un ítem que son ajustadas para dos grupos de forma independiente y expresadas en la misma métrica (Herrera et al., 2007).	Sencillo y fácil de calcular. Exigencias en el tamaño de muestra. Algunos estudios no favorecen su uso (Herrera et al., 2007).
Diferencia del parámetro b .	Comparación de los parámetros de dificultad para dos grupos controlando por el nivel de habilidad.	Debilidades para detectar DIF no uniforme y puede ser engañoso en situaciones en las que se ajusten mejor modelos de dos o tres parámetros (Camilli & Shepard, 1994).
Índice de DIF no compensatorio (NCDIF) Raju, van der Linden & Fleer (1995)	Se basa en cuantificar las diferencias entre las CCI de los grupos focal y de referencia. Asume que todos los ítems, excepto el que es objeto de estudio, no presentan DIF y se considera no aditivo ni compensatorio (Oshima, Raju,	Es preciso cuando se compara con la prueba del χ^2 de Lord y la medida entre las áreas (Raju et. al., 1995). Puede verse afectado por variables como razón de tamaños entre los grupos y su valor parece estar relacionado con ciertos rangos de dificultad y discriminación

Procedimiento (Autor)	Descripción	Algunas ventajas y desventajas
	Flowers & Slinde, 1998).	(Bolt, 2002; Oshima, Raju & Nanda, 2006). También, la prueba estadística que se ha propuesto no se aproxima a una distribución χ^2 (Oshima et al., 2006; Raju et al, 1995).
SIBTEST Shealy & Stout (1983)	Consiste en una técnica basada en la teoría multidimensional propuesta por Ackerman (1992) que permite la identificación del DIF en uno o conjuntos de ítems (bundle ítems) de forma simultánea (Herrera, 2005).	Cuenta con una medida de tamaño del efecto del DIF. Más sensible a las diferencias entre dificultad y discriminación, e identifica erróneamente ítems con DIF cuando poseen parámetros de dificultad y discriminación similares; sin embargo en condiciones de modelos de tres parámetros o discriminación media o baja, el error tipo I disminuye (Akelo, 2008, citado por Santana, 2009).

Por otra parte, dado que el MH es catalogado como el *Gold Standar* de los procedimientos para la detección de DIF y es uno de los procedimientos que se usará en la presente investigación se describirá con más detalle. Así mismo, se presentará el procedimiento diferencia de la dificultad, también por ser el otro método elegido para este trabajo.

Mantel-Haenszel (MH).

De acuerdo con Elosua (2006) el MH, es un procedimiento no paramétrico modificado por Holland & Thayer (1988) que evalúa la igualdad entre las proporciones de respuestas correctas e incorrectas de los grupos focal y de referencia estratificados tomando en cuenta los niveles en que se ha dividido la habilidad, dando como resultado odds-ratio, cuyo estimador es el α_{MH} . El estadístico para probar $H_0: \alpha_{MH} = 1$ contra $H_1: \alpha_{MH} \neq 1$ es el χ^2 de MH que sigue una distribución χ^2 con un grado de libertad (Herrera et al., 2005). La evaluación del tamaño del DIF se realiza a partir de una transformación a la escala delta de los odds-ratio, cuyo resultado se denomina delta mantel (Δ_{MH}), e indica la diferencia entre las dificultades del grupo de referencia y el focal; los valores positivos indican que el ítem es más fácil para el grupo focal y valores negativos que lo es para el grupo de referencia (Kamata & Vaughn, 2004). De acuerdo con la clasificación establecida por el ETS, un ítem presenta DIF severo (Categoría C) cuando la prueba estadística es

significativa y el valor del Δ_{MH} es superior a 1.5, y es moderado (Categoría B) cuando está entre 1 y 1.5 y la prueba estadística también es significativa (Elosua, 2006).

De acuerdo con los resultados de Arias (2008), la prueba estadística del MH, el χ^2_{MH} , se deja afectar por variables como la razón de tamaños, el ajuste y el impacto encontrando que se presenta un mayor error tipo I cuando la razón es menor a 250, hay impacto y desajuste del modelo, aunque mostró una adecuada potencia para la identificación de DIF no uniforme. En contraste, la métrica del Δ_{MH} tuvo bajas tasas de error tipo I y un buen poder en la identificación de DIF uniforme en las mismas condiciones. De acuerdo con lo anterior, la combinación de la prueba estadística del MH y del Δ_{MH} como medida de tamaño del efecto, permiten una mayor potencia de prueba y una menor proporción del error tipo I, para las condiciones de SABER 11°. Entidades como el ETS usan la métrica del Δ_{MH} para identificar ítems clasificados como C, los cuales son eliminados y luego revisados para establecer cuáles son las razones del DIF (Arias, 2008). La figura 2, ilustra el procedimiento del MH, teniendo en cuenta su métrica.

Diferencia de la dificultad.

Este procedimiento se enmarca en el modelo de Rasch o de un parámetro (b) de la TRI. En este modelo se asume que la probabilidad de acertar a una pregunta depende únicamente de la habilidad del evaluado y la dificultad de la pregunta, por lo que la detección de DIF se hace a partir de la comparación de la dificultad del ítem de los dos grupos (Angoff, 1993, citado por Berrío, 2008). La prueba estadística de este procedimiento viene dada por la ecuación 1.

$$t = \frac{b_1 - b_2}{\sqrt{SE_{b_1} + SE_{b_2}}}$$

(1)

Donde:

b_1 es la dificultad para el grupo 1.

b_2 es la dificultad para el grupo 2.

SE_{b_1} es el error estándar de la dificultad para el grupo 1.

SE_{b_2} es el error estándar de la dificultad para el grupo 2.

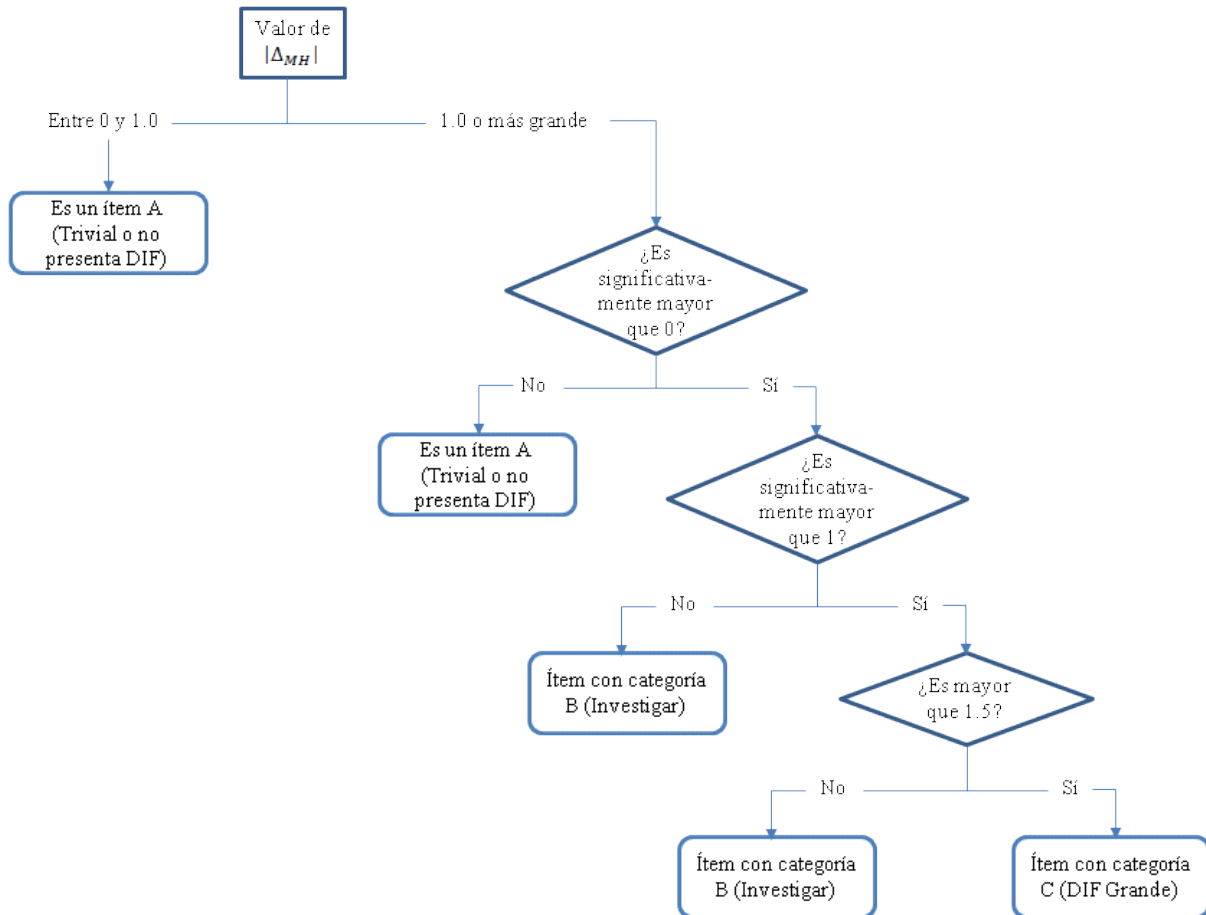


Figura 2. Procedimiento para la identificación de ítems con DIF basado en el Δ_{MH} . Adaptado de Kamata & Vaughn (2004).

En el estudio de Berrío (2008) se encontró que la razón de tamaños entre los grupos, el ajuste del modelo, el impacto y el porcentaje de ítems con DIF producían un aumento en el error tipo I de la prueba estadística; además, la interacción entre la razón de tamaños, el impacto y el ajuste del modelo incidían sobre la potencia de la prueba. Con el fin de establecer criterios empíricos sobre el DIF en los ítems, que controlaran el error tipo I sin perder potencia, Berrío (2008) propuso una métrica análoga (D) al Δ_{MH} , la cual se describe

en la ecuación 2. De acuerdo con la autora de la métrica (comunicación personal, Noviembre 25 de 2010) se recomienda un punto de corte $|0.22|$ debido a que es el que permite un buen control del error tipo I sin perder potencia de prueba.

$$\text{Métrica } D = \left[\ln \left(\frac{b_{r15,2.5}}{b_{f15,2.5}} \right) \right] * -2.35 \quad (2)$$

Donde:

$b_{r15,2.5}$ es el parámetro de dificultad del grupo de referencia con media 15 y desviación 2.5.

$b_{f15,2.5}$ es el parámetro de dificultad del grupo focal con media 15 y desviación 2.5, en la misma escala que el grupo de referencia.

En esta métrica, si el signo de la D es negativo favorece al grupo focal, si es positivo favorece al de referencia. En términos generales, Berrío (2008) encontró que esta métrica ofrecía un mejor control del error tipo I y era más robusta frente a variables como la razón de tamaños, impacto y ajuste del modelo, comparada con la prueba estadística. En términos generales, Berrío señala que la prueba estadística presenta tasas adecuadas de error tipo I y detección correcta cuando las razones de tamaño están entre $1/500$ y $1/20$, en cualquier condición de impacto, con 10% de ítems con DIF, DIF uniforme y con modelos ajustados. Aunque esto puede variar en otras condiciones, destacándose que en esos casos la razón de $1/20$ es la que presenta constancia en cuanto a su buen comportamiento en interacción con otras variables.

En cuanto a la métrica propuesta para este procedimiento, Berrío (2008) indica que se obtuvieron tasas de error tipo 1 cercanas al criterio liberal de Bradley (1978) cuando el modelo estaba ajustado, las razones de tamaño oscilaban entre $1/20$ y $1/250$ y bajo cualquier condición de impacto; en general se observó que las tasas de falsos positivos se controlan más que las obtenidas con y sin purificación del procedimiento. Además, la potencia no se vio afectada por las variables manipuladas en su estudio, encontrándose tasas entre 0.63 y .69 en condiciones de 10% de ítems con DIF. En cuanto al DIF no

uniforme, las tasas estuvieron entre .13 y 1 cuando el modelo estaba desajustado. La mencionada autora señala que esta métrica se puede aplicar en situaciones de ajuste del modelo, independientemente de si hay impacto y con razones entre 1/20 y 1/250. Se destaca la estabilidad de la métrica a través de diferentes razones de tamaño entre los grupos cuando se presenta un modelo de la TRI ajustado.

Sesgo

El sesgo puede definirse como un “error sistemático que distorsiona el significado de las puntuaciones y que está causado por la intervención de habilidades espurias junto a la habilidad principal en un ítem” (Ackerman, 1992; Mellenbergh, 1989; Shealy & Stout, 1993; citados por Elosua, López, & Torres, 2000, p. 198), y dado que es un error sistemático se puede incluir su evaluación en el análisis de validez. Se cree comúnmente que el sesgo se debe a alguna característica no deliberada del ítem de la prueba (Banks, 2006) que da una ventaja injusta a un grupo de examinados sobre otro (Clauser & Mazor, 1998).

De acuerdo con Herrera (2005), el sesgo se puede diferenciar del DIF, en dos aspectos. El primero radica en que el sesgo incluye además de detectar los ítems con DIF, identificar las causas de dicho funcionamiento, lo que va más allá del procedimiento estadístico (Gómez e Hidalgo, 1997; Muñiz, 1997) y que va en conjunción con un análisis lógico (Camilli & Shepard, 1994). Dicho análisis lógico implicaría la identificación de los factores que producen el sesgo y la discusión sobre la importancia de dichos factores en el constructo que pretende medir la prueba. El segundo aspecto, se refiere a que si un ítem presenta DIF, no necesariamente implica que esté sesgado, ya que puede estar midiendo algo más de lo que se pretende medir; solamente si un ítem es relativamente más difícil para un grupo que para otro y la fuente de esa dificultad es irrelevante para el constructo que se quiere medir, el ítem estará sesgado (Camilli & Shepard, 1994). Dicho de otra manera, si la causa del DIF fuera relevante para la variable que se está midiendo, el ítem puede determinarse como no sesgado (Muñiz, 1997). Esto último conllevará a que quienes

tengan la misma cantidad de atributo en lo que está midiendo el ítem, tendrán la misma probabilidad de puntuar y no provocará diferencias irrelevantes entre grupos.

Una integración de los conceptos mencionados anteriormente la ofrecen Gómez & Navas (1998) al afirmar que el objetivo principal de un estudio de sesgo es discernir cuándo las diferencias de grupo reflejan diferencias existentes entre los grupos de habilidad, conocimiento o experiencia (impacto), o sesgo, es decir, cuándo reflejan diferencias artificiales originadas en el proceso de medida como tal. En este sentido, y en relación directa con la validez, lo importante es “determinar si la causa por la que los grupos puntúan diferente en un ítem es relevante o irrelevante frente al constructo medido, ya que en el primer caso se tratará de impacto y en el segundo de sesgo” (Gómez & Navas, 1998, p. 686).

A continuación se presentan algunos ejemplos que pueden ayudar a diferenciar entre lo que es DIF, sesgo e impacto. Zieky (1992) plantea los dos siguientes ítems:

- 1) ¿Cuánto es $5.3 \times 1,000$?
 - a. 53
 - b. 530
 - c. 5,300
 - d. 53,000
- 2) ¿Cuántos metros hay en 5.3 kilómetros?
 - a. 53
 - b. 530
 - c. 5,300
 - d. 53,000

Si los ítems están pretendiendo medir habilidades para multiplicar y presentan DIF, en el primer caso el ítem no presenta sesgo o es justo. Sin embargo el segundo puede ser el ejemplo contrario porque conocer el sistema métrico no es parte de lo que se pretende medir y existen grupos de personas que están menos familiarizadas con el sistema métrico presentado en el ítem 2. Dicho sesgo desaparecería en el segundo ítem, si éste pretendiera

medir habilidades de conversión en el sistema métrico (Zieky, 1992). También este mismo autor afirma que el contexto del ítem es importante al momento de revisar la equidad, y lo ejemplifica en la siguiente pregunta:

“Un modem de un computador puede enviar 1,200 bits por segundo. ¿Cuánto le tomará enviar un mensaje de 720,000 bits?”

Aunque para responderlo, si se trata de habilidades para multiplicar, no se requiere saber de computadores, este contexto puede hacer que los evaluados eviten responder el ítem porque se sienten intimidados si no están familiarizados con el tema, así posean las habilidades aritméticas necesarias para contestarlo correctamente.

Otro ejemplo sobre la diferencia entre sesgo e impacto, puede tomarse de Camilli & Shepard (1994) en el que se manifiesta que un ítem de comprensión de lectura relacionado con béisbol puede ser más difícil para las mujeres que para los hombres debido a su falta de familiaridad con la terminología del béisbol. Dado que conocer de béisbol no es necesario para responder un ítem de comprensión lectora, ni está relacionado con dicho constructo, el ítem se podría considerar sesgado. Sin embargo, si la prueba está evaluando conocimientos en deporte, el ítem no estará sesgado, sino que por el contrario probablemente esté revelando diferencias reales, es decir, impacto.

El ETS (2009) establece tres grandes fuentes de varianza irrelevante que se pueden presentar al medir los constructos: cognitivas, afectivas o físicas. Estas fuentes lo llevaron a establecer lineamientos para la revisión de la equidad en sus evaluaciones.

Las fuentes cognitivas se presentan cuando para contestar correctamente el ítem se requiere de conocimientos o habilidades que no están relacionados con el propósito de la prueba y que no están distribuidos equitativamente entre los grupos. Las fuentes afectivas se dan si el lenguaje o las imágenes ocasionan emociones fuertes que pueden interferir con la habilidad para responder un ítem correctamente, por ejemplo contenido ofensivo, molesto o controversial, que dificulta que el evaluado se concentre en el significado de una lectura o la respuesta a un ítem, llegando incluso a responder de forma emocional más que

lógica. Finalmente, las fuentes físicas ocurren cuando aspectos de la prueba interfieren con la habilidad del evaluado para prestar atención, ver, escuchar o percibir los ítems, lo cual se presenta con mayor frecuencia en evaluados con discapacidad.

Entre las fuentes cognitivas de varianza irrelevante señaladas por el ETS (2009) se encuentran: Lenguaje difícil innecesario, regionalismos y tópicos específicos como aspectos militares, herramientas especializadas, deportes o religión. Estos tópicos dejan de ser irrelevantes si se encuentran directamente relacionados con lo que se quiere medir y por tanto con la validez del constructo.

El material que es sensible para ciertos grupos; la experiencia previa de los evaluados; franqueza (*directness*) del material; aspectos extremadamente amenazadores de accidentes, enfermedades, desastres naturales o muerte; apoyo a ideologías o causas particulares; la evolución; diferencias entre grupos; imágenes para poblaciones internacionales; suntuosidad; entre otros, pueden encontrarse entre las fuentes emocionales de varianza irrelevante, a menos que estos temas se encuentren relacionados con la validez del constructo, en cuyo caso se recomienda tratarlos de tal forma que causen el menor impacto emocional posible.

Respecto a las fuentes físicas de varianza irrelevante, existen características de los ítems como los elementos de ayuda y aspectos innecesarios que pueden no encontrarse relacionados con lo que se quiere medir, por ejemplo, tiras cómicas a partir de las cuales los evaluados deben escribir un texto o estímulos visuales redundantes. En general, frente a las posibles fuentes de varianza irrelevante en los ítems, el ETS da algunas recomendaciones, las cuales pueden consultarse en ETS (2009).

En cuanto a la metodología del análisis del sesgo, Arias (2008) expresa de manera gráfica (figura 3) el proceso de identificación de sesgo en un ítem a partir de las recomendaciones que hace el ETS en Zieki (1993). También señala, de acuerdo con Hambleton, Clauser, Mazor y Jones (1993), que el análisis del sesgo requiere un análisis profundo de los parámetros de la prueba, resultados estadísticos de DIF y evaluación de expertos en el área que permita identificar y corregir las posibles fuentes de sesgo.

Pero la revisión o juicio de expertos no es la única forma de evaluar sesgo en los ítems después de que son identificados con DIF. Padilla, García y Gómez (2007) señalan que a través de otros procedimientos como las entrevistas cognitivas y grupos focales se pueden identificar fuentes de error de medida en las diferentes fases del proceso. Igualmente, también se realizan análisis a través de metodologías como pensar en voz alta (thinking aloud), en la que participan directamente los evaluados.

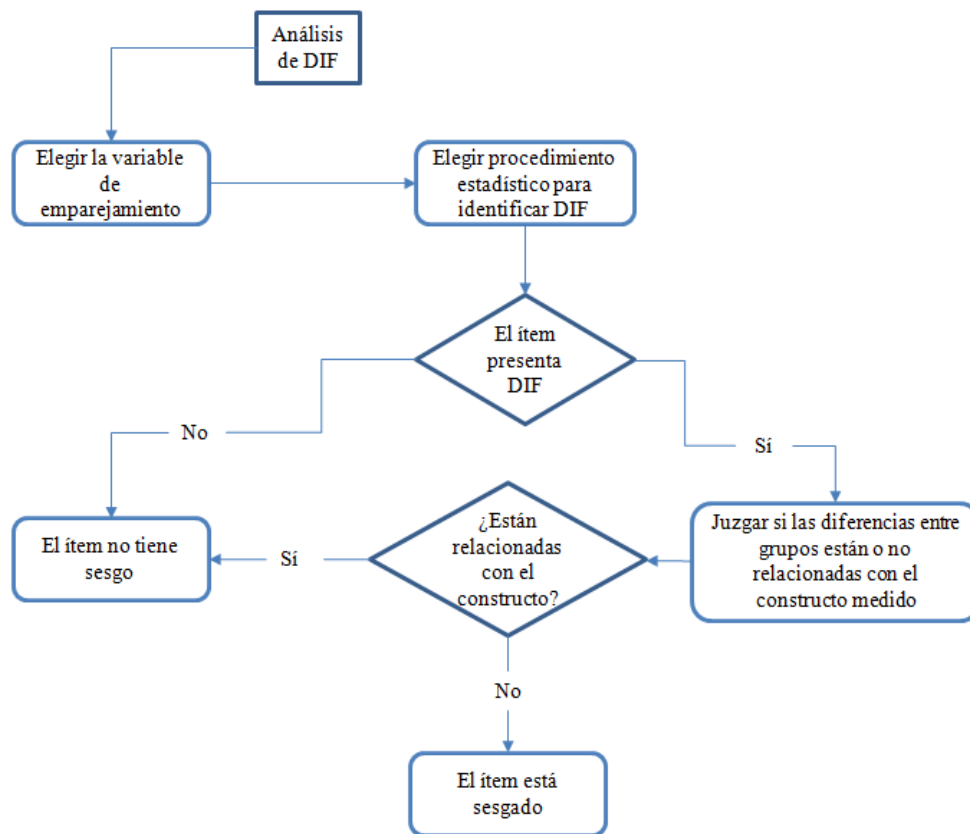


Figura 3. Proceso de identificación de sesgo en un ítem. Adaptado de Arias (2008).

De acuerdo con Ercikan, Arim, Law, Domene & Lacroix (2010) la revisión por expertos es el método más comúnmente usado para identificar características de los ítems, incluidas las fuentes de DIF en estudios de evaluación multilingüe o que se enfocan en diferencias étnicas o de género. Estas revisiones incluyen la evaluación de contenido del ítem, lenguaje y formato como indicadores de fuentes de DIF. Esta metodología ha mostrado su utilidad en la identificación de fuentes de DIF en diferentes estudios aplicados

como los de Zieky (1992), Acosta & Padilla (2004, citados por Padilla et al., 2007), Gierl & Nyla (2001) y Sireci & Allouf (2003), aunque otros señalan que deben ser complementados con otros tipos de análisis (Ercikan et al, 2010).

De acuerdo con Padilla et. al (2007) en el estudio de Acosta & Padilla (2004) se usó el juicio de expertos para identificar posibles fuentes de DIF entre las distintas versiones de la adaptación de la versión alemana del Inventario de Autogobierno al español, mientras que el trabajo de Zieky (1992) presentó la evaluación de equidad en los ítems en el contexto de un examen de licencia para docentes en la que se usó paneles de revisores.

En este segundo estudio se trabajó con 400 revisores y 7000 ítems. Todos los revisores recibieron un entrenamiento a través de correo y de una discusión de más de tres horas antes de iniciar la evaluación de los ítems. En la discusión se invirtió una gran cantidad de tiempo abordando el tema de la equidad por ser el juicio más complicado y controversial de los que tenían que hacer. El entrenamiento fue diseñado para sensibilizar a los revisores frente a las causas potenciales de inequidad, con el fin de disminuir la probabilidad de que cualquier ítem injusto pasara sin ser detectado y para ayudarlos a diferenciar entre problemas de equidad y otros aspectos de calidad del ítem. También se entrenaron en diferenciar cuándo un aspecto hacía que se presentaran diferencias relevantes (válidas) o irrelevantes (inválidas) entre los grupos en un ítem. Debido a que no todos los ítems podrían ser analizados en los encuentros posteriores al entrenamiento, los revisores evaluaron cada ítem de forma individual. En general, este trabajo permitió identificar aspectos de los ítems que pueden llevar a que los expertos consideren como sesgados ítems que realmente no lo son. Por ejemplo, durante el entrenamiento, los revisores tendieron a ver como injustos ítems que eran difíciles o ambiguos, o que presentaban distractores fuertes o claves (respuestas correctas) débiles, siendo catalogados como injustos para todos. Frente a ello, los entrenadores tuvieron que hacer la aclaración que si un ítem afectaba a los grupos por igual no era injusto. También los evaluadores tendieron a clasificar ítems como injustos si consideraban que algunos grupos particulares de evaluados podrían haber tenido menos oportunidad de aprender lo abordado por el ítem o porque no era enseñado en sus

escuelas; sin embargo, si el ítem evaluaba algo que era importante para lo que se pretendía medir, no se podía considerar como injusto.

En su trabajo Gierl & Nyla (2001) tenían como objetivo identificar fuentes de DIF en los ítems y en conjuntos de ellos que responden a un mismo contexto (*bundle items*) en la traducción de pruebas de desempeño a través de análisis substantivos y estadísticos. El primer tipo de análisis fue llevado a cabo por un comité de 11 miembros, los cuales encontraron cuatro fuentes de DIF. Después, dos traductores certificados categorizaron otro conjunto de ítems con DIF para grado sexto y noveno de las pruebas de Desempeño en Matemáticas y Estudios Sociales (Mathematics and Social Studies Achievement Tests) en las cuatro fuentes y el grupo al cual favorecían. Los investigadores hicieron un análisis estadístico para cada categoría de ítems a través del SIBTEST. Los traductores predijeron correctamente ocho de los 13 ítems, respecto al grupo al que favorecían. Los autores concluyen que la combinación del análisis substantivo y de los análisis estadísticos ayudan a los investigadores a encontrar y confirmar fuentes de DIF que pueden ser usadas para diseñar guías y principios de construcción de ítems que permitan reducir el DIF en pruebas traducidas.

El estudio de Sireci & Allouf (2003) también trabajó con expertos para la identificación de las posibles fuentes de DIF al comparar grupos de evaluados rusos y hebreos en la subprueba verbal de la Prueba Psicométrica de Entrada (Psychometric Entrance Test –PET-), una prueba de altas consecuencias usada para la admisión a las universidades de Israel. Posterior al análisis de DIF y a la identificación de la dirección del mismo, cinco traductores Ruso-Hebreo analizaron el tipo y el contenido de los ítems con DIF, sin conocer su clasificación. Cada traductor revisó de forma individual un cuestionario sobre 60 ítems, los cuales incluían ítems DIF e ítems sin DIF. Después de la revisión, se conformó un comité de ocho personas que incluían a los cinco traductores y a tres investigadores en lenguaje hebreo. Durante la reunión del comité, se les brindó información sobre el DIF encontrado en los ítems y cada revisor que hubiese hecho una suposición de DIF expresó su opinión. Los ítems sin DIF sirvieron para propósitos de comparación. El objetivo general del proceso era encontrar acuerdos sobre las causas de DIF en los ítems

que lo presentaron. Como resultado, los investigadores plantearon cuatro hipótesis posibles de las causas de DIF: cambios en la dificultad de las palabras y oraciones, cambios en el contenido, cambios en el formato y diferencias en la relevancia cultural.

Erickan et al. (2010) investigaron sobre el uso de protocolos de pensar en voz alta como una estrategia para evaluar y confirmar fuentes de DIF. Dichos protocolos fueron usados para investigar en qué medida las características de los ítems que son identificadas por expertos como fuentes de DIF son soportadas por los procesos de pensamiento de los evaluados en las versiones inglesa y francesa de una evaluación nacional canadiense. Los ejercicios de pensar en voz alta fueron grabados, transcritos y revisados por los miembros del grupo de investigación, y las transcripciones y las notas que habían tomado los administradores del ejercicio fueron los datos usados para el análisis. La codificación de los datos tomó como base las preguntas hechas a los evaluados durante el ejercicio. El análisis consistió en un proceso iterativo de identificación de categorías dentro de las preguntas, a través de las respuestas y observaciones de los administradores. Como resultado encontraron que 10 de los 20 ítems de matemáticas y ciencias, nueve de selección múltiple y once de respuesta corta abierta, en los que los expertos bilingües habían identificado fuentes de DIF, fueron confirmados por el ejercicio de pensar en voz alta. De acuerdo con ello, los autores concluyen que la revisión por parte de expertos no puede ser considerada como evidencia suficiente para decidir si un ítem con DIF está sesgado y que dichas opiniones necesitan tener en cuenta los procesos cognitivos de los evaluados.

Adicionalmente, Ercikan et al. (2010) señalan que las revisiones por expertos tienen dos limitaciones: a) determinadas fuentes de DIF identificadas por los expertos son reales de forma tentativa hasta que evidencia adicional muestre que efectivamente son fuentes de DIF, y b) los revisores no responden la pregunta de por qué una característica de un ítem hace que presente una alta dificultad o discriminación para un grupo u otro. Sin embargo debe tenerse en cuenta que la representatividad de la muestra en dicho estudio y el tamaño de la misma (36 evaluados de un grupo y 12 del otro), puede mostrar diferencias respecto a los análisis de DIF.

Independientemente de cuál sea el método que se elija o cuál sea el más adecuado, la identificación de las fuentes de DIF es un aspecto crucial para la interpretación del mismo, pero es un reto teniendo en cuenta los múltiples factores que afectan la equivalencia de los ítems (Ercikan et al., 2010).

Sesgo Cultural

Solano & Nelson (2001) plantean el término validez cultural en el contexto de las pruebas de ciencias, la cual se define como la efectividad con la que dichas evaluaciones “abordan las influencias socioculturales que forman el pensamiento del estudiante y la forma en la cual los estudiantes hallan el sentido a los ítems de ciencias y los responden” (p. 555). Una forma de enfrentar las influencias socioculturales en las pruebas es identificar si existen potenciales fuentes de DIF, es decir, sesgo en los ítems. Es bien conocido que factores como la lengua materna de los evaluados, las experiencias de vida y las prácticas de socialización pueden originar agudas interacciones entre los antecedentes de los individuos y el contenido de una prueba (Ross & Okabe, 2006).

Para Banks (2006) los ítems sesgados culturalmente tienen características que no están relacionadas con el desempeño en el constructo que está siendo medido, sino que son sensibles a grupos culturales particulares y afectan su ejecución. En el caso de pruebas de selección múltiple, ciertos grupos culturales pueden interpretar el ítem o las opciones de respuesta en formas que no fueron anticipadas durante la construcción de la prueba y, de acuerdo con ello, ser atraídos por distractores que contienen estímulos culturales específicos (Veale & Foreman, 1983). Los investigadores frecuentemente rechazan las fuentes de sesgo cultural porque son difíciles de investigar en términos metodológicos y su poder inferencial está fuertemente restringido por otros factores de confusión (Wu & Ercikan, 2006), es decir que, los sesgos culturales pueden estar relacionados con otras variables.

A pesar de la necesidad de que las pruebas educativas no estén sesgadas culturalmente, no hay una definición explícita de cultura en este contexto, según Banks (2006), esto no es tarea fácil. Dicha autora plantea que la cultura es un “sistema interrelacionado de ideas,

valores, símbolos, productos, comportamientos y cosas similares, las cuales todas funcionan interdependientemente” (p. 116). En este sentido, el hecho de que los grupos culturales difieran en su desarrollo y el uso de ciertos aspectos culturales (Naylor, 1997; Thomas, 1999, citados por Banks, 2006), y la forma cómo cada grupo experimenta esas diferencias en su vida diaria, pueden explicar su disparidad en la ejecución en las pruebas. De acuerdo con esta propuesta, Banks (2006) ofrece como ejemplo el hecho de que cuando se responde a un ítem de selección múltiple, grupos culturales distintos pueden prestar atención a estímulos diferentes, es decir, es posible que si se incluye una opción que ilustre un aspecto de la cultura de un grupo, los miembros de ésta tiendan a preferirla sin importar si es la respuesta correcta o incorrecta.

El estudio en el cual se enmarcan la anterior definición de cultura fue desarrollado por Banks (2006) con el fin de proponer una estructura para evaluar hipótesis sobre sesgo cultural en pruebas educativas. Básicamente trabajó con la prueba de Terra Nova (CTB/McGraw-Hill, 1999) para ilustrar aspectos de las culturas hispanas o afrodescendientes que pudieran mediar en la elección de opciones de respuestas correctas e incorrectas. A partir de una revisión de descripciones de estas culturas, encontró cuatro aspectos de la cultura hispanoamericana que podrían ser factores de sesgo: familisimo, presentisimo, personalisimo y espiritismo; y cuatro de la cultura afroamericana: comunalismo, perspectiva social del tiempo, oralidad y armonía.

El **familisimo** “enfatisa el valor de poner las necesidades familiares antes de los intereses personales. Los valores familiares acentúan la codependencia o interdependencia, la cooperación, el compartir la solución a los problemas y recursos materiales” (pág. 118). Un posible ejemplo de la manifestación del familisimo es expuesto por Banks en el siguiente ítem (p. 119), el cual proviene de la prueba de lectura para el grado noveno de la Evaluación de conocimientos y habilidades de Texas:

“¿Cuál es el tema general expresado en el artículo²?”

- a) Las relaciones entre hermanos son usualmente más fuertes que entre hermanas.
- b) Centrarse en metas profesionales en lugar de la familia conduce a la decepción.
- c) Para tener éxito, un matrimonio debe basarse en la honestidad.
- d) La felicidad se encuentra en estar satisfecho con los logros personales.”

La opción b de este ítem describe lo que puede suceder si no se ponen las necesidades familiares antes de los intereses personales, por lo que allí estaría expresado el familisimo. La opción correcta es la c.

El **presentisimo** “enfatisa la importancia del aquí y ahora. Los valores se centran en las necesidades actuales y en alcanzar metas a corto plazo antes que embarcarse en temas más amplios” (pág. 118). Por su parte, el **espiritismo** se caracteriza por “las interconexiones entre los seres humanos y lo natural y súper natural. Más que controlar ciertos aspectos del ambiente, se cree que los individuos están pensados para estar en armonía con ellos” (pág. 118). El personalísimo corresponde a la noción de la buena comunicación interpersonal, en donde ésta implica compartir sentimientos, aportarle al otro y validar el valor propio del otro. Un ejemplo de personalisimo es ilustrado por Banks (p. 119) en el siguiente ítem tomado de la prueba de lectura para el grado once del grupo 1 del American College Testing:

“Los detalles y eventos en el pasaje sugieren que la amistad entre el narrador y la señora Senett sería más exactamente descrita como:

- a) estimulante, marcada por un amor compartido por aventuras excéntricas.
- b) indiferente, marcada por la insensibilidad ocasional a las necesidades del otro.
- c) considerada, que se destaca por el intercambio de favores entre amigos.

² El ítem tiene como contexto un artículo.

- d) emocional, basada en el compromiso de los amigos que desean compartir sus cargas con otro.

Este ítem puede ser catalogado como **personalísimo** porque las opciones c y d describen personas intercambiando favores y sentimientos. La opción correcta es la c.

El **comunalismo** corresponde a la interconexión entre las personas y “acentúa la importancia del deber de cada persona hacia un grupo social, la cooperación, la interdependencia, la cohesión e identidad de grupo, la construcción de estrategias colaborativas y la puesta en común de los recursos” (pág. 119). El siguiente ítem es tomado por Banks de la prueba de Inglés/Lenguaje del *Arts Indiana State Wide Testing for Educational Progress* y puede ser clasificado como comunalismo puesto que las opciones c y d describen el deber hacia un grupo social y el mantener lazos sociales:

“¿Cuál de los siguientes es un tema en la historia?

- a) superar el peligro
- b) evitar la soledad
- c) mostrar lealtad
- d) mantener la amistad”

La **perspectiva social del tiempo** se manifiesta en la “noción de que el tiempo debe ser considerado en términos de las actividades sociales en las que se está comprometido, más que en el sentido estricto de cumplir con horarios rígidos. El énfasis está centrado en las tradiciones y costumbres del pasado que guían eventos futuros” (pág. 119).

La **oralidad** se refiere al valor que se otorga al “conocimiento que es obtenido o transmitido a través de medios de comunicación oral. Las formas de comunicación oral son consideradas más apropiadas porque se piensa que son mucho mejores para transmitir significados que las palabras escritas” (pág. 119). Un ejemplo de oralidad es el siguiente

ítem³ tomado por Banks de la prueba de lenguaje para el grado cuarto del *Arts Wisconsin Knowledge and Concepts Examinations* (pág. 121), en el cual la opción c representa el dialecto cultural de los afroamericanos porque la omisión de la palabra “is” está permitida al hablar en presente continuo en ese dialecto:

“Which sentence is complete and written correctly?”

- a) Molding a pinch pot out of clay.
- b) Work on a good surface, the pot can be made more easily.
- c) The moist clay drying in the air.
- d) After making a pinch pot, try making a coil pot, too.”

Finalmente, la **armonía** acentúa “la importancia de dar homenaje a un ser supremo que controla todos los aspectos del mundo. Hay una creencia de que los humanos y lo natural y lo súper natural están interconectados. Más que controlar ciertos aspectos del medio ambiente, los individuos son pensados para ajustarse al universo adoptando creencias metafísicas tales como la suerte o el destino para explicar las circunstancias de la vida” (pág. 119).

Banks (2006) encontró en su estudio que al analizar en conjunto las respuestas correctas no mostraron resultados significativos, pero al analizar los distractores si se encontraron resultados importantes. Los hallazgos mostraron que cinco de los ítems clasificados como posiblemente sesgados por oralidad y tres de los clasificados por comunalismo fueron respondidos diferencialmente entre blancos y afroamericanos. En dichos ítems, las dos culturas mostraron odds⁴ más grandes en relación con la elección de

³ Este ítem no fue traducido al español pues se consideró que al hacerlo se podría perder el sentido de ejemplificar la oralidad.

⁴ Los odds se definen como una proporción o probabilidad dividida por su complemento. En el contexto del DIF, se define como qué tantas veces es más probable que un grupo responda correctamente que incorrectamente una pregunta ($\Omega = \frac{p}{q}$) (Camilli & Shepard, 1994). En el estudio de Banks (2006) hace referencia a qué tantas veces es más probable que un individuo seleccione determinado distractor que está relacionado con su cultura a que no lo haga.

los distractores que se consideraban directamente relacionados con su cultura, lo que lleva a la autora a sugerir que no sólo se debe realizar análisis de DIF para las opciones de respuesta correcta, sino también para los distractores.

Otro estudio sobre fuentes culturales de DIF en las pruebas educativas es el de Wu & Ercikan (2006), quienes tenían como objetivo mostrar la utilidad del método de comparación de múltiples variables usando regresión logística para identificar fuentes de DIF. Como ejemplo, los autores utilizaron el Tercer Estudio Internacional de Matemáticas y Ciencias (TIMSS) llevado a cabo en 1999, tomando como grupos los evaluados de Taiwán y Estados Unidos, siendo el primero el grupo de referencia y el segundo el grupo focal. Los autores tenían como hipótesis que los examinados de los países del Asia del Este tenían mayor probabilidad de responder correctamente los ítems, debido a horas extras de lecciones después de la escuela, fenómeno que se hace más evidente en dichas culturas con el pasar de los años, porque se presume que la educación pública no es muy adecuada y a la presencia de competitividad frente a los cupos para la educación superior. De acuerdo con los autores, el estudio señala que después de haber dividido el examen en diferentes áreas de contenido, las horas extras de enseñanza mostraron ser una potencial fuente de DIF, sobre todo en las áreas de álgebra, sentido de fracciones y números, medición y geometría. Sin embargo, ésta parece ser más una variable que introduce diferencias válidas entre los países que una fuente de DIF que indique la existencia de sesgo, por lo que podría haber una aparente confusión entre sesgo e impacto en este estudio.

Una investigación más reciente relacionada con sesgo cultural en el contexto de las pruebas internacionales es la de da Costa & Araújo (s.f.), en la cual se realizó un análisis de DIF tomando como grupos a estudiantes nativos europeos (referencia, $n = 182122$) y a estudiantes inmigrantes (focal, $n = 15763$). Las autoras manifiestan que las poblaciones inmigrantes son ampliamente diversas en términos de los países de origen y que llevan su capital cultural al país al cual llegan. Dicho capital humano es transmitido de generación en generación, así como sus creencias y actitudes, estatus ocupacional de los padres y nivel educativo (Leibig & Widmaier, 2009, citado por da Costa & Araújo, s.f.). Se analizaron 95 ítems de la prueba de lectura en 29 países europeos, los cuales correspondían a ítems de

selección múltiple y respuesta construida, enfocándose en si el DIF podría estar asociado con el formato del ítem, el tipo de texto, aspecto evaluado y situación. El análisis de DIF fue realizado en términos del parámetro de dificultad, encontrándose diferencias entre nativos e inmigrantes y que hay ítems más fáciles para un grupo y otros para otro. Respecto al formato del ítem se encontró que el número más alto de ítems con DIF se presentó en los de selección múltiple. En cuanto al aspecto evaluado, los nativos se desempeñaron mejor en “integrar e interpretar” y “reflexionar y evaluar” mientras que los ítems de “acceso y recobro” fueron más fáciles para los inmigrantes. Con relación al formato del texto, el número más alto de ítems más fáciles para inmigrantes correspondió al formato continuo y no continuo, mientras que el formato mixto y múltiple favoreció a los nativos. Respecto al tipo de texto, los nativos se desempeñaron mejor en los textos descriptivos mientras que los textos expositivos, de instrucción y narración, fueron más fáciles para los inmigrantes. Finalmente, frente a la situación, los estudiantes inmigrantes se desempeñaron mejor en ítems relacionados con situaciones educativas y ocupacionales, mientras que los nativos en ítems relacionados con dominios personales y públicos, lo que según los autores podría estar asociado con el propósito de la lectura, por ejemplo, recreativo en el caso de los nativos. En términos generales, los autores señalan que los inmigrantes están más relacionados con aspectos de la lectura que son típicamente vistos en la escuela o en los libros de texto y con el “leer para aprender”.

Un trabajo aplicado al análisis de DIF y posibles fuentes culturales del mismo en matemáticas es el de Yildirim (2006). En este estudio se pretendía evaluar la equivalencia de los ítems de matemáticas de TIMSS 1999 y el Programa para la Evaluación Internacional de Estudiantes (PISA, por sus siglas en inglés) tomando las versiones turca e inglesa, con el fin de averiguar si el desempeño en matemáticas presenta aspectos culturales específicos. Para el análisis de DIF, el autor usó el análisis factorial restringido, el MH y la razón de verosimilitud con TRI, presentándose un alta tasa de acuerdo entre ellas para PISA, aunque mucho menor para TIMSS, probablemente debido a que para este último programa de evaluación la proporción de ítems con DIF fue más alta y se presentaron diferencias en la discriminación y el pseudo azar. Los resultados indicaron que “los ítems

que requerían competencias de reproducción de conocimiento práctico, conocimiento de hechos, ejecución de procedimientos de rutinas y aplicación de habilidades técnicas tuvieron menos probabilidad de estar sesgados en contra de los estudiantes Turcos” (pág. V). En contraste los ítems que exigían a los estudiantes comunicarse matemáticamente, comparar resultados o que presentaban contextos del mundo real tuvieron menos probabilidad de favorecer a los estudiantes turcos.

Otras variables que se han considerado como posibles fuentes de sesgo cultural o que influyen en las respuestas de los estudiantes en las evaluaciones son: el manejo de objetos poco conocidos en una cultura (Tellegen & Laros, 2004); forma en que se toman las decisiones (Rhodes, 1988, citado por Snetzler & Qualls, 2000); experiencias, creencias, interferencia fonológica, ortográfica o semántica, estructura pragmática, convenciones textuales y género de los textos (Luykx, Lee, Mahotiere, Lester, Hart & Deaktor, 2007); y la epistemología del estudiante y estilos de aprendizaje (Solano y Nelson, 2001).

Respecto a la primera, de acuerdo con Tellegen & Laros (2004) un ítem puede considerarse sesgado si uno de los grupos muestra problemas con el reconocimiento de una determinada foto o gráfico mientras que el otro grupo no. También puede considerarse sesgado si se usan diseños viejos de los objetos o que no tienen un uso amplio, el uso de figuras que son difíciles de reconocer u objetos que presentan diseños ampliamente diferentes entre las dos culturas, por ejemplo un cronómetro, un termómetro, una bañera, etc.

En cuanto a la forma en que se toman las decisiones, en algunas investigaciones realizadas con nativos norteamericanos (Rhodes, 1988, citado por Snetzler & Qualls, 2000) se ha encontrado que pruebas que implican respuestas rápidas, adivinación, tomar riesgos y la eliminación de opciones en preguntas de selección múltiple contradicen el proceso de toma de decisiones de estas personas. Esto se debe a que en su cultura las decisiones se toman de forma lenta y segura y están basadas en combinar elementos a partir de diferentes opciones posibles en una respuesta apropiada.

De acuerdo con Luykx et al. (2007) las experiencias particulares han sido consideradas como aspectos que influyen en las respuestas de los estudiantes en las evaluaciones de ciencias. Cuando los evaluados se enfrentan a preguntas en las que aparentemente no tienen el conocimiento suficiente para responder, suelen acudir a sus creencias culturales o experiencias en sus hogares. También existen aspectos fonológicos u ortográficos de la primera lengua de los evaluados que causan interferencia en sus respuestas si están presentando una evaluación en un idioma que no es el materno. En el caso de la semántica, algunos estudiantes pueden interpretar términos científicos de acuerdo con el significado que le dan en su vida diaria.

Por otra parte para estos últimos investigadores, la estructura pragmática puede ser un factor cultural importante al responder los ítems, específicamente se refiere a que los evaluados “necesitan reconocer la estructura pragmática bajo la cual los ítems deben ser interpretados en lugar de simplemente tomar las palabras por su valor aparente” (p. 914), lo que implica que deban ‘decodificar’ la intención de los ítems. Las convenciones textuales de los ítems y gráficas también pueden interferir con las respuestas de los evaluados y no estar relacionadas con lo que se quiere medir. Igualmente, pueden interferir aspectos como textos en negrita, la agrupación jerárquica de las preguntas relacionadas o el formato de las mismas. El género de los textos puede llevar a que no se brinde la respuesta esperada a una pregunta, por ejemplo, si se pregunta por la conclusión de un problema basado en una historia, la respuesta que brinda el evaluado puede ser simplemente cómo terminaría la historia mas no la solución al problema planteado, es decir, el texto es de tipo científico, pero es considerado por el evaluado como literario.

Para Solano & Nelson (2001), quienes citan a Delpit (1998), la forma en que el conocimiento es valorado, cómo es pensado y aprendido y lo que es exitoso dentro de un sistema escolar puede ser resultado de los valores, conocimiento y estilos de la población mayoritaria y puede alienar a las minorías culturales, lingüísticas y étnicas. Esto también puede aplicarse a los estilos de aprendizaje, por ejemplo,

“los adultos y los niños pueden compartir la misma actividad por períodos largos de tiempo en el cual hay interacciones verbales pero no están centrados en el proceso de aprendizaje. Este estilo puede no ser óptimo en un sistema escolar occidental tradicional que organiza la enseñanza alrededor de períodos de clase cortos y frecuentes en los cuales se espera que los estudiantes escuchen pasivamente a los profesores, sigan instrucciones y brinden largas respuestas verbales a las preguntas (Lipka, 1998)” (pág. 556).

También Solano & Nelson (2001) afirman que la epistemología propia de los antecedentes culturales de los estudiantes puede influir en la forma en que solucionan los ítems. Esto se evidencia en un ejemplo que Solano y Nelson toman de Greenfield (1997), a partir del cual concluyen que ignorar la epistemología de los estudiantes puede llevar a entender equivocadamente sus habilidades. De acuerdo con ese ejemplo, podría decirse que la forma en la que se realizan las preguntas puede ser contraria o diferente a la forma cómo los evaluados conciben el mundo y el conocimiento. Asimismo, la epistemología del estudiante puede asociarse con las experiencias vividas por el evaluado que hacen que la respuesta que da sea correcta en su contexto cultural, lo que se infiere del ejemplo que los autores toman de Solano & Nelson (2000).

Una fuente de DIF que podría hacer parte de los sesgos culturales, es la que se presenta en las pruebas adaptadas de otros idiomas o cuyo lenguaje de aplicación no corresponde con la lengua materna de los examinados, dado que, como señalan Snetzler & Qualls (2000), hay que reconocer la inseparable asociación entre lenguaje y antecedentes étnico-culturales, cuyos efectos no pueden ser completamente separados. Un estudio de Ross & Okabe (2006) que pretendía comparar la forma “convencional” y subjetiva de encontrar ítems sesgados a partir de un panel de profesores, con la forma objetiva de resultados estadísticos obtenidos a partir de software para detectar DIF, planteó como ejemplo el análisis de una prueba de comprensión de lectura cuyo grupo focal eran estudiantes mujeres y el de referencia hombres que estaban aprendiendo inglés como lengua extranjera. En su revisión de la literatura, Ross & Okabe (2006) mencionaron tres posibles sesgos que podrían introducirse en pruebas sobre lenguas extranjeras según el grupo cultural o país al

cual se perteneciera: transferencia, exposición o práctica y los factores de socialización. La transferencia se refiere al parecido o raíces léxicas comunes que muestran diferentes lenguas, como por ejemplo el inglés y el español (conflagration - conflagración); y a la ortografía similar, aspecto que se da entre japonés y chino, razón por la cual a un chino le resulta más fácil aprender japonés que a un norteamericano. La exposición o práctica se presenta como sesgo, por ejemplo, cuando hay hogares bilingües y los factores o patrones de socialización involucran la orientación académica establecida en la educación de los estudiantes (p. ej. ciencias o humanidades) (Pae, 2004, citado por Ross & Okabe, 2006) que puede corresponder con las prácticas de socialización por género. Aunque el objetivo de la investigación de Ross & Okabe (2006) no era la identificación de ítems sesgados a nivel de lenguaje, debe mencionarse que como conclusión a nivel metodológico encontraron que la forma subjetiva de identificarlos tendía a sobreestimar la cantidad de ítems que en realidad presentaban DIF y que eran detectados por medios objetivos.

Otro estudio desarrollado en el contexto de pruebas internacionales y los sesgos provenientes del lenguaje es el de Elosua (2006) quien, aunque detectó ítems con DIF, al comparar evaluados españoles e ingleses, no pudo identificar la fuente del mismo. Sin embargo, en otra investigación desarrollada por dicha autora y algunos colaboradores (Elosua et. al, 2000), se encontró que el idioma podría ser una fuente de sesgo, cuando se administra la prueba en un lenguaje que no es el de escolarización. Otras investigaciones mencionan factores de sesgo en el lenguaje adicionales a los ya descritos como: forma en que están escritos los ítems (Price & Oshima, 1997) y el tomar en cuenta los antecedentes de la cultura a la que se pertenece en el momento de construir los ítems (Al-Fallay, 1999).

Finalmente, respecto a aspectos relacionados con el lenguaje Ercikan et al. (2010) afirman que se ha encontrado que los Aprendices de Lengua Inglesa (ELLs por sus siglas en inglés) que toman pruebas en una segunda lengua obtienen desempeños más altos en ítems menos complejos lingüísticamente y que la complejidad lingüística innecesaria es un problema muy importante que afecta la comparabilidad entre ELLs y no ELLs (Abedi, Lord, & Plummer, 1995 y Abedi, Leon, Wolf, & Farnsworth, 2008, citados por Ercikan et al., 2010).

Como puede observarse son muchos los estudios y contextos variados en los que se ha estudiado el sesgo cultural o aspectos relacionados. Todos ellos enmarcan la presente investigación con miras a confirmar las fuentes reportadas en la literatura o a sumar unas nuevas y propias del contexto colombiano. Aquí puede verse además que buena parte de los estudios se centran en el lenguaje o diferencias idiomáticas que de una u otra forma representan elementos culturales, si se tiene en cuenta lo señalado por Snetzler & Qualls (2000), puesto que existe una clara asociación entre lenguaje y antecedentes culturales o étnicos, por lo que las potenciales fuentes de DIF en el lenguaje también pueden considerarse como fuentes de sesgo cultural.

Examen de Estado de la Educación Media ICFES - SABER 11°

Teniendo en cuenta el documento sobre los antecedentes y marco legal del Examen de Estado, elaborado por el Grupo de Evaluación de la Educación Básica y Media del ICFES en 1999, el Examen surge cuando la Asociación Colombiana de Universidades y el Fondo Universitario firman el Acuerdo No. 65 de 1966 a partir del cual se crea el Servicio de Admisión Universitaria y Orientación Profesional. En septiembre de 1968 se aplican los primeros Exámenes Nacionales, siendo pruebas unificadas para toda la población y calificadas a partir de escalas nacionales, y el Servicio mencionado anteriormente fue reestructurado y convertido en el Servicio Nacional de Pruebas (SNP), dependencia del ICFES.

En 1980, con el decreto 2343 (República de Colombia, 1980) se reglamentan los exámenes de Estado para el ingreso a la educación superior y se declaran como pruebas académicas de cobertura nacional, de carácter oficial y obligatorio, cuyo objetivo era comprobar los niveles mínimos de aptitudes y conocimientos de quienes deseaban ingresar a las IES. Igualmente, se pretendía que este examen fuera un punto de referencia para dichas instituciones para la admisión de estudiantes, y que sus resultados fueran ponderados por factores socio-culturales con miras a determinar puntajes de admisión mínimos para adquirir la calidad de estudiante mediante la matrícula (artículo 169).

Posteriormente, a través del artículo 14 de la Ley 30 de 1992 (República de Colombia, 1992), la cual organiza el servicio público de la Educación Superior, se reafirma el Examen de Estado como requisito para ingresar a ese nivel educativo. Así mismo, se establece que el examen tienen por objeto: “a) Comprobar niveles mínimos de aptitudes y conocimientos, b) Verificar conocimientos y destrezas para la expedición de títulos a los egresados de programas cuya aprobación no esté vigente⁵, c) Expedir certificación sobre aprobación o desaprobarción de cursos que se hayan adelantado en instituciones en disolución cuya personería jurídica ha sido suspendida o cancelada, y d) Homologar y convalidar títulos de estudios de Educación Superior realizados en el exterior cuando sea pertinente a juicio del Consejo Nacional para la Educación Superior (CESU).” (pág. 10). En 1994, con el artículo 99 de la Ley 115 (Ley general de educación) se reglamenta su uso para la asignación de becas (República de Colombia, 1994).

Con la Ley 1324 de 2009 (República de Colombia, 2009) “se fijan parámetros y criterios para organizar el sistema de evaluación de resultados de la calidad de la educación, se dictan normas para el fomento de una cultura de la evaluación, en procura de facilitar la inspección y vigilancia del Estado y se transforma el ICFES” (p. 1). En esta Ley se afirma que la evaluación que se efectúa a través de los exámenes de Estado se realizará bajo los principios de independencia, igualdad, comparabilidad, periodicidad, reserva individual, pertinencia y relevancia; donde el principio de igualdad se refiere a garantizar a todos los entes evaluados la misma protección y trato al practicar la evaluación y al producir y dar a conocer sus resultados. A partir de esta Ley el examen de Estado para la educación media también puede ser utilizado como forma de acreditar – validar – que se han obtenido los conocimientos y competencias que se esperan de las personas que terminan el nivel de educación media. Adicionalmente, las instituciones se ven en la obligatoriedad de presentar a dichas evaluaciones a todos los estudiantes, salvo circunstancias excepcionales. Actualmente, el examen de Estado de la educación media ICFES – SABER 11° está reglamentado por el decreto 869 de 2010 (República de Colombia, 2010).

⁵ Literal declarado exequible por la Corte Constitucional a través del proceso D-4715.

El examen de Estado aplicado durante las décadas del 80 y 90, estaba compuesto por nueve pruebas agrupadas en 5 áreas: Ciencias Naturales, Matemáticas, Ciencias Sociales y una prueba electiva (Grupo de Evaluación de la Educación Básica y Media, 1999), teniendo como núcleo básico las pruebas la aptitud verbal y matemática, posicionadas a partir de los años 40 por los modelos psicométricos que consideraban que la aptitud estaba ligada al logro educativo (Torrado, 1998).

Posteriormente, sufre una reforma en el contexto de la renovación de los propósitos educativos generada a partir de la Constitución Política de 1991 y la Ley General de Educación, las recomendaciones de la Misión para la Modernización de la Universidad Pública y la Misión de Ciencia, Educación y Desarrollo, los cambios producidos en el contexto mundial de las áreas que conforman el examen y los nuevos modelos psicométricos para la medición y evaluación educativa, las exigencias surgidas a partir de la globalización, el trabajo interno del ICFES, las pruebas similares aplicadas en el ámbito internacional y la investigación desarrollada por el ICFES desde 1991 como parte de la Evaluación de la Calidad de la Educación (Grupo de Evaluación de la Educación Básica y Media, 1999). El proyecto de reconceptualización del examen de Estado se inicia en 1995, teniendo en cuenta la experiencia de las pruebas SABER encaminadas a la Evaluación de la Educación Básica, lo que se hacía en otros lugares del mundo en evaluación y sus perspectivas epistemológicas, cognitivas y técnicas; generando como resultado una prueba que evalúa competencias antes que información y datos que el estudiante guarda (Grupo de la Evaluación de la Educación Básica y Media, 2005).

De acuerdo con el ICFES (s.f.), la prueba de Estado tenía como propósitos: a) Servir como un criterio para el Ingreso a la Educación Superior, b) Informar a los estudiantes acerca de sus competencias en cada una de las áreas evaluadas, con el ánimo de aportar elementos para la orientación de su opción profesional, c) Apoyar los procesos de autoevaluación y mejoramiento permanente de las instituciones escolares, d) Constituirse en base e instrumento para el desarrollo de investigaciones y estudios de carácter cultural, social y educativo y e) Servir de criterio para otorgar beneficios educativos.

Aunque los propósitos enunciados en el decreto 869 de 2010 no son muy disímiles a los expuestos en el párrafo inmediatamente anterior, se destacan diferencias o aspectos adicionales relacionados directamente con la evaluación de la calidad de la educación como: “Comprobar el grado de desarrollo de las competencias de los estudiantes que están por finalizar el grado undécimo de la educación media” (literal a), “Monitorear la calidad de la educación de los establecimientos educativos del país, con fundamento en los estándares básicos de competencias y los referentes de calidad emitidos por el Ministerio de Educación Nacional.” (literal d), “Proporcionar información para el establecimiento de indicadores de valor agregado, tanto de la educación media como de la educación superior.” (literal e), y “Servir como fuente de información para la construcción de indicadores de calidad de la educación, así como para el ejercicio de la inspección y vigilancia del servicio público educativo.” (literal f). También se mantienen los relacionados con la selección y nivelación académica: “Proporcionar a las instituciones educativas información pertinente sobre las competencias de los aspirantes a ingresar a programas de educación superior, así como sobre las de quienes son admitidos, que sirva como base para el diseño de programas de nivelación académica y prevención de la deserción en este nivel.” (literal b).

A partir del año 2000, el examen de Estado de la educación media está conformado por un núcleo común que mide competencias en las áreas fundamentales de la Educación Básica y Media; y un componente flexible que mide competencias en niveles más complejos (las áreas de profundización) y frente a problemas actuales (interdisciplinar) (ICFES, s.f.). En el componente flexible, el estudiante escoge un área de profundización o un área interdisciplinar. A partir del año 2006 se dejó de evaluar Geografía e Historia y se pasó a evaluar Ciencias Sociales. También se incluyó la prueba de Inglés y se dejó de evaluar dos pruebas en el componente flexible. La estructura de prueba y el tiempo dedicado a cada sesión para el examen aplicado en el primer semestre del 2007 aparece en la figura 4.

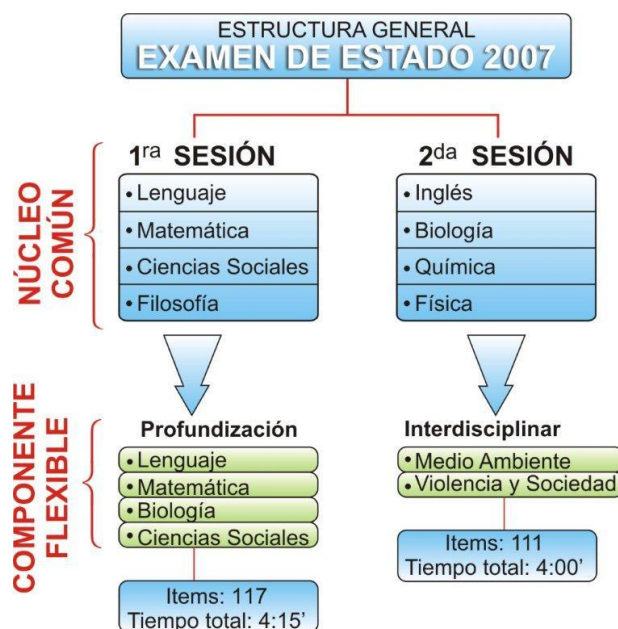


Figura 4. Estructura del Examen de Estado aplicado en el primer Semestre de 2007. P. Pedraza (Comunicación personal, Agosto 22, 2012).

Actualmente, el MEN es la entidad que le indica al ICFES qué es lo que desea evaluar en los exámenes de Estado, lo cual debe ser consultado con el Consejo Nacional de Educación Superior (CESU) (República de Colombia, 2009). Sin embargo, el número de pruebas y componentes son determinados por el ICFES a través de acuerdo de su junta directiva (República de Colombia, 2010). Los ítems analizados en el presente estudio se tomaron de las pruebas de Lenguaje, Matemáticas, Biología y Ciencias Sociales de la segunda aplicación de 2006 y la primera de 2007.

Prueba de Lenguaje.

De acuerdo con ICFES (s.f.b), teniendo en cuenta los estándares básicos de competencias establecidos por el Ministerio de Educación Nacional “la pedagogía de la lengua castellana debe centrarse en el desarrollo de la competencia comunicativa básica de los sujetos: el perfeccionamiento de su capacidad para identificar el contexto comunicativo en el que se encuentran y, en consecuencia, saber cuándo hablar, sobre qué, de qué manera, cómo reconocer las intenciones que subyacen a todo discurso, cómo evidenciar los aspectos conflictivos de la comunicación, en fin, cómo actuar sobre el mundo a partir de la lengua y,

desde luego, del lenguaje.” Los siguientes son las competencias y los componentes del área de Lenguaje.

Competencia interpretativa. “Alude fundamentalmente a la constitución de los diversos sentidos que circulan en los textos. Interpretación que no debe entenderse como “captar el sentido asignado por el autor a un escrito” sino como una acción caracterizada por la participación del lector en su construcción” (ICFES, s.f.b).

Competencia argumentativa. “Es una acción contextualizada que busca explicar las ideas que articulan y dan sentido a un texto. Así, el estudiante (lector) no argumenta desde un discurso previamente elaborado sino en razón de las ideas expuestas en el escrito, las cuales actualizan sus saberes respecto al tema abordado” (ICFES, s.f.b).

Competencia propositiva. Se refiere a “Es una acción fundada en la interpretación. Se caracteriza por ser una actuación crítica que exige la puesta en escena de los saberes del lector, lo cual permite el planteamiento de opciones o alternativas ante las situaciones o problemáticas presentes en un texto. La propuesta o alternativa está sujeta al contexto creado por el texto.” (ICFES, s.f.b).

Componente de la Función semántica de la información local. “Este grupo de preguntas indaga por la función que cumplen los elementos microtextuales y locales en la construcción del sentido del texto.” (ICFES, s.f.b.).

Componente de la Configuración del sentido global del texto. Las preguntas de este componente indagan por “el universo de sentido que cada texto propone. También invitan a la realización de lecturas sintagmáticas y paradigmáticas, con el fin de establecer relaciones entre lo dicho y lo sugerido.” (ICFES, s.f.b.).

Componente del sentido del texto hacia otros textos. “Este grupo de preguntas indaga por lo dicho en el texto en relación con otros textos.” (ICFES, s.f.b).

Prueba de Matemáticas.

De acuerdo con Acevedo, Montañéz, Huertas & Pérez (2007), en la prueba de matemáticas se evalúa el conocimiento matemático del estudiante y los procesos que median en la construcción del pensamiento matemático. Los siguientes son los componentes y competencias del área de Matemáticas.

Competencia el razonamiento y la argumentación. Están relacionados con “aspectos como el dar cuenta del cómo y del porqué de los caminos que se siguen para llegar a conclusiones, justificar estrategias y procedimientos puestos en acción en el tratamiento de situaciones problema, formular hipótesis, hacer conjeturas, explorar ejemplos y contraejemplos, probar y estructurar argumentos, generalizar propiedades y relaciones, identificar patrones y expresarlos matemáticamente y plantear preguntas. Saber qué es una prueba de matemáticas y cómo se diferencia de otros tipos de razonamiento y distinguir y evaluar cadenas de argumentos.” (págs. 23 y 24, Acevedo et al., 2007).

Competencia la comunicación y la representación. La comunicación y la representación “están referidas, entre otros aspectos, a la capacidad del estudiante para expresar ideas, interpretar, usar diferentes tipos de representación, describir relaciones matemáticas, relacionar materiales físicos y diagramas con ideas matemáticas, modelar usando lenguaje escrito, oral, concreto, pictórico, gráfico y algebraico, manipular proposiciones y expresiones que contengan símbolos y fórmulas, utilizar variables y construir argumentaciones orales y escritas, traducir, interpretar y distinguir entre diferentes tipos de representaciones, interpretar lenguaje formal y simbólico y traducir de lenguaje natural al simbólico formal” (pág. 23, Acevedo et al., 2007)

Competencia la modelación y planteamiento y resolución de problemas. “se relaciona, entre otros, con la capacidad para formular problemas a partir de situaciones dentro y fuera de la matemática, traducir la realidad a una estructura matemática, desarrollar y aplicar diferentes estrategias y justificar la elección de métodos e instrumentos para la solución de problemas, justificar la pertinencia de un cálculo exacto o aproximado en la solución de un problema y lo razonable o no de una respuesta obtenida. Verificar e

interpretar resultados a la luz del problema original y generalizar soluciones y estrategias para dar solución a nuevas situaciones problema.” (pág. 23, Acevedo et al., 2007).

Componente Numérico variacional. Este componente “indaga por la comprensión de los números y de la numeración, el significado del número, la estructura del sistema de numeración; el significado de las operaciones, la comprensión de sus propiedades, de su efecto y de las relaciones entre ellas; el uso de los números y las operaciones en la resolución de problemas diversos, el reconocimiento de regularidades y patrones, la identificación de variables, la descripción de fenómenos de cambio y dependencia; conceptos y procedimientos asociados a la variación directa, a la proporcionalidad, a la variación lineal en contextos aritméticos y geométricos, a la variación inversa y al concepto de función.” (pág. 23, Acevedo et al., 2007).

Componente Geométrico-métrico. Este componente se relaciona con “la construcción y manipulación de representaciones de los objetos del espacio, las relaciones entre ellos y sus transformaciones” (págs. 23 y 24, Acevedo et al., 2007).

Componente aleatorio. Este componente “indaga por la representación, lectura e interpretación de datos en contexto; el análisis de diversas formas de representación de información numérica, el análisis cualitativo de regularidades, de tendencias, de tipos de crecimiento, y la formulación de inferencias y argumentos usando medidas de tendencia central y de dispersión y el reconocimiento, descripción y análisis de eventos aleatorios.” (pág. 24, Acevedo et al., 2007).

Prueba de Biología.

De acuerdo con Toro, Reyes y Martínez (2007) “Las competencias específicas en biología permiten establecer relaciones entre diferentes ramas de la biología para entender la vida, los organismos, sus interacciones y sus transformaciones, para este fin se abordan los procesos vitales, la descripción y organización de los seres vivos; la importancia de las especies en los ecosistemas en la medida que modelan y transforman su entorno, las relaciones entre los seres vivos con los ecosistemas en los procesos de intercambio de

energía.” (Pág. 49). Los siguientes son las competencias y los componentes de las pruebas de Ciencias Naturales:

Competencia Identificar. Se refiere a la “capacidad para reconocer y diferenciar fenómenos, representaciones y preguntas pertinentes sobre estos fenómenos” (pág. 18, Toro, Reyes, Martínez, Castelblanco, Cárdenas, Granés y Hernández, et al., 2007).

Competencia Indagar. Se entiende como la “capacidad para plantear preguntas y procedimientos adecuados y para buscar, seleccionar, organizar e interpretar información relevante para dar respuesta a esas preguntas” (pág. 18, Toro et al. 2007).

Competencia Explicar. Se refiere a la “capacidad para construir y comprender argumentos, representaciones o modelos que den razón de fenómenos.” (pág. 18, Toro et al., 2007).

Componente Celular. “La unidad estructural y funcional de todos los seres vivos es la célula, la unidad de vida más sencilla que puede vivir con independencia. Los procesos de todo el organismo son la suma de las funciones coordinadas de sus células constitutivas. Estas unidades celulares varían considerablemente en tamaño, forma y función.” (pág. 52, Toro et al., 2007).

Componente Organísmico. “Este componente hace referencia a la comprensión y el uso de nociones y conceptos relacionados con la composición y el funcionamiento de los organismos, a sus niveles de organización interna, su clasificación, sus controles internos (homeostasis) y a la reproducción como mecanismo para mantener la especie. Involucra el conocimiento de la herencia biológica, las adaptaciones y la evolución de la diversidad de formas vivientes” (pág. 54, Toro et al., 2007).

Componente Ecosistémico. Este componente se refiere a “la organización de grupos de especies, a las relaciones que establecen los organismos con otros organismos, al intercambio que establecen entre ellos, con su ecosistema y con el ambiente en general, al establecimiento y conservación de los ecosistemas. También considera el papel de las

especies en lo que se relaciona con la transformación de los ecosistemas, los ecosistemas del mundo; y los procesos de intercambio de energía. Hace referencia al concepto de evolución aludiendo a sus causas y consecuencias en el nivel ecosistémico.” (pág. 56, Toro et al., 2007).

Prueba de Sociales.

De acuerdo con Ortiz, Ayala, Chaparro, Sarmiento y Restrepo (2007) la prueba de ciencias sociales valora “las competencias (y en ellas son necesarias habilidades y conocimientos teóricos, prácticos, metodológicos y axiológicos) en un área que ofrece posibilidades para la comprensión, confrontación y construcción de significados del mundo social”. (Pág. 9). Las competencias y componentes de la prueba de Ciencias Sociales son los siguientes:

Competencia Interpretar. Esta competencia comprende “Describir, identificar, reconocer, deducir, inducir, clasificar y jerarquizar elementos y factores de distintas estructuras sociales” (pág. 32, Ortiz et al., 2007).

Competencia Argumentar. Esta competencia evalúa “Plantear causas, efectos, razones, juicios, relaciones y explicaciones, en forma contextualizada de diferentes procesos y estructuras sociales” (pág. 32, Ortiz et al., 2007).

Competencia propositiva. Se refiere a la “capacidad predictiva y heurística (dados hechos y tendencias, imaginar resultados posibles). Plantear alternativas, indicar soluciones o posibilidades de acción y de reflexión frente a distintos problemas, situaciones y fenómenos sociales)” (pág. 32, Ortiz et al., 2007).

Componente El espacio, el territorio, el ambiente y la población. Este componente se refiere a “las relaciones entre los actores sociales (población y sujetos) y las condiciones de la acción (espacio y tiempo)” (pág. 39, Ortiz et al., 2007).

Componente El poder, la economía y las organizaciones sociales. Estas preguntas evalúan “las relaciones entre los actores sociales (población y sujetos) y los sistemas de la

acción (poder económico, poder político, sociedad, y familia y comunidad)” (pág. 39, Ortiz et al., 2007).

Componente *El tiempo y las culturas*. Estos ítems evalúan “las relaciones entre los actores sociales (población y sujetos) y las significaciones de la acción (dimensiones de la cultura: científicas, tecnológicas y técnicas; dimensiones estéticas y expresivas; dimensiones éticas o integrativas; y dimensiones trascendentes, filosóficas, religiosas o sapienciales)” (pág. 39, Ortiz et al., 2007).

En cuanto a los análisis estadísticos realizados al examen de Estado, hasta hace algunos años se basaban en la Teoría Clásica de los Tests y actualmente en modelos de Rasch; sin embargo no hay estudios publicados o disponibles para el público sobre posible sesgo en los ítems (Herrera, Gómez, Quintero, Arias, Berrío & Cervantes, 2007).

Grupos indígenas y educación en Colombia

No se tiene conocimiento exacto de cuántos indígenas habían en América a la llegada de Colón, los datos van desde tres millones y medio a cien millones (Uribe, 1993, citada por Rodríguez, 2007). Pero sin importar el número, existían una amplia variedad de culturas que fueron resquebrajadas por la llegada de los europeos. Se estima que la población indígena fue reducida a un 10% y con ello una gran cantidad de saberes y costumbres, lo que se vio aumentado dado el afán de asimilar a sus miembros a otra cultura y minimizando sus posibilidades de expresión (Rodríguez, 2007). En el siglo XVI, los lineamientos españoles buscaban agrupar la mayor cantidad de indígenas para la explotación de la tierra, dando origen de esta forma a los resguardos que terminaron convirtiéndose en una forma de segregación de los indígenas (Rodríguez, 2007). Los resguardos fueron disueltos en el siglo XIX después de la independencia y distribuidos entre las familias, las cuales tenían la potestad de venderlas, aunque la disolución no se hizo efectiva en todo el territorio nacional, si acabó la propiedad tradicional colectiva de los indígenas (Rodríguez, 2007). Debido a los malos tratos a los que fueron sometidos los indígenas, estas comunidades decidieron mantener sus raíces culturales de forma codificada en sus danzas, cantos y ritos (Rodríguez., 2007). Respecto a la representación de los

indígenas en la población nacional, para el siglo XVIII, correspondía a un 15% (Hernández, Salamanca & Ruiz, 2007) y al final del siglo XIX el 25% de los colombianos eran indígenas (Friedemann & Arocha, 1985, citados por Rodríguez, 2007).

Desde 1890 hasta la Constitución Política de 1991, los indígenas tuvieron que someterse a la Ley 089 de noviembre de 1890 por “la cual se determina la manera cómo deben ser gobernados los salvajes que vayan reduciéndose a la vida civilizada” (República de Colombia, 1890). Esta ley reglamentaba que la forma en que debían ser gobernadas las comunidades correspondería a un acuerdo entre el Gobierno y las autoridades eclesiásticas. Con ella se normativizan los cabildos, los cuales serían nombrados por los indígenas de acuerdo con sus costumbres y con potestad, entre otras, para castigar las faltas que cometieran los indígenas. Adicionalmente, fueron considerados como menores de edad para el manejo de las porciones de sus resguardos (artículo 40). En el primer censo del siglo XX, en 1912, que contó con información sobre indígenas, se estimó que la población de estas comunidades era aproximadamente el 6.8% de la población colombiana (Uribe, 1998, citados por Hernández et al., 2007b).

El trabajo de las organizaciones indígenas propició un “proceso de reafirmación cultural y conciencia de identidad que terminó con el reconocimiento del país como pluriétnico y multilingüe” (Hernández et al., 2007a) en la Constitución de 1991. A partir del censo llevado a cabo en el 2005, se estimó que el 3.4% de la población era indígena, ubicándose la mayoría en resguardos, parcialidades o en territorios no delimitados legalmente (Hernández, s.f.). También se encontró que existen en Colombia 87 pueblos indígenas identificados. Aunque el castellano es el idioma oficial de Colombia, las lenguas indígenas son también oficiales en su territorio, de las cuales actualmente se hablan 64 amerindias y diferentes dialectos que se agrupan en 13 familias lingüísticas (Hernández et al., 2007a). En los anexos 1, 2 y 3 se muestra la distribución de la población indígena según etnias, los resguardos indígenas en las zonas territoriales establecidas por el Departamento Administrativo Nacional de Estadísticas (DANE) y por departamentos, y la organización lingüística de las comunidades.

Actualmente, la mayoría de la población indígena se encuentra en el área rural del país, en la selva, la Orinoquía, los Andes, los valles interandinos y el Caribe (Hernández et al., 2007a). De acuerdo con lo que muestra la figura 5, los departamentos con mayor porcentaje de indígenas son: Amazonas, Guainía, Nariño, Vaupés y Vichada, siendo los de la Guajira, Cauca y Nariño los que sumados cuentan con aproximadamente la mitad del país (Hernández et al., 2007a). En algunas cabeceras municipales y en las ciudades vive una minoría de indígenas, que ha venido en aumento como resultado de la migración a las zonas urbanas debidas a los cambios culturales, el agotamiento de las tierras de los resguardos y por el desplazamiento forzado causado por el conflicto interno colombiano (Hernández et al., 2007a).

De acuerdo con Hernández, Salamanca & Ruiz (2007c) los hombres representan el 50.4% de la población indígena y las mujeres el 49.6%, siendo mayor el porcentaje de hombres frente al nacional. En cuanto a la edad, el 40% de la población indígena es menor de 15 años, evidenciando altas tasas de natalidad y mortalidad, esto tiende a ser más marcado en las zonas que no son cabeceras municipales, puesto que en estas últimas se presenta un menor porcentaje en las edades de 0 a 20 años, esto puede reflejar que hay migración de personas adultas jóvenes en edad de trabajar, especialmente mujeres, desde los territorios rurales ocasionando cambios culturales de importancia en la población indígena (Hernández, s.f.). También las comunidades indígenas presentan las más altas tasas de fecundidad frente a la tasa nacional y otros grupos étnicos como los afrocolombianos; la explicación de este fenómeno son los factores directos (como la exposición a las relaciones sexuales y el uso de métodos anticonceptivos) y factores como la educación y la participación en la fuerza de trabajo (Hernández, s.f.).

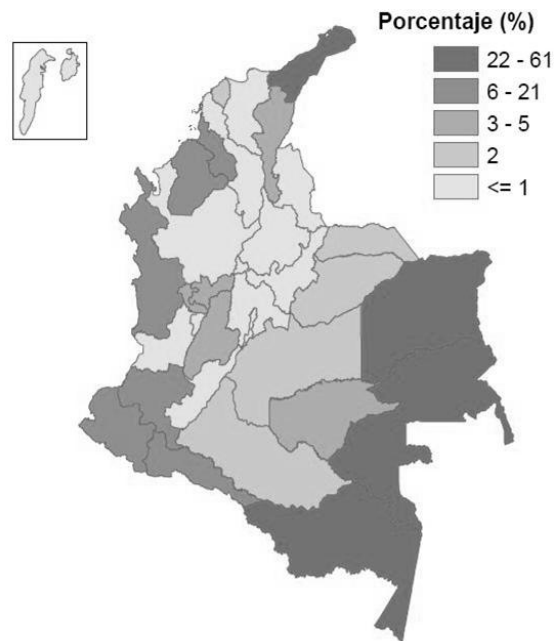


Figura 5. Participación de indígenas respecto al total departamental. Tomado de DANE (2006).

En educación, el 6.7% de los indígenas manifiesta que ha tenido que cambiarse de residencia por esa necesidad, frente a un 3.8% de la población nacional (Hernández et al., 2007c). De hecho los establecimientos educativos en los resguardos indígenas son pocos, atienden el nivel de educación básica primaria y sólo algunos cuentan con programas de etnoeducación que permiten la integración de su tradición, principalmente oral, con los conocimientos de la sociedad restante (Hernández, s.f.). El alfabetismo en los indígenas es del 71.4 %, mientras que el nacional es del 91.6%, adicionalmente se aprecia que hay una diferencia entre hombres (74.1%) y mujeres (68.6%), que no refleja la situación nacional, en la que los dos sexos tienen muy similares tasas (91,3% y 91,8%, respectivamente). También, el alfabetismo es mayor en las cabeceras municipales, pero no se sabe si es en castellano o en la lengua propia de su grupo (Hernández, s.f.).

De acuerdo con Hernández (s.f.) los indígenas presentan las tasa más bajas de asistencia escolar en todos los grupos de edad, seguidos por los afrocolombianos y la población nacional. Respecto al nivel educativo, como lo muestra la figura 6, la primaria es el nivel máximo alcanzado por la mayoría de los grupos, siendo la población indígena la

que presenta el mayor porcentaje en esa categoría y en el de personas que no han estudiado (Hernández, s.f.).

En el contexto educativo, y específicamente para la asignación de cupos para estudiantes indígenas en las universidades, Quintero (2006), de acuerdo con el artículo 2 del Decreto 2164 de 1995, define las comunidades indígenas como el grupo de familias de ascendencia amerindia que “tienen conciencia de identidad y comparten valores, rasgos, usos o costumbres de su cultura, así como forma de gobierno, gestión, control social o sistemas normativos propios que las distinguen de otras comunidades, tengan o no títulos de propiedad o que no pueden acreditarlos legalmente o que sus resguardos fueron disueltos, divididos o declarados vacantes” (p. 2). Por otra parte, los resguardos indígenas actualmente son entidades sociales y políticas regidas por el fuero indígena que se basa en las tradiciones culturales de cada comunidad y se encuentran reglamentadas por el Decreto 2001 de 1998 (Medellín & Fajardo, 2006) y reconocidos constitucionalmente. Los resguardos son una institución legal y sociopolítica especial, conformada por una o varias comunidades indígenas, que poseen un territorio con título de propiedad comunitaria, inembargable e intransferible, regido por una justicia propia (Achury, 2000; Hernández et al., 2007a).

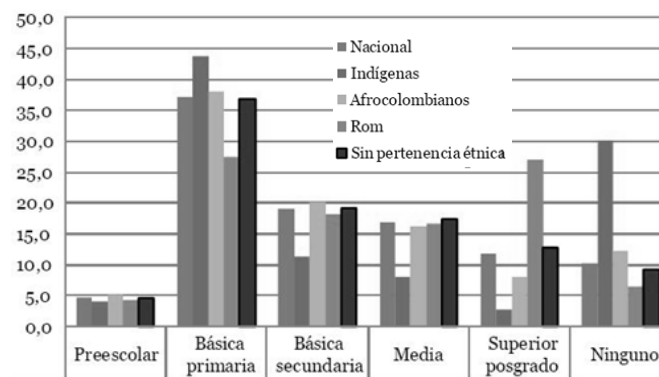


Figura 6. Porcentaje de nivel educativo por etnia. Tomado de Hernández (s.f.) con datos del censo general del DANE en 2005.

La educación en el caso de los indígenas, según Gross (2000, citado por Rodríguez, 2011) consistió en asimilarlos de forma que se favoreció el mestizaje biológico y cultural.

Dicho proceso se dio en tres etapas: la evangelización durante la colonia, la integración durante la República y el reconocimiento de los derechos (Rojas y Castillo, 2005, citados por Rodríguez, 2011). En la colonia la evangelización definió la forma en la que el Estado y las comunidades indígenas se relacionaron y la educación determinó las prácticas que se impusieron en los territorios de dichas poblaciones, la limitación de ciertas expresiones culturales y se originaron “los pasos necesarios para civilizarlos” (pág. 25, Rojas & Castillo, 2005, citados por Rodríguez, 2011). Durante la segunda etapa y desde la propuesta de la Iglesia, la educación ayudó a reforzar modelos espacio-temporales, pautas cognitivas, valoraciones, prohibiciones y formas de producción que resquebrajaron las formas propias de los indígenas y se dejaron de lado aspectos como la lengua, la cultura y las relaciones con las autoridades tradicionales (Rodríguez, 2011). En la última etapa se da el reconocimiento de los derechos étnicos y surge la política de etnoeducación (Rodríguez, 2011).

La Constitución Política de Colombia estableció “el derecho de los grupos étnicos con tradiciones lingüísticas a una educación bilingüe, la participación de las comunidades en la dirección y administración de la educación y el derecho a una formación que respete y desarrolle su identidad cultural” (República de Colombia, 1995, segundo considerando). Sin embargo, desde dos décadas antes (desde 1971) se ha venido trabajando en educación indígena, teniendo como base programas educativos bilingües, dirigidos a la construcción de una educación propia que responda a las características socioculturales de los grupos. Estos procesos se han retomado por el Estado en acuerdo con los pueblos indígenas y se han concretado en la política de etnoeducación vigente (Rojas, 2001). La Ley 115 introdujo cambios en diversos campos, incluyendo la reglamentación de la educación para grupos étnicos, a través de la etnoeducación. Esta política se encuentra enmarcada internacionalmente en la Declaración de la Década Mundial de los pueblos indígenas (1994-2004) hecha por la ONU, la Conferencia Mundial contra el racismo, la xenofobia y la intolerancia y la conmemoración de la esclavitud en Colombia el 21 de mayo de 2001 (MEN, 2001).

De acuerdo con la Ley 115 (República de Colombia, 1994), la etnoeducación se refiere a la educación ofrecida a grupos o comunidades que integran la nacionalidad y que poseen una cultura, una lengua, unas tradiciones y unos fueros propios y autóctonos. Este tipo de educación debe estar ligada al ambiente, el proceso productivo, al proceso social y cultural, respetando las creencias y tradiciones. En los territorios en los que exista tradición lingüística la enseñanza debe ser bilingüe, siendo la lengua materna del grupo el fundamento escolar, sin que vaya en detrimento de lo expuesto artículo 21 de la Ley 115, es decir, del desarrollo de habilidades comunicativas en lengua castellana. Así mismo, en concertación con los grupos indígenas, el Estado debe prestar asesoría especializada en el desarrollo del currículo, la elaboración de libros de texto y materiales educativos y en programas de investigación y capacitación etnolingüística. Finalmente, los educadores que laboren en territorios de grupos étnicos serán seleccionados en concertación con los miembros de los mismos, preferiblemente radicados en la comunidad y deberán acreditar formación en etnoeducación, conocimientos básicos del grupo étnico, su lengua materna y castellano. Los artículos de esta Ley relacionados con etnoeducación fueron el fundamento para la reglamentación de esta forma de educación a través del Decreto 804 de 1995.

También los calendarios educativos para los establecimientos educativos que atiendan grupos étnicos deberán ser definidos de acuerdo con las formas de trabajo, los calendarios ecológicos, las concepciones de espacio y tiempo y las condiciones geográficas y climáticas (República de Colombia, 1995, artículo 17). Esto igualmente aplica para la infraestructura física necesaria para la atención educativa de los grupos étnicos, que adicionalmente debe tener en cuenta los usos y costumbres de las comunidades (República de Colombia, 1995, artículo 19).

De acuerdo con el MEN (2001), la política de etnoeducación pretende contar con una educación acorde con las características, necesidades y aspiraciones de los grupos étnicos que desarrolle la identidad cultural, la interculturalidad y el multilingüismo. Se encuentra dirigida a los grupos indígenas, a las comunidades rom o gitanas y a los afrocolombianos. Así mismo, señala que existe la dificultad de que la etnoeducación no está presente en los currículos, por lo que el MEN en el año de 1999 realizó una premiación a los Proyectos

Educativos Institucionales (PEI) que tuvieran en cuenta la etnoeducación. Entre los proyectos relacionados con grupos indígenas se destacó el PEI “Pensamiento educativo indígena” de las comunidades Yukuna en el departamento del Amazonas, el cual en un calendario ecológico une lo pedagógico con lo comunitario tomando en cuenta aspectos del territorio y agroecológicos y manteniendo los procesos de aprendizaje en el tiempo que los estudiantes no van al colegio sino que se dedican a la caza, pesca y recolección.

A pesar de la reglamentación vigente y descrita anteriormente, autores como Rodríguez (2011) plantean que las organizaciones indígenas han estado trabajando en un sistema de educación propio, aunque el MEN ha mantenido la propuesta de la etnoeducación, la cual, según la mencionada autora, está en discordancia conceptual y metodológica con la política educativa para la Nación y es similar a las normativas que la precedían.

MÉTODO

Para el cumplimiento del objetivo del presente estudio se diseñó una estrategia metodológica compuesta por el uso de métodos cuantitativos y cualitativos en dos fases. Este tipo de investigaciones suelen conocerse como de enfoque mixto porque combinan estrategias cuantitativas y cualitativas. En este caso, la primera buscaba contar con un criterio empírico de posibles diferencias en el comportamiento psicométrico de los ítems y, la segunda, explicaciones a ese comportamiento a través de sesiones de análisis de sesgo que comprendió la revisión de ítems y grupos focales con expertos.

Población y Muestra.

La población correspondió a las personas que presentaron el examen de Estado SABER 11° en el segundo semestre de 2006 y en el primer semestre de 2007, el cual fue aplicado en dos sesiones en ambas aplicaciones. Sin embargo, para la composición de las muestras se excluyeron aquellos registros correspondientes a estudiantes que hubieran dado respuestas no válidas (omisión o multimarca) en más del 10% de respuestas en cada prueba, por lo que el número de examinados no es el mismo en todas ellas. Además para la composición de los grupos focal y de referencia se adoptaron diferentes estrategias de validación de la información.

Dado que el dato que reportan los evaluados en SABER 11° sobre la etnia a la que pertenecen, puede no ser totalmente confiable fue necesario acudir a criterios externos para validar dicha información. Para esta labor se acudió a tres diferentes fuentes: los datos registrados por los estudiantes en el formulario y dos bases de datos del MEN, una con información sobre matrícula de los estudiantes de grado 11 y ciclo seis⁶ en el año 2006 que eran reportados como pertenecientes a alguna etnia y otra sobre los establecimientos educativos (EE) que habían reportado al MEN tener estudiantes pertenecientes a una etnia

⁶ El ciclo seis hace parte de la educación por ciclos para jóvenes y adultos que se ofrece bajo algunos modelos de educación flexible y equivale al grado 11 (MEN, s.f.).

particular en el año 2007. Debe mencionarse que las bases del MEN no contaban con información de los departamentos de Atlántico y Bolívar, por lo que no hubo evaluados clasificados como indígenas para estos departamentos. Sin embargo, esto no es preocupante puesto que de acuerdo con Hernández et al. (2007a), no existen resguardos indígenas ubicados en dichos departamentos y sólo una etnia se encuentra en el departamento del Atlántico.

Para la selección de la muestra del 2006 se tomó el número de identificación de los estudiantes que aparecían en el archivo de inscripción al examen SABER 11° y se cruzó con el mismo número que aparecía en el archivo de matrícula para este año teniendo en cuenta que fueran reportados como pertenecientes a una etnia en particular, de allí concordaron 1.061 estudiantes que se tomaron como indígenas. Con base en los mismos datos de matrícula se escogieron establecimientos educativos (EE) que reportaron tener por lo menos un estudiante de alguna etnia, y se observó cuáles estudiantes que habían reportado pertenecer a alguna de ellas en el formulario de inscripción al examen estudiaban en dichos EE. Quienes cumplieron este último criterio también fueron considerados como indígenas.

Teniendo en cuenta los anteriores criterios, la muestra para la aplicación del segundo semestre de 2006 estuvo constituida por 410.548 evaluados después de la depuración, de los cuales 3.319 fueron clasificados como indígenas y 407.229 como no indígenas. El anexo 4 muestra el porcentaje de estudiantes por cada etnia que componen el grupo de indígenas, si el estudiante había sido reportado en la base de matrícula como indígena se clasificó bajo la etnia con la que allí aparecía, si provenía del cumplimiento del segundo criterio se clasificó bajo la etnia que había sido reportada en el formulario de inscripción.

Para la muestra de examinados en el 2007, la validación del dato de pertenencia a una etnia se hizo teniendo en cuenta las bases que contenían los EE que habían reportado tener estudiantes de alguna etnia en particular en el año 2006 y 2007. Si un evaluado señalaba en el formulario de inscripción pertenecer a alguna etnia, se contrastaba si su EE se encontraba

en dicha base. Si éste era el caso, el estudiante era clasificado como indígena y en la etnia que había reportado en el formulario.

El anexo 5 muestra el porcentaje de estudiantes por cada etnia que componen el grupo de indígenas. De acuerdo con los anteriores criterios, la muestra quedó constituida por 67.703 estudiantes, 759 indígenas y 66.944 no indígenas.

Instrumentos.

Los instrumentos para la elaboración del presente trabajo fueron las pruebas de Lenguaje, Matemáticas, Biología, Física, Química, Filosofía y Ciencias Sociales del Examen de Estado SABER 11° aplicado en el segundo semestre de 2006 y el primero de 2007. Todas las pruebas cuentan con 24 ítems, a excepción de Ciencias Sociales que tiene 30. Se utilizaron programas especializados para los análisis de los datos y la aplicación de la rutina estadística de detección de DIF, que en este caso fue el R-2.12.1 (2011) y el Winsteps 3.63.0 (2006). El R fue usado para realizar todos los análisis, a excepción de la Diferencia de la Dificultad que fue ejecutado con el Winsteps.

También se desarrollaron protocolos para las sesiones del análisis de sesgo tomando en cuenta diversas referencias (Kramer, 2009; Nagle y Williams, 2012; y Snijkers, 2002). Estos protocolos iban desde las instrucciones para las sesiones de preparación con los participantes, hasta el desarrollo final de la sesión, con el fin de que pudieran ser usados posteriormente para otras investigaciones. También incluían los formatos de acuerdo de confidencialidad, un cuestionario sobre los ítems de las pruebas, definiciones de fuentes de sesgo o factores culturales potenciales fuentes de sesgo encontrados en la literatura, y un formato para registrar las respuestas de cada uno de los participantes a las preguntas sobre los ítems que eran dicotómicas.

El protocolo general puede consultarse en el anexo 6 y el acuerdo de confidencialidad en el anexo 7. Las preguntas incluidas en el cuestionario para la revisión individual sobre los ítems estaban relacionadas con si el ítem favorecía a algún grupo particular (pregunta 1), a cuál (pregunta 2), cuáles aspectos del ítem hacían que se favoreciera a un grupo

(pregunta 3) y cómo se podría mejorar el ítem para que fuera justo para ambos grupos (pregunta 4). Este formato puede ser consultado en el anexo 8.

Las definiciones sobre fuentes de sesgo derivadas de la literatura se elaboraron a partir de la revisión bibliográfica de esta investigación y pretendió abarcar aquellas que estaban mejor descritas en los estudios, éstas pueden ser consultadas en el anexo 9. El formato para registrar las respuestas a las preguntas dicotómicas sobre los ítems de la revisión individual se encuentran en el anexo 10 e incluían casillas para registrar cuántos acuerdos se presentaban respecto a si existía favorabilidad o no, el grupo al que favorecía el ítem y el lugar de éste en la discusión en el grupo focal.

Adicionalmente se usó una grabadora de voz para tener registro de todas las opiniones de los participantes en las sesiones. Para el análisis de las respuestas a las preguntas abiertas del formato del anexo 8 y de las opiniones de los participantes durante el grupo focal se usó el software ATLAS.ti versión 5.2.0, el cual permite hacer un análisis cualitativo visual de los datos. Se eligió este software porque permitía valorar las expresiones de los participantes en su conjunto, sin generar particiones que podrían llevar a analizar los argumentos sin un contexto.

Procedimiento

Como se mencionó anteriormente este estudio se desarrolló en dos fases. La primera consistió en la aplicación de una rutina estadística para la detección de los ítems con DIF a través de dos procedimientos y la segunda a una estrategia de tipo cualitativo donde se revisaron los ítems y se efectuaron grupos focales con expertos en el área disciplinar de evaluación de la prueba y conocedores de población indígena.

Fase 1. Identificación de Ítems con DIF.

El propósito de esta fase fue identificar los ítems que presentaban DIF y hacer un análisis general de los resultados a la luz de variables como impacto y ajuste de los ítems, que habían sido trabajadas por Arias (2008) y Berrío (2008). Así mismo, se buscó evaluar

el estado de algunos supuestos de la TRI que enmarcan una de las metodologías para identificar DIF y la concordancia entre los métodos usados. Como producto final se escogieron algunos ítems que se consideraron debían ser revisados para intentar identificar posibles fuentes de sesgo cultural.

Para cada una de las pruebas se extrajeron 20 muestras utilizando bootstrapping como método de remuestreo con el fin probar la estabilidad de los resultados obtenidos a partir de los análisis de DIF. El muestreo por bootstrapping consiste en realizar un muestreo aleatorio con reemplazamiento, en el que cada muestra es del mismo tamaño de la original. La muestra inicial también hizo parte de los análisis, es decir, que en total se tuvieron 21 muestras por cada prueba.

Verificación de la Unidimensionalidad. En primer lugar se aplicaron tres procedimientos con el fin de establecer el grado de unidimensionalidad en las pruebas objeto de este trabajo.

El primer método fue un análisis de componentes principales realizado para la muestra original de datos y con base en la matriz de correlaciones tetracóricas⁷. Este análisis se ejecutó de dos formas. La primera forma corresponde al análisis que comúnmente se lleva a cabo con este método, es decir, se obtienen los vectores y valores propios de la matriz de covarianzas sin transformaciones adicionales y se estiman las cargas de las variables en cada componente. Este procedimiento fue llevado a cabo con el paquete stats del software R. En contraste, la segunda forma arroja un subconjunto de sólo los mejores factores o componentes, re-escalando los vectores propios por la raíz cuadrada de los valores propios con el fin de producir cargas más características de un análisis factorial en los componentes (Revelle, 2012). En este segundo procedimiento, se usó rotación varimax y se efectuó con el paquete psych del software R.

⁷ Las correlaciones tetracóricas se usan cuando las variables son dicotómicas.

El segundo método fue el análisis paralelo que es una forma de determinar el número de factores o componentes en una matriz de correlación, examinando el diagrama de sedimentación o “scree-plot” de los valores propios sucesivos en donde el lugar en el que se rompe la forma del diagrama de sedimentación señala el número de componentes o factores apropiados a extraer. En el análisis paralelo se compara el “scree plot” de los datos observados con los de una matriz aleatoria simulada del mismo tamaño que la original (Revelle, 2012). Para este análisis se usó la matriz de correlaciones tetracóricas, la cual fue estimada con el paquete polycor. El análisis paralelo se llevó a cabo con el paquete psych del R y la comparación se hizo con 200 ensayos o réplicas.

Finalmente, el tercer análisis fue el propuesto por Linacre (2006), el cual sugiere realizar un análisis por componentes principales de los residuales, en donde éstos últimos corresponden a la varianza no explicada por la dimensión de Rasch, deben ser aleatorios y no mostrar estructura. De acuerdo con K. Konrad (Comunicación personal, Octubre 18, 2011) puede considerarse que se cuenta con un grado adecuado de unidimensionalidad si la varianza explicada por la prueba es del 40%. Sin embargo también se tuvo en cuenta lo propuesto por Linacre (2006) en cuanto a que siempre hay un grado de multidimensionalidad y la idea es preguntarse si la multidimensionalidad propuesta en la prueba es lo suficientemente grande para dividir los ítems en dos pruebas separadas o construir una nueva prueba. Linacre manifiesta que si se tienen dudas, se puede usar el primer contraste de los residuales. Así, se dividen los ítems de la prueba en dos subpruebas con base en las cargas positivas y negativas en el primer contraste, es decir, los que cargan positivamente forman una subprueba y los demás otra subprueba. Se estiman las puntuaciones para los individuos en cada una de las subpruebas, se efectúa una correlación de Pearson entre ellas y se estima el siguiente índice de correlación de acuerdo con lo que muestra la ecuación 3:

$$rT_xT_y = \frac{r_{XY}}{\sqrt{r_{xx}r_{yy}}} \quad (3)$$

Donde:

r_{XY} Es la correlación de Pearson,

r_{xx} Es la confiabilidad para la subprueba 1,

y

r_{yy} Es la confiabilidad para la subprueba 2.

Si rT_xT_y es cercano a uno, las dos subpruebas están midiendo el mismo atributo. Esto mismo se puede hacer para los siguientes contrastes, pero si los valores propios son menores a 1.4, puede carecer de importancia evaluarlos hasta ese punto. Para este análisis, la confiabilidad se estimó con el Kuder-Richardson 20: KR-20.

Identificación de Impacto. Como ha sido reportado en la literatura, los procedimientos que se usaron para la detección de ítems con DIF en el presente trabajo suelen verse afectados cuando hay impacto (Berrío, 2008; Arias, 2008), por lo que se consideró necesario estimar si existían diferencias reales entre los grupos de estudiantes indígenas y no indígenas. Para ello, siguiendo a Camilli & Shepard (1994) se tomó la diferencia entre el promedio del número de respuestas correctas de los dos grupos en cada una de las pruebas y dado que, por el tamaño de la muestra, cualquier mínima diferencia podría ser estadísticamente significativa se usó la *d de Cohen* (1988) como medida de tamaño del efecto, la cual se describe en la ecuación 4. Debido a que no se podía suponer que las varianzas de los dos grupos eran iguales, se estimó la varianza conjunta, aunque en ningún caso su diferencia con la obtenida por el grupo de referencia superó el valor de 0.01. Cohen (1992) considera que una $d = 0.20$ representa un efecto pequeño, una $d = 0.50$ un efecto medio y una $d = 0,80$ un efecto grande.

$$d = \left| \frac{m_A - m_B}{\sigma} \right| \quad (4)$$

Donde:

d Es el índice de tamaño del efecto para pruebas t de medias en unidades estándar.

m_A, m_B Es la media poblacional expresada en la unidad original de medida.

σ Es la desviación de cualquiera de las dos poblaciones, si se considera que tienen varianzas iguales.

Evaluación del Ajuste del modelo. Como medidas del ajuste del modelo se utilizaron el INFIT y el OUTFIT. El primero es un estadístico de ajuste que brinda información ponderada, “la cual es más sensible al comportamiento inesperado que afecta las respuestas dadas a los ítems cercanos al nivel de medida de las personas” (Linacre, 2006, p. 192). El *OUTFIT* también es un estadístico de ajuste, que a diferencia del *INFIT* es “más sensible al comportamiento inesperado de personas en los ítems lejos de su nivel de medida” (Linacre, 2006, p. 193).

Se consideró que un ítem estaba ajustado si la medida de ajuste estaba entre 0,8 y 1,2, con respecto a los dos índices. Se escogieron estos criterios con base en el trabajo de Smith, Shumacker y Busch (1998) en donde a través de simulaciones se encontró que el outfit presentaba tasas de error tipo I menores al 1% con 20 ítems y 1000 personas cuando se usaba un criterio de ajuste entre 0,8 y 1,2. En cuanto al *infit*, ese mismo trabajo indicó que las tasas de error tipo I eran cercanas a 0% con diferentes criterios, por lo que se escogió el mismo que el aplicado con el outfit.

Detección de DIF. De acuerdo con los resultados de Arias (2008) y Berrío (2008) se aplicó una rutina para la detección de ítems con DIF que estima el χ^2_{MH} y el Δ_{MH} . Además, se usó el software Winsteps para la estimación de la diferencia de la dificultad y se calculó la métrica que Berrío desarrolló para este método. Teniendo en cuenta que el Δ_{MH} como medida de tamaño del efecto permite un mayor control del error tipo I y una adecuada potencia para la detección de DIF uniforme en condiciones similares a las de SABER 11° (potencias de 0.79 y de 0.62 cuando se ajusta un modelo de un parámetro y razones de 1/100 y de 1/250, respectivamente) (Arias, 2008), se eligió este método para la detección de DIF en el presente trabajo. A pesar del bajo poder que presenta el Δ_{MH} para detectar DIF no uniforme, se considera como un aspecto a tener en cuenta, aunque no crucial para el desarrollo de esta investigación, puesto que el ICFES usa el modelo de Rasch para los análisis estadísticos de las pruebas de SABER 11°. En otras palabras, si se presenta DIF éste debe ser considerado como cambios en el parámetro de dificultad (DIF uniforme) y no en la

discriminación (DIF no uniforme), puesto que esta última se considera constante en el modelo de Rasch.

Se usó como variable criterio para el emparejamiento de los grupos el número de respuestas correctas y como puntos de corte para los estratos 3, 6, 9, 12, 14, 15, 17, 20, 23, 26 y 30, cuando la prueba fue de 30 preguntas y 3, 6, 9, 11, 12, 14, 17, 20 y 24, cuando fue de 24. Estos puntos de corte se establecieron tratando de mantener un número suficiente de evaluados en cada estrato.

Básicamente se siguió el procedimiento descrito en Arias (2008) que comprendió un procedimiento bietápico de purificación en donde en una primera etapa se eliminaron del cálculo de la variable criterio aquellos ítems que fueron clasificados en categoría C del Δ_{MH} . El *script* usado fue desarrollado por Cervantes (2007). La evaluación del tamaño del DIF, después de la etapa de purificación, se realizó a partir de la métrica del Δ_{MH} y se tuvo en cuenta la clasificación propuesta por el ETS. Adicionalmente para que un ítem fuera considerado con DIF severo o moderado el χ^2_{MH} de MH debía ser significativo con $p < 0.05$.

El segundo procedimiento utilizado fue la diferencia de dificultad descrito y usado en el estudio de Berrío (2008). Este procedimiento se aplicó sin etapa de purificación en la presente investigación por dos razones: a) esta forma ha sido la que suele usarse en el ICFES y en los estudios empíricos, y b) en el trabajo de Berrío (2008) se encontró que las tasas de error tipo I no mejoran sustancialmente con el procedimiento con o sin purificación, mientras que la métrica si se comporta mejor en este sentido. Para la métrica D se usó como punto de corte $|0.22|$, debido a que, de acuerdo con lo expuesto en la revisión bibliográfica, es el que permite un buen control del error tipo I sin perder potencia de prueba. Si el signo de la D es negativo favorece al grupo focal, si es positivo favorece al de referencia.

Dados los tamaños de muestra, las razones entre los grupos de indígenas y no indígenas son 1:124 (un estudiante indígena por 124 no indígenas) para la segunda aplicación del año 2006 y 1:89 para la primera del 2007. En este sentido, los hallazgos de

los estudios de Arias (2008) y Berrío (2008), muestran que las métricas Δ_{MH} y D se comportan adecuadamente con razones de tamaño de 1:20, 1:100 y 1:250, por lo que las razones de tamaño encontradas en los datos reales están en dicho rango y no es necesario hacer procedimientos adicionales.

Para evaluar el acuerdo entre los procedimientos se utilizó el Kappa de Cohen tomando los ítems que fueron clasificados en las categorías B y C por el Δ_{MH} y como medida de tamaño del efecto del Kappa se tuvieron en cuenta los criterios establecidos por Fleiss (1981, citado por Losada & Arnau, 2000) en los cuales un valor de kappa entre 0,4 y 0,6 es regular, entre 0,6 y 0,75 son buenos y excelentes por encima de 0,75.

Fase 2. Análisis substantivo de los ítems con DIF.

En esta fase se pretendió analizar los ítems que habían sido detectados con el fin de identificar posibles fuentes de sesgo cultural. Este análisis se hizo a partir de la revisión individual de los ítems por un grupo de expertos y grupos focales con estas mismas personas.

Para el análisis substantivo de los ítems que fueron detectados con DIF se escogieron personas que hubiesen participado en la construcción de ítems de las pruebas de Lenguaje, Matemáticas, Biología y Ciencias Sociales, por ser las áreas que más presentaron ítems con DIF y adicionalmente por ser las que tradicionalmente se han considerado como las más importantes. Todos estos participantes habían culminado su formación profesional en Licenciaturas en las diferentes áreas o tenían un título en un programa de educación superior afín, por ejemplo, eran Biólogos o Filósofos, aunque uno no cumplió este requisito pues se encontraba culminando su licenciatura. La gran mayoría estaban realizando sus estudios de Maestría o Doctorado. Su participación fue importante en este proceso porque ellos conocen el objeto de medida y están familiarizados con el proceso de elaboración de preguntas.

Además se contó con dos personas que conocían las comunidades indígenas desde perspectivas distintas: la antropología y la lingüística. Ambas habían desarrollado trabajo

de campo en comunidades indígenas y sabían hablar varias lenguas nativas. En total fueron 15 personas participantes distribuidas así: dos expertas en comunidades indígenas, tres autores de ítems de Lenguaje, tres de Matemáticas, cuatro de Ciencias Sociales y tres de Biología, en este último grupo participó además el coordinador del área de Biología en el ICFES. Cada uno estuvo presente en el análisis substantivo de los ítems de su área, y los participantes expertos en comunidades indígenas estuvieron presentes en todas las sesiones de análisis.

Cada una de las sesiones contó con la participación de un moderador, que en este caso fue la autora del presente trabajo y un asistente. Este último fue una persona con conocimientos en psicometría y construcción de ítems. Ambos tomaron notas durante el desarrollo de toda la sesión acerca de las apreciaciones sobre los ítems que hacían los participantes.

Con base en la literatura y en ejercicios como los de Sireci & Allouf (2003) y Zieky (1992) se planearon las sesiones de análisis de sesgo. Esta parte del trabajo se puede dividir en cuatro subfases: sesión de información, revisión individual de los ítems, grupos focales y transcripción de las opiniones de los participantes en las sesiones y análisis de las mismas.

Sesión de información. Esta sesión inició contextualizando el proyecto “Identificación de Ítems con Sesgo Cultural en las Pruebas de los Exámenes de Estado en Colombia”, sus antecedentes y los objetivos del presente trabajo. Se expuso el concepto básico de sesgo, tratando de que no implicara demasiados tecnicismos que interfirieran con los propósitos de esta fase. Se brindaron ejemplos sobre la diferenciación entre sesgo e impacto y se presentaron las fuentes potenciales de sesgo que se encuentran descritas en el anexo 9 junto con ejemplos que se habían reportado en la literatura. Se describieron los objetos y propósitos de medición de cada una de las pruebas a analizar, incluyendo sus competencias y componentes, y se les indicó a los participantes las condiciones generales del trabajo (revisión individual de los ítems y discusión en los grupos focales, confidencialidad, puntualidad y reconocimiento monetario y académico por participación en la investigación). También se ilustraron algunas características básicas de los grupos focal y

de referencia, tales como proporción en la población y en el estudio. Esta sesión duró aproximadamente una hora y fue desarrollada por aparte para cada una de las áreas.

Revisión individual de ítems. En un día diferente, posterior a la sesión de información, se realizó la sesión de análisis de sesgo que se compuso de una revisión individual de los ítems por parte de cada participante y un grupo focal. Durante la primera hora los participantes revisaron los ítems de forma individual, entre los cuales se encontraban ítems que habían sido detectados con DIF por alguna de las dos métricas y otros adicionales que no cumplían esta condición con el fin de que sirvieran de comparación.

Los ítems se presentaron a través de *video beam* y los participantes tuvieron alrededor de 5 minutos para revisar cada uno y consignar sus respuestas en el anexo 8. Esta revisión individual se realizó por dos razones: a) es posible que algunos de los participantes no se sintieran cómodos al expresar a un grupo sus ideas, por lo que se quisieron recoger sus opiniones en el formato, y b) dada la cantidad de ítems en algunas de las pruebas que era necesario revisar, se estimó que tal vez en dos horas no era factible dialogar sobre ellos en el grupo focal.

Grupos focales. Después de la revisión de ítems hubo un descanso de 15 minutos, en el cual la moderadora y el asistente estimaron el número de acuerdos para cada ítem, respecto a si presentaba favorecimiento a un grupo particular y a cuál grupo, con el fin de determinar cuáles ítems serían revisados en el grupo focal, en caso de que el tiempo no alcanzara para revisarlos todos. De esta forma, si eran cinco participantes y para un ítem tres estaban de acuerdo con que favorecía a un grupo y para otro ítem cuatro estaban de acuerdo en que no favorecía a un grupo particular, se revisaba primero el que presentaba solamente tres acuerdos.

Un grupo focal es una discusión grupal dirigida que debe durar entre una y dos horas, cuyo objeto es la prueba y se recomienda que el número de participantes sea entre 5 y 10 (Padilla et al, 2007; Snijkers, 2002). Estos grupos suelen ser homogéneos, compuestos por expertos, usuarios de los datos (estadísticos, investigadores, etc.) o quienes responden el cuestionario (Snijkers, 2002). Stewart, Shamdasani & Rook (2007) afirman que los grupos

focales, desde la investigación en sociología, psicología clínica y las ciencias de la salud son grupos que se reúnen y que comparten una identidad común, metas y situación concreta. Se sugiere que los participantes hayan contestado la prueba anteriormente e idealmente no conocerse entre sí (Padilla et al., 2007; Snijkers, 2002) y que el moderador siga un guion que contenga las preguntas importantes para la evaluación de la prueba.

Los grupos focales fueron escogidos por dos razones: a) a diferencia del juicio de expertos y de las entrevistas cognitivas, brindan evidencia sobre la perspectiva social y no sobre la dimensión cognitiva (Padilla et al., 2007), siendo el caso de este trabajo, y b) se querían conocer todas las opiniones sobre los ítems sin necesidad de llegar a un acuerdo, aspecto que es típico en el juicio de expertos.

La moderadora realizó todas las sesiones de grupos focales siguiendo el guión que se encuentra en el anexo 6. En términos generales, se proyectó cada ítem por medio del *vídeo beam* y se realizó la pregunta central: “¿*Qué aspectos del ítem hacen que se favorezca o no a un grupo particular?*” (pregunta 1). Posteriormente se ilustró el porcentaje de elección de cada opción de respuesta para los dos grupos (indígenas y no indígenas) con el fin de identificar si existía algún aspecto de las opciones no válidas que hicieran que uno de los grupos se inclinara por una de ellas en particular, también se informó sobre el grupo al que favorecía el ítem. La pregunta para los participantes en ese momento fue: “¿*Qué indicios nos brindan los porcentajes por opción de respuesta sobre la razón por la cual este ítem favorece al grupo XXXX?*” (pregunta 2”).

Posteriormente, se indagó cómo podrían definirse esos aspectos que potencialmente estaban favoreciendo a un grupo particular y qué nombres se les podría dar, indicando que podían usar las definiciones que se encontraban en la literatura (anexo 9), aunque no estaban limitados a ellas: “¿*Cómo podríamos definir esos aspectos que hacen que un ítem sea más favorable para un grupo que para otro?*” (pregunta 3). Finalmente, se preguntó sobre cómo podrían evitarse esos posibles favorecimientos: “¿*Cómo podemos mejorar esta pregunta para evitar que presente favorabilidad hacia algún grupo?*” (pregunta 4).

Análisis de los datos. Las respuestas a cada una de las preguntas de la revisión individual de los ítems y de los grupos focales fueron transcritas. Estos datos fueron analizados con el software ATLAS.ti, tomando como fuentes cada uno de los participantes de la sesión de análisis de sesgo. Se crearon ocho unidades hermenéuticas, dos por cada prueba que iba a ser analizada. La primera contenía las opiniones que los participantes habían registrado en el formato de revisión individual y la segunda las que habían sido expresadas en el grupo focal, con el fin de contar con toda la información sobre los ítems. De acuerdo con Muñoz (2005), una unidad hermenéutica se compone de varios elementos principales: documentos primarios, citas, códigos y anotaciones.

Los documentos primarios son los datos tal cual fueron transcritos, aunque también pueden ser imágenes o vídeo. Las citas son fragmentos de los documentos primarios que presentan algún significado y se consideran una primera reducción de los datos brutos. Los códigos generalmente son la unidad básica de análisis, pueden verse como conceptualizaciones, resúmenes o agrupaciones de citas, por lo que se consideran un segundo nivel de reducción de datos. El último elemento principal son las anotaciones que se consideran como comentarios cualitativamente superiores a los elementos mencionados anteriormente, porque comprenden desde notas recordatorias hasta conclusiones y pueden ser usadas como puntos de partida para la redacción de un informe (Muñoz, 2005).

Para cada unidad hermenéutica se tomaron como fuentes los participantes en la sesión de análisis de sesgo y se crearon códigos para ellos. En el caso de las unidades que fueron creadas para los formatos de revisión individual, se crearon dos documentos primarios que contenían las respuestas de las preguntas 1 a 3, y uno adicional con las respuestas a la pregunta 4. Las unidades para las opiniones expresadas en los grupos focales estuvieron compuestas por las intervenciones de todos los participantes. En un primer análisis se crearon códigos para todas las categorías que provenían de las fuentes de sesgo y factores culturales derivados de la literatura y que se encuentran en el anexo 9. También se crearon códigos para las categorías: posible fuente de sesgo no reportada en la literatura, ítems sin DIF, ítems con DIF y formas para evitar favorecimiento de las preguntas hacia un grupo particular (cómo mejorar la pregunta).

Cuando el ítem en revisión presentaba DIF, las citas que provenían de las preguntas 3 del formato del anexo 8 y de las preguntas uno a tres del grupo focal fueron tomadas para ser clasificadas en las categorías de fuentes de sesgo derivadas de la literatura y de “posibles fuentes de sesgo no reportadas en la literatura”. Las citas que provenían de la cuarta pregunta del anexo 8 y de la pregunta cuatro de los grupos focales fueron clasificadas en la categoría “cómo mejorar la pregunta”. Cuando el ítem no presentaba DIF, todas las citas fueron clasificadas en la categoría “Ítems sin DIF”, para ser posteriormente revisadas y realizar comparaciones con las que provenían de los ítems con DIF. Este primer análisis para los grupos focales lo llevó a cabo un asistente de investigación, quien se entrenó en el manejo del ATLAS.ti y en la teoría que sustenta el presente trabajo, posteriormente fue revisado por la autora. El análisis de los formatos también fue hecho en ATLAS.ti, aunque solo por la investigadora.

El segundo análisis consistió en la revisión de aquellas citas que habían sido clasificadas en la categoría “posible fuente de sesgo no reportada en la literatura” con el fin de identificar categorías emergentes en ellas. Luego se hizo la comparación de las citas que finalmente habían quedado en cada categoría con aquellas que estaban en la categoría “Ítems sin DIF” con el fin de considerar si había aspectos similares en las mismas y tenerlos en cuenta al momento de definir las fuentes finales de sesgo.

Finalmente, un tercer análisis buscó identificar si existían fuentes de sesgo transversales a las pruebas que fueron analizadas. Para ello se revisaron las salidas del programa ATLAS.ti para cada prueba y unidad hermenéutica, las cuales contienen una organización de los datos por categorías. Al terminar los análisis cualitativos se elaboraron pautas que guían la construcción de ítems libres de sesgo cultural con una estructura similar a la del documento del ETS (2009).

RESULTADOS

Detección de Ítems con DIF.

Este capítulo muestra los resultados del análisis llevado a cabo para detectar ítems con DIF a partir de los dos procedimientos propuestos, así como las detecciones a través de variables como impacto y ajuste de los ítems. También se mencionan aspectos relacionados con la verificación de la unidimensionalidad de las pruebas, el ajuste de los ítems y el acuerdo entre los dos procedimientos de detección de DIF.

Verificación de la Unidimensionalidad. La tabla 2 muestra el porcentaje de varianza explicado por los dos primeros factores y sus valores propios en cada uno de los procedimientos de ACP por prueba y aplicación. Como puede apreciarse, si se tienen en cuenta criterios convencionales como los propuestos por Abad, Garrido, Olea & Ponsoda (2006) según los cuales una prueba tendría un buen indicador de unidimensionalidad si el primer factor explica el 25% (después de eliminar ítems con cargas inferiores a 0,10 en el primer factor) de la varianza, ninguna de las pruebas usadas para este estudio muestran un alto grado de unidimensionalidad. Aunque la tabla 2 muestra los porcentajes de varianza incluyendo todos los ítems, debe mencionarse que se efectuó el ejercicio quitando aquellos con cargas menores a 0,1 en una de las pruebas y el porcentaje de varianza aumentó mínimamente. En este sentido, las pruebas con menor grado de unidimensionalidad son Filosofía y Física.

Igualmente, si se tienen en cuenta los criterios de Hattie (1984, 1985) y Gorsuch (1983), citados por Iraurgi (2008), según los cuales una razón superior a cuatro entre los valores propios del primer y segundo factor sería evidencia de unidimensionalidad, nuevamente la prueba de Física presenta uno de los menores grados de unidimensionalidad. Finalmente, las varianzas explicadas por las pruebas en las dos aplicaciones varían, siendo en unos casos mayor en la segunda aplicación de 2006 y en otros en la primera de 2007. Comparando los dos años, las varianzas explicadas son mayores en 2006 para las pruebas

de Biología, Filosofía y Física, y en 2007 lo son para las pruebas de Lenguaje, Matemáticas y Ciencias Sociales.

De otra parte, la tabla 3 muestra los resultados del análisis paralelo, indicando el número de componentes y factores que sería adecuado extraer, es decir, aquellos que son estadísticamente diferentes de los que se obtendrían a partir de una matriz del mismo tamaño pero sin correlación entre las variables. La prueba de Física se destaca nuevamente por ser la que presenta un mayor número de factores y componentes sugeridos por el análisis paralelo, lo que estaría de acuerdo con los resultados del ACP, en donde se encontró que esta prueba presenta el menor porcentaje de varianza explicado por los dos primeros componentes. No se observan diferencias superiores a dos factores o componentes sugeridos por el análisis entre las dos aplicaciones, exceptuando el número de componentes señalados para Matemáticas.

Tabla 2. Porcentaje de varianza explicado por los dos primeros componentes en los análisis de componentes principales (ACP)

Prueba	Factor	ACP		ACP *		Valores propios	
		II Aplicación de 2006	I Aplicación de 2007	II Aplicación de 2006	I Aplicación de 2007	II Aplicación de 2006	I Aplicación de 2007
Biología	I	15,05%	10,31%	13,00%	10,00%	2,77	1,6
	II	4,56%	4,72%	6,00%	5,00%	0,2	0,19
Filosofía	I	11,97%	7,87%	11,00%	7,00%	2,03	0,98
	II	5,41%	4,78%	6,00%	6,00%	0,37	0,18
Física	I	7,16%	6,46%	7,00%	6,00%	0,79	0,66
	II	5,35%	5,45%	5,00%	6,00%	0,31	0,33
Lenguaje	I	12,00%	14,53%	11,00%	15,00%	2,02	2,72
	II	5,27%	4,85%	7,00%	5,00%	0,32	0,23
Matemáticas	I	11,67%	16,56%	11,00%	16,00%	2,05	3,22
	II	5,98%	5,79%	7,00%	6,00%	0,48	0,47
Química	I	10,71%	13,48%	8,00%	12,00%	1,71	2,4
	II	5,53%	4,66%	8,00%	6,00%	0,43	0,23
Ciencias Sociales	I	9,00%	16,21%	8,00%	11,00%	1,8	4,12
	II	4,19%	4,64%	5,00%	10,00%	0,32	0,54

Nota: *Se refiere a la segunda forma de llevar a cabo el ACP en el que se maximizan las cargas en los componentes y

Prueba	Factor	ACP	ACP *	Valores propios
---------------	---------------	------------	--------------	------------------------

que fue llevado a cabo con rotación varimax.

Finalmente, la tabla 4 muestra que el índice propuesto por Linacre (2006) es alto para las pruebas de Filosofía y Ciencias Sociales y moderado para Matemáticas en la aplicación de 2006, mientras que en 2007 muestran este comportamiento las cuatro pruebas para las cuales resultaba adecuado calcularlos, dados los valores propios de sus contrastes de los residuales. Sin embargo debe tenerse en cuenta que en todas estas pruebas, a excepción de Ciencias Sociales, las correlaciones de Pearson son bajas entre las dos partes de la prueba surgidas de los ítems que cargan positivamente y los que cargan negativamente en el primer contraste de los residuales. En este sentido, el aumento en el índice propuesto por Linacre frente a la correlación de Pearson es debido al efecto correctivo que sobre esta última proporcionan las bajas confiabilidades en los dos subtests. En otras palabras, cuando uno o los dos instrumentos que se están correlacionando presentan confiabilidades bajas, el índice de Linacre es mucho mayor que la correlación de Pearson y su resultado debe observarse con cuidado porque el error de medición en ambas subpruebas es muy alto.

Tabla 3. Número de factores y componentes sugeridos por el análisis paralelo.

Prueba	Número de Factores a extraer		Número de componentes a extraer	
	II Aplicación de 2006	I Aplicación de 2007	II Aplicación de 2006	I Aplicación de 2007
Biología	13	11	7	6
Filosofía	11	12	8	7
Física	11	13	10	11
Lenguaje	12	11	8	6
Matemáticas	12	11	9	5
Química	11	11	8	6
Ciencias Sociales	13	14	8	7

Tabla 4. Resultados del análisis de unidimensionalidad a partir del ACP de los residuales.

Prueba	II Aplicación de 2006				I Aplicación de 2007			
	KR 20	KR 20	rT_xT_y	r	KR 20	KR 20	rT_xT_y	r
	Subtest 1	subtest 2			Subtest 1	subtest 2		
Biología	--	--	--	--	0,26	0,59	0,18	0,78
Filosofía	0,17	0,52	0,13	0,67	--	--	--	--
Física	0,00	0,06	0,37	0,02	--	--	--	--
Lenguaje	--	--	--	--	0,20	0,62	0,09	0,86
Matemáticas	0,13	0,50	0,17	0,45	0,25	0,64	0,19	0,72
Química	--	--	--	--	--	--	--	--
Ciencias Sociales	0,37	0,59	0,34	0,83	0,35	0,70	0,19	0,97

r = Índice de correlación de Pearson. rT_xT_y = Índice propuesto por Linacre (2006). *Nota:* Las celdas que aparecen con -- corresponden a análisis en los que el valor propio del primer contraste es igual o inferior a 1.4. No se muestran estos resultados siguiendo las recomendaciones de Linacre (2006) en las que señala que es probable que después de dicha magnitud en los valores propios, no hay mayor motivación para llevar a cabo el análisis del contraste.

En términos generales, aplicando diferentes procedimientos para evaluar la unidimensionalidad de las pruebas, se observó que ninguna de ellas presentó un alto grado puesto que: a) el porcentaje de varianza fue menor al sugerido por algunos autores (Abad y cols., 2006), b) los factores y componentes sugeridos por el análisis paralelo fueron mucho más de uno, y c) el índice propuesto por Linacre (2006), si bien en algunas pruebas fue alto o moderado, éste podría estar artificialmente aumentado debido a la baja confiabilidad de las dos partes de las pruebas que se están correlacionando. Adicionalmente, se aprecia que las pruebas con menor grado de unidimensionalidad tienden a ser las de Física y Filosofía.

Los hallazgos sobre la unidimensionalidad de las pruebas usadas en este trabajo presentan dos tipos de implicaciones: una en el proceso de construcción de las pruebas y, otra en los procedimientos usados para identificar DIF y posteriormente sesgo. La primera implicación se refiere a que es necesario revisar la definición del constructo y la construcción de ítems a partir del mismo, aspecto que es competencia del ICFES. La segunda, si bien la unidimensionalidad es uno de los supuestos de la TRI y por ende de uno de los procedimientos usados en el presente trabajo para la identificación de ítems con DIF,

la falta de ésta puede estar relacionada con sesgo. Por esta razón habrá que evaluar posteriormente si las posibles fuentes de sesgo que se encuentren al finalizar este estudio explican el bajo grado de unidimensionalidad, aunque este aspecto rebasa los límites del presente trabajo.

Identificación de Impacto. La tabla 5 ilustra el porcentaje de muestras en las cuales no se presentó impacto o se presentó impacto pequeño teniendo en cuenta la muestra original y las muestras que se obtuvieron por bootstrapping, el promedio de las diferencias de respuestas correctas entre los dos grupos de todas las muestras y la media de la d de Cohen (1988) que se usó como medida del tamaño del impacto para dichas diferencias. Con el fin de determinar el grado de impacto, se tuvieron en cuenta los criterios de Cohen (1992), de tal forma que las categorías quedaron establecidas así: a) una $d < 0.2$ señala que no hay impacto, b) si $0.20 \leq d < 0.50$ representa un impacto pequeño, c) si $0.50 \leq d < 0.80$ habría impacto mediano, y d) una $d \geq 0,80$ representa un impacto grande.

De acuerdo con estos resultados, se encuentra que en términos generales las mayores diferencias se observan en la segunda aplicación de 2006 puesto que cuenta con las tasas más altas de muestras con impacto pequeño y las mayores diferencias. En cuanto a las pruebas de Biología, Lenguaje, Matemáticas y Ciencias Sociales presentan las diferencias más amplias en ambas aplicaciones.

Tabla 5. Porcentaje de muestras clasificadas de acuerdo con su categoría de impacto, promedio de las diferencias de número de respuestas correctas entre los dos grupos y de la d de Cohen.

Prueba	II Aplicación de 2006				I Aplicación de 2007			
	% con $d < 0.2$	% con $0.20 \leq d < 0.50$	\bar{x} de las diferencias	\bar{x} de la d de Cohen	% con $d < 0.2$	% con $0.20 \leq d < 0.50$	\bar{x} de las diferencias	\bar{x} de la d de Cohen
Biología	0,00	100,00	1,18	0,34	23,81	76,19	0,72	0,22
Filosofía	0,00	100,00	0,87	0,29	100,00	0,00	0,20	0,07
Física	100,00	0,00	0,33	0,13	100,00	0,00	0,15	0,06
Lenguaje	0,00	100,00	1,12	0,35	0,00	100,00	0,94	0,28
Matemáticas	0,00	100,00	0,86	0,31	4,76	95,24	0,75	0,22
Química	38,10	61,90	0,57	0,20	100,00	0,00	0,21	0,06

Prueba	II Aplicación de 2006				I Aplicación de 2007			
	% con $d < 0.2$	% con $0.20 \leq d < 0.50$	\bar{x} de las diferencias	\bar{x} de la d de Cohen	% con $d < 0.2$	% con $0.20 \leq d < 0.50$	\bar{x} de las diferencias	\bar{x} de la d de Cohen
Sociales	0,00	100,00	1,32	0,34	33,33	66,67	0,92	0,22

Evaluación del Ajuste del modelo. La tabla 6 muestra la proporción de ítems por categoría de ajuste a través de las diferentes muestras teniendo en cuenta el *outfit*. En términos generales, la primera aplicación de 2007 mostró mayor proporción de ítems desajustados por *outfit*. Se destaca la prueba de Sociales por mostrar la mayor cantidad de ítems desajustados en las dos aplicaciones, diferenciándose en que para 2006 mostró tanto ítems con desajuste por arriba como por debajo de los rangos de outfit mencionados como criterios anteriormente, mientras que para la primera aplicación de 2007 únicamente se presentaron ítems desajustados con desajuste por arriba. No se presentaron ítems desajustados por *infit*.

Tabla 6. Porcentaje de ítems a través de las muestras de acuerdo con el ajuste al modelo evaluado a través del outfit.

Prueba	II Aplicación de 2006			I Aplicación de 2007		
	No desajuste	Desajuste por Abajo	Desajuste por arriba	No desajuste	Desajuste por Abajo	Desajuste por arriba
Biología	95,83%	0,00%	4,17%	100,00%	0,00%	,00%
Filosofía	100,00 %	0,00%	,00%	96,23%	0,00%	3,77%
Física	100,00%	0,00%	,00%	100,00%	0,00%	,00%
Lenguaje	99,01%	0,00%	,99%	99,21%	0,00%	0,79%
Matemáticas	100,00%	0,00%	,00%	95,83%	0,00%	4,17%
Química	100,00%	0,00%	,00%	100,00%	0,00%	,00%
Sociales	93,33%	3,33%	3,33%	93,97%	0,00%	6,03%

Detección de ítems con DIF a través del MH y la diferencia de la dificultad. Las tablas 7 y 8 ilustran el porcentaje promedio de ítems detectados con las pruebas estadísticas y los criterios de las métricas de la diferencia de la dificultad y el MH por prueba, a través de las 21 muestras en las dos aplicaciones y teniendo en cuenta el impacto. Como un primer resultado se observa que la tasa de detección de ítems con DIF a través de la prueba estadística es más alta para el método de diferencia de la dificultad que para el MH. Esta

situación tiende a repetirse cuando se comparan las métricas, siendo mayor el porcentaje de ítems clasificados como con DIF a través de D que con el Δ_{MH} .

En cuanto a las aplicaciones se observa que hubo una mayor tasa de detección con la prueba estadística en 2006, a excepción de Lenguaje, sin importar el procedimiento que se utilice. Este comportamiento puede deberse a la gran diferencia en la cantidad de evaluados entre las dos aplicaciones, siendo 2006 más de seis veces mayor que 2007. En contraste, la métrica del Δ_{MH} presenta un comportamiento diferente, puesto que en la primera aplicación del 2007 el promedio del porcentaje de ítems detectados a través de las muestras es mayor, exceptuando la prueba de Sociales. Respecto a la métrica D , ésta repite el mismo patrón que su prueba estadística, excepto para las pruebas de Biología y Lenguaje. Es de resaltar el hecho que ambas métricas detectan menos ítems con DIF que sus pruebas estadísticas tomadas aisladamente.

Al observar los resultados por prueba se aprecia que Biología tiene la tasa de detección más alta usando la métrica D en la aplicación de 2007 seguido de Química para 2006, y Lenguaje para 2007. Cuando se empleó el Δ_{MH} , se encontró que la prueba con la tasa más alta de detección promedio fue Ciencias Sociales para la segunda aplicación de 2006 y Lenguaje para la de 2007. Finalmente, teniendo en cuenta los ítems que son detectados a través de las muestras por los dos procedimientos, se observa que la tasa de detección más alta la presenta Lenguaje en la segunda aplicación de 2007.

Respecto al impacto, se observa que en la mayoría de comparaciones de las pruebas estadísticas y las métricas entre las muestras con y sin impacto, se presenta un mayor porcentaje de ítems con DIF cuando hay impacto pequeño, es decir, que las tasas de detección suelen ser más altas cuando existen diferencias pequeñas en la media de respuestas correctas (inferior a 0.2 desviaciones estándar). Se destaca este comportamiento en la prueba de Sociales de la I aplicación de 2007.

Las tablas 9 y 10 muestran el porcentaje de ítems detectados teniendo en cuenta el ajuste del modelo, en ellas se observa que pocos ítems mostraron desajuste; sin embargo,

cuando se tiene en cuenta la prueba estadística de los dos procedimientos hubo un 3,3% de los ítems de la prueba de Sociales en la segunda aplicación del 2006 que fueron detectados con DIF y que presentaron desajuste (outfit) por debajo del criterio establecido. Un comportamiento similar se observó en Biología en 2006 en donde un 4,2% de los ítems fueron detectados con DIF con la prueba estadística y presentaban desajuste por encima del criterio establecido.

Tabla 7. Porcentaje promedio de ítems detectados con DIF en las muestras para la II aplicación de 2006

Variable y prueba	$\chi^2_{MH} *$	B** del Δ_{MH}	Prueba estadística* de la diferencia de la dificultad	D de diferencia de la dificultad	Ambas métricas
General***					
Biología	67,06%	0,00%	68,65%	5,56%	0,00%
Filosofía	57,54%	0,00%	72,02%	5,95%	0,00%
Física	61,11%	0,20%	61,71%	8,53%	0,20%
Lenguaje	55,36%	0,20%	60,12%	9,52%	0,20%
Matemáticas	65,28%	0,40%	73,61%	7,14%	0,00%
Química	56,55%	0,00%	63,49%	14,48%	0,00%
Sociales	63,17%	3,49%	70,16%	3,17%	3,17%
Impacto***					
No impacto					
Química	55,73%	0,00%	64,06%	14,06%	0,00%
Impacto pequeño					
Química	57,05%	0,00%	63,14%	14,74%	0,00%

Nota: * $p < 0.05$, ** No hubo ítems detectados en la Categoría C del Δ_{MH} . ***Se omiten los porcentajes en las pruebas en las que la condición de impacto fue la misma en todas las muestras.

Tabla 8. Porcentaje promedio de ítems detectados con DIF en las muestras para la I aplicación de 2007

Variable y prueba	$\chi^2_{MH} *$	Δ_{MH}		Prueba estadística* de la diferencia de la dificultad	D de diferencia de la dificultad	Ambas métricas
		B	C			
General**						
Biología	40,87%	5,36%	0,00%	47,22%	25,79%	5,36%
Filosofía	47,22%	2,18%	0,00%	49,21%	3,77%	1,98%
Física	36,71%	1,19%	0,00%	37,10%	4,17%	,20%
Lenguaje	59,52%	8,93%	0,79%	63,29%	10,32%	6,55%
Matemáticas	34,92%	1,98%	0,60%	46,83%	3,17%	2,58%
Química	37,30%	3,37%	0,00%	39,68%	7,34%	2,58%
Sociales	51,43%	2,70%	0,00%	58,10%	1,59%	1,43%
Impacto**						
No impacto						
Biología	42,50%	5,83%	0,00%	47,50%	22,50%	5,83%
Matemáticas	33,33%	0,00%	0,00%	45,83%	4,17%	0,00%
Sociales	46,67%	1,90%	0,00%	52,86%	0,95%	0,48%
Impacto pequeño						
Biología	40,36%	5,21%	0,00%	47,14%	26,82%	5,21%

Variable y prueba	$\chi^2_{MH} *$	Δ_{MH}		Prueba estadística* de la diferencia de la dificultad	D de diferencia de la dificultad	Ambas métricas
		B	C			
Matemáticas	35,00%	2,08%	0,63%	46,88%	3,13%	2,71%
Sociales	53,81%	3,10%	0,00%	60,71%	1,90%	1,90%

Nota: * $p < 0.05$, ** Se omiten los porcentajes en las pruebas en las que la condición de impacto fue la misma en todas las muestras.

Tabla 9. Porcentaje de ítems detectados con DIF según categoría de desajuste (outfit) para la II aplicación de 2006

Variable y prueba	$\chi^2_{MH} *$	B** del Δ_{MH}	Prueba estadística* de la diferencia de la dificultad	D de diferencia de la dificultad	Ambas métricas
No desajuste (0,8 < Outfit < 1,2)					
Biología	62,90%	0,00%	64,48%	4,56%	0,00%
Filosofía	57,54%	0,00%	72,02%	5,95%	0,00%
Física	61,11%	0,20%	61,71%	8,53%	0,20%
Lenguaje	54,76%	0,20%	59,13%	9,52%	0,20%
Matemáticas	65,28%	0,40%	73,61%	7,14%	0,00%
Química	56,55%	0,00%	63,49%	14,48%	0,00%
Sociales	59,37%	3,49%	64,13%	3,17%	3,17%
Desajuste por Abajo (Outfit < 0,8)***					
Sociales	3,33%	0,00%	3,33%	0,00%	0,00%
Desajuste por arriba (Outfit > 1,2)***					
Biología	4,17%	0,00%	4,17%	0,99%	0,00%
Lenguaje	0,60%	0,00%	0,99%	0,00%	0,00%
Sociales	0,48%	0,00%	2,70%	0,00%	0,00%

Nota: * $p < 0.05$, ** No hubo ítems detectados en la Categoría C del Δ_{MH} . ***Se omiten los porcentajes en las pruebas en las que la condición de desajuste fue la misma para todos los ítems.

Tabla 10. Porcentaje de ítems detectados con DIF según categoría de ajuste (outfit) para la I aplicación de 2007

Variable y prueba	$\chi^2_{MH} *$	Δ_{MH}		Prueba estadística* de la diferencia de la dificultad	D de diferencia de la dificultad	Ambas métricas
		B	C			
No desajuste (0,8 < Outfit < 1,2)						
Biología	40,87%	5,36%	0,00%	47,22%	25,79%	5,36%
Filosofía	46,63%	2,18%	0,00%	48,81%	3,77%	1,98%
Física	36,71%	1,19%	0,00%	37,10%	4,17%	0,20%
Lenguaje	58,73%	8,93%	0,79%	62,5%	9,92%	6,55%
Matemáticas	34,92%	1,98%	0,60%	45,04%	3,17%	2,58%
Química	37,30%	3,37%	0,00%	39,68%	7,34%	2,58%
Sociales	50,63%	2,70%	0,00%	57,30%	1,59%	1,43%
Desajuste por arriba (Outfit > 1,2)**						
Filosofía	0,58%	0,00%	0,00%	0,39%	0,00%	0,00%
Lenguaje	0,79%	0,00%	0,00%	0,79%	0,40%	0,00%
Matemáticas	0,00%	0,00%	0,00%	1,79%	0,00%	0,00%
Sociales	0,79%	0,00%	0,00%	0,94%	0,00%	0,00%

Nota: * $p < 0.05$. **Se omiten los porcentajes en las pruebas en las que la condición de desajuste fue la misma para todos los ítems.

La tabla 11 muestra los ítems por aplicación y prueba que se consideran deben ser revisados por haber sido detectados por lo menos una vez a través de las muestras por

alguna de las métricas. De acuerdo con ello, se observa que hay más ítems de la primera aplicación de 2007 que requieren revisión, lo que puede deberse al hecho de que la razón entre los grupos focal y de referencia es menor en 2007, aspecto que se ha demostrado influye en la potencia del Δ_{MH} (Arias, 2008). Sin embargo, debe tenerse en cuenta que en la aplicación del 2006 se cuenta con mayor diversidad de etnias que en el 2007, por lo que la heterogeneidad de la primera puede estar enmascarando ítems que podrían presentar DIF para etnias particulares. Por otra parte, se aprecia que Lenguaje y Biología son las pruebas con mayor cantidad de ítems para revisión y que ambas aplicaciones presenta un número similar de ítems que favorecen a cada grupo. El anexo 11 ilustra el número de veces en el que cada ítem fue detectado con DIF a través de las muestras con las pruebas estadísticas y las métricas de los procedimientos utilizados.

Tabla 11. Ítems sugeridos para revisión por haber sido detectados con DIF en alguna de las muestras por al menos una de las métricas de los métodos usados.

Prueba	II Aplicación de 2006			I Aplicación de 2007		
	Ítem(s) que favorece(n) a			Ítem(s) que favorece(n) a		
	Focal	Referencia	Total	Focal	Referencia	Total
Biología	4, 11, 15	1	4	2, 4, 7, 8, 9, 20	10, 13, 16, 23, 24	11
Filosofía	2, 19	3, 5	4	5, 10, 11	7, 21	5
Física	7, 13	24	3	3, 10, 11, 14, 15	8, 16, 18	8
Lenguaje	16, 19	2, 10, 23	5	4, 11, 20, 23	6, 9, 14, 17, 19, 24	10
Matemáticas	20	5, 7, 8	4	11, 20	2	3
Química	3, 5, 14, 21, 24	20, 23	7	11, 12, 24	5, 7, 8, 21	7
Sociales		9, 10	2	6, 20, 21	5, 25, 26, 29	7
Total	15	14	29	26	25	51

Ejemplos de las curvas características de ítems que fueron detectados con DIF se encuentran en las figuras 7 y 8, el primero favorece al grupo focal y el segundo al grupo de referencia. Estos ítems se encuentran graficados teniendo en cuenta el modelo de Rasch y las dificultades arrojadas para cada uno en el procedimiento para detectar DIF que se efectuó con el software WinSteps.

Finalmente, el anexo 12 muestra los promedios de las dificultades de los ítems con DIF en las muestras en las que fueron detectados por alguna métrica, la dificultad total y para cada uno de los grupos; también se muestra el promedio de las diferencias entre ellas. Aunque el modelo de Rasch sostiene que los ítems tienen la discriminación especificada en el modelo ($a = 1$), este parámetro varía empíricamente, por lo que el Winsteps realiza una estimación de la discriminación post-hoc después de haber ajustado un modelo de Rasch (Linacre, 2006), este parámetro también es ilustrado en la tabla 14.

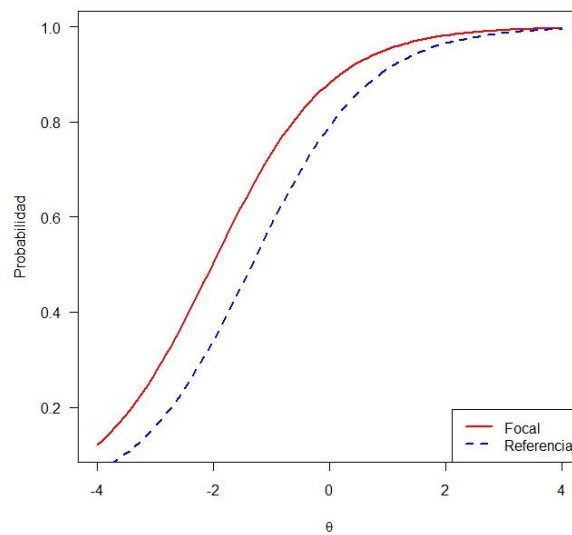


Figura 7. CCI del ítem 20 de la prueba de Lenguaje de la segunda aplicación de 2007 y la muestra 15 que favorece al grupo focal.

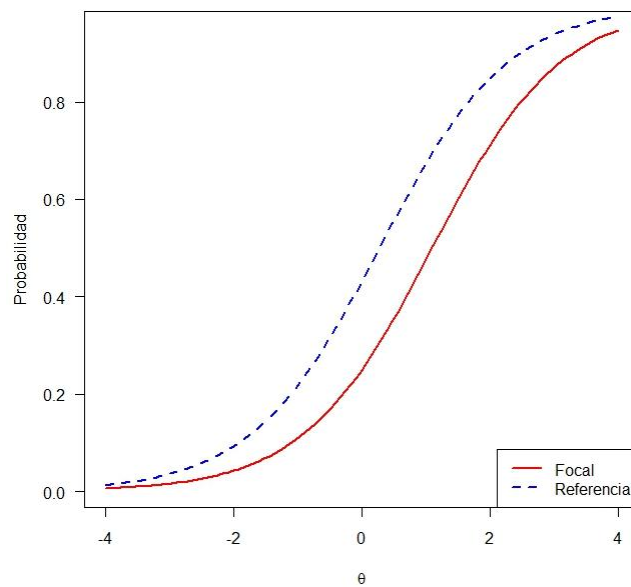


Figura 8. CCI del ítem 2 de la prueba de Matemáticas de la segunda aplicación de 2007 y la muestra 14 que favorece al grupo de referencia.

Herrera (2005) estableció para su trabajo algunos criterios de clasificación de la dificultad y la discriminación, estos fueron para el parámetro b : baja si $b \leq -1.5$, media si $-1.5 < b < 1.5$ y alta si $b \geq 1.5$; y para el parámetro a : baja si $a \leq 0.5$, media si $0.5 < a < 1$ y alta si $a \geq 1$. De acuerdo con dichos criterios de clasificación de las dificultades y las discriminaciones, la mayoría de los ítems detectados con DIF presentan una dificultad media (90%) y en cuanto a la discriminación el 52.5% de ellos se encuentran en la categoría alta y un 1.3% en baja. Sin embargo, debe tenerse en cuenta que la estimación de la discriminación fue tomada por la obtenida a través del Winsteps que es un software que trabaja con el modelo de un parámetro de la TRI, por lo que las estimaciones usadas en este trabajo, corresponden a una estimación a posteriori del modelo. Las diferencias de las dificultades se encontraron en promedio entre $|0.12|$ y $|0.65|$, teniendo en cuenta las muestras en las que fueron detectados con alguna de las métricas, mientras que para los no identificados con DIF estuvieron entre $|0.02|$ y $|0.5|$.

Medida de acuerdo entre los dos procedimientos. La tabla 12 muestra los resultados del acuerdo entre los dos procedimientos para detectar DIF, tanto con la prueba estadística como entre las métricas. Teniendo en cuenta los criterios establecidos por Fleiss (1981,

citado por Losada & Arnau, 2000), el acuerdo entre pruebas estadísticas es excelente para la aplicación de 2007 y bueno si se toman las dos aplicaciones en conjunto y la de 2006. Para las métricas, el acuerdo es regular en la I aplicación de 2007, mientras que para 2006 y las aplicaciones en conjunto, no alcanza siquiera el criterio de regular.

Tabla 12. Índice Kappa en la detección de ítems a través de las muestras entre pruebas estadísticas y métricas del MH y de la diferencia de la dificultad

Aplicación	Acuerdo entre pruebas estadísticas	Acuerdo entre Δ MH y D diferencia de la dificultad
General	0,733** (0,008)	0,328** (0,022)
II Aplicación de 2006	0,658** (0,015)	0,133** (0,026)
I Aplicación de 2007	0,786** (0,010)	0,472** (0,03)

Errores estándar en paréntesis. ** $p < 0,01$.

Análisis de sesgo.

A continuación se describen los resultados de las reuniones de análisis de sesgo para cada área. La primera parte corresponde a la clasificación de los ítems con DIF de acuerdo con su componente y competencia. La segunda parte muestra los resultados para cada área tomando en cuenta las expresiones brindadas por los participantes en los formatos y grupos focales, también lo que se obtuvo en relación con los ítems sin DIF y las posibles alternativas para evitar potenciales sesgos y mejorar los ítems. También se ilustran algunos ítems con DIF, los cuales fueron reproducidos con autorización del ICFES. Finalmente, se presentan posibles factores de sesgo transversales a las áreas evaluadas.

Debe mencionarse que fue necesario crear una categoría adicional en el análisis, que no estaba prevista en el método, ésta se denominó “Participante dice que sin favorecimiento” pues se consideró importante tener en cuenta este aspecto, para aquellos ítems que, aunque presentaban DIF, algunos de los participantes en algún momento lo consideraron como que no beneficiaban a algún grupo particular. Para cada análisis se presentan las afirmaciones que respaldan los factores culturales que se han considerado como potenciales fuentes de sesgo desde la revisión de la literatura y aquellos emergentes. Al final del capítulo se

definen aquellos factores nuevos y se redefinen los derivados de la literatura. No se analizaron ítems sin DIF para los grupos focales de Lenguaje, Matemáticas y Biología.

Clasificación de los Ítems Detectados con DIF por Componente y Competencia. La tabla 13 muestra la clasificación de los ítems con DIF por competencia que se consideraron debían ser revisados durante el análisis de sesgo. De acuerdo con ella, se puede apreciar que para Biología el número de ítems que presentan un comportamiento psicométrico que favorece a los indígenas es mayor frente al que favorece al grupo de referencia. Se destaca que en las competencias explicar e identificar hay cuatro y tres ítems, respectivamente, que favorecen a los indígenas, mientras que se presenta dos y uno para el otro grupo. Esto podría indicar que el grupo focal se encuentra favorecido por ítems que requieran capacidad para generar y entender argumentos, representaciones o modelos, así como que impliquen reconocer, diferenciar y representar fenómenos.

En la prueba de Sociales se observa que existen más ítems que favorecen a los no indígenas. En el caso de la competencia propositiva se observa un poco más favorecido el grupo de referencia, mientras que sucede lo contrario para la competencia interpretativa. Lo anterior indicaría que los ítems que demandan predecir sobre hechos e imaginar sus resultados, así como plantear alternativas, soluciones y reflexiones sobre fenómenos sociales, podrían favorecer a los no indígenas.

En la prueba de Matemáticas, el número de ítems que favorece a cada grupo es similar, sin embargo se destaca la competencia la comunicación y la representación por presentar tres ítems que favorecen al grupo de referencia. Lo anterior indicaría que preguntas que requieran que expresen ideas, usar diferentes tipos de representación y relacionar materiales físicos con aspectos matemáticos a través del lenguaje oral, escrito o gráfico, favorecería a los no indígenas.

Tabla 13. Número de Ítems que fueron detectados con DIF clasificados por competencia

Prueba y Competencia	Indígenas	No indígenas	Total
-----------------------------	------------------	---------------------	--------------

Prueba y Competencia	Indígenas	No indígenas	Total
Biología	9	6	15
Explicar	4	2	6
Identificar	3	1	4
Indagar	2	3	5
Sociales	3	6	9
Argumentativa	1	2	3
Interpretar	2	1	3
Propositiva		3	3
Matemáticas	3	4	7
El razonamiento y la argumentación	1		1
La comunicación y la representación	1	3	4
La modelación y planteamiento y resolución de problemas	1	1	1
Lenguaje	5	10	15
Argumentativa		2	2
Interpretar	4	4	8
Propositiva	1	4	5
Total	20	26	46

En Lenguaje se aprecia que 10 ítems benefician a los no indígenas y cinco al grupo focal. Se destacan las competencias argumentativa y propositiva por presentar dos y cuatro ítems, respectivamente, a favor de los no indígenas. Lo anterior indica que los ítems que demandan explicar las ideas que articulan y dan sentido a un texto, así como interpretar para generar una actitud crítica a través de los saberes de quien lee, benefician al grupo de referencia.

La tabla 14 muestra los ítems detectados con DIF clasificados por componente. En la prueba de Biología sobresalen los componentes Celular y Ecosistémico por presentar cuatro y tres ítems que benefician a los indígenas. Esto querría decir que el grupo focal se ve favorecido por ítems relacionados con la célula y su funcionamiento, así como la composición, organización de grupos de especies y sus relaciones con el ambiente en general.

En la prueba de Sociales se observa que hay más ítems que benefician al grupo de referencia que al grupo focal, específicamente aquellos relacionados con el componente “El

tiempo y las culturas”. Esto indica que los ítems que evalúan las relaciones entre los actores sociales y el significado de la acción en las dimensiones científicas, tecnológicas, éticas, religiosas o sapienciales, parecen desfavorecer a los indígenas. Para la prueba de Lenguaje se observa que los ítems que pertenecen al componente de la Función de los elementos locales favorecen a los no indígenas. Es decir, aquellos ítems que indagan sobre la función de los elementos locales y microtextuales al darle sentido al texto.

Tabla 14. Número de Ítems que fueron detectados con DIF clasificados por componente.

Prueba y Componente	Indígenas	No indígenas	Total
Biología	9	6	15
Celular	4	2	6
Ecosistémico	3	1	4
Organísmico	2	3	5
Sociales	3	6	9
El espacio, el territorio, el ambiente y la población	1		1
El poder, la economía y las organizaciones sociales	1	1	2
El tiempo y las culturas	1	5	6
Matemáticas	3	4	7
Aleatorio		3	1
Geométrico-métrico	2		1
Numérico-variacional	1	1	2
Lenguaje	5	10	15
Configuración del sentido global del texto	4	3	7
Del sentido hacia otros textos.		2	2
Función semántica de los elementos locales	1	5	4
Total general	20	26	46

Sesión de la prueba de Lenguaje. Las respuestas consignadas por los participantes en los formatos se clasificaron en siete categorías que podrían ser potenciales fuentes de sesgo, seis que de una manera u otra se desprendían de las reportadas en la literatura y una categoría emergente. Entre las primeras se encontraron: espiritismo, estilos de aprendizaje, experiencias más frecuentes en un grupo que en otro, objetos poco conocidos o usados por el grupo, perspectiva social del tiempo y pragmatismo. Como un nuevo potencial factor de

sesgo se encontró los problemas de construcción, siendo éste último el que presentó un 25% de las afirmaciones clasificadas en alguna categoría, mientras que las demás sólo obtuvieron una afirmación, por esta razón se describe a continuación.

Los problemas de construcción se evidencian a través de afirmaciones que los participantes realizaron como: *“Un evaluado indígena que no comprenda bien la redacción del texto elegirá D porque hay un guiño falso en esa opción, pues pertenece a un campo de sentido fácilmente distinguible de las otras opciones: es la única opción referida a una persona y no a objetos o sustancias abstractas.”* y *“su lectura puede ser un poco demorada por la naturaleza del mismo.”*. Estas frases se refieren a un ítem que favorece a los no indígenas y que se muestra en la figura 8a, cuya respuesta correcta era la A y en la que se observó que mientras un 9.6% de no indígenas eligieron la D, un 14% de indígenas escogieron esta opción. Así mismo debe señalarse un posible problema de diagramación en este ítem, pues en la opción B no tenía un punto, aspecto que, según uno de los participantes, podría ser interpretado como una señal de que esa es la clave; de hecho el 34.4% de los indígenas la elige contra un 28.4% de los no indígenas.

En los ítems sin DIF que fueron usados para hacer comparaciones, no se encontraron afirmaciones que contradijeran lo expuesto sobre problemas de construcción. En cuanto a las recomendaciones que hicieron los participantes sobre cómo mejorar las preguntas evitando posibles sesgos, se encuentra que: *“El ítem podría explicar un poco dicho contexto histórico (brevemente), para hacer la pregunta o remitirse a lo que el texto proporciona.”*, *“Hacer una prueba de inferencia sobre información del mismo texto y no de su contexto cultural.”* y *“Hacer una descripción corta de los adjetivos usados en las opciones.”*.

9. En la expresión: “mientras intento engañar con él las horas largas de la noche”, la palabra subrayada se refiere al
- A. lienzo.
 - B. tiempo
 - C. lecho.
 - D. adúltero.

Figura 8. Ítem con DIF en la prueba de Lenguaje que favorece a los no indígenas.

En cuanto a los grupos focales, se encontraron como posibles fuentes de sesgo relacionadas con las que se reportan en la literatura: comunalismo, creencias, epistemología de las comunidades de origen, espiritismo, estilos de aprendizaje, experiencias más frecuentes en un grupo que en otro, interferencia fonológica, ortográfica o semántica, objetos poco conocidos por el grupo, oralidad, perspectiva social del tiempo y pragmatismo. Entre las categorías emergentes como posibles fuentes de sesgo se hallaron: la diagramación, el colonialismo religión – Estado, problemas de construcción y tecnicismos. Entre todas estas fuentes se destacan los problemas de construcción, pues de las categorizaciones hechas sobre las afirmaciones, obtuvo el 26.67%, razón por la cual será descrita a continuación.

Los problemas de construcción se evidenciaron en las siguientes afirmaciones: *“creería que la ‘C’ se puede cambiar y la palabra momentáneo se podría cambiar para dejarla igual que la otra y visualmente no sobresalga, debe ser equilibrada en la frase y en lo semántico para que no se descarte de entrada por extensión o porque es muy corta.”*, *“esta clase de textos son para ellos (indígenas) los más difíciles porque lo tienen que releer, ellos están más atentos a la escritura.”* – aunque el ítem sobre el que se hace esta apreciación favorece a los indígenas -, y *“Se podrían tratar los mismos temas pero la complejidad del texto debería radicar no tanto en su lectura, sino en las preguntas que se respondan”*, *“confunde la extensión y uno sabe que el tiempo de respuesta es uno o dos minutos... También está mal construido, hay una (opción) que no concuerda con las otras tres...”*.

Respecto a cómo mejorar las preguntas, se destacan afirmaciones relacionadas con hacer los textos más sencillos y específicos, y que se tenga cuidado al momento de tener opciones de respuesta de orden categorial diferente. Lo anterior se puede afirmar a partir de sugerencias como: *“concentrándose en algún tema o aspecto”*, *“podría conservar el contenido pero con una estructura más sencilla”*, *“dejar dos sustantivos y dos abstractos en las opciones.”*, *“parear ‘B’ con ‘D’, porque ‘A’ y ‘C’ están pareadas porque ambas son objetos... toca parear esa diferencia, que ‘C’ y ‘D’ sean de la misma categoría.”* y *“para arreglar el ítem se debe hacer una descripción corta de los adjetivos que se utilizan,*

así queda más al alcance y el indígena sabría que le están evaluando respecto de lo sensorial”.

Como resultado general de la sesión de análisis de sesgo de la prueba de Lenguaje, es importante mencionar que hubo pocas afirmaciones registradas en los formatos de revisión individual, lo que puede deberse a lo manifestado por los participantes al terminar la primera parte de la sesión, pues indicaron que ellos esperaban ver aspectos más claros en los ítems que les pudieran dar ideas sobre por qué un ítem favorecía a un grupo u a otro. Finalmente, si bien se supone que los problemas de construcción desfavorecen a los grupos por igual, de acuerdo con lo expresado por los participantes en la sesión de lenguaje, parece que exacerban posibles comportamientos psicométricos diferentes en los ítems al comparar su desempeño entre los dos grupos estudiados.

Sesión de la prueba de Matemáticas. De acuerdo con las respuestas consignadas en los formatos, se identificaron cinco posibles fuentes de sesgo derivadas de la literatura: epistemología de las comunidades de origen, estilos de aprendizaje, experiencias más frecuentes en un grupo que en otro y objetos poco conocidos o usados por el grupo. Nuevamente como un factor emergente de sesgo se encuentran los problemas de construcción. Entre las clasificaciones realizadas sobre las afirmaciones, sobresalen las experiencias más frecuentes en un grupo que en otro, pues esta posible fuente cuenta con un 40% de ellas, por lo que se describe a continuación.

Las experiencias más frecuentes en un grupo que en otro fueron evidenciadas a través de afirmaciones como: *“Dado que el futbol tiene amigos y enemigos personalmente considero que a los enemigos les molestaría el contexto por su desconocimiento.”*, *“Evaluados que están familiarizados con la metodología utilizada en la FIFA para elegir al campeón mundial, realizan una lectura más sencilla de la gráfica e interpretan con mayor facilidad lo que pregunta el ítem.”* y *“El tipo de contexto es más familiar para las personas no indígenas.”*. Todas estas afirmaciones se realizaron sobre un ítem con un contexto que explica cómo se asignan los equipos en un mundial de fútbol a cada uno de los grupos y la forma en que se desarrolla el mismo. Respecto a los ítems sin DIF, no se

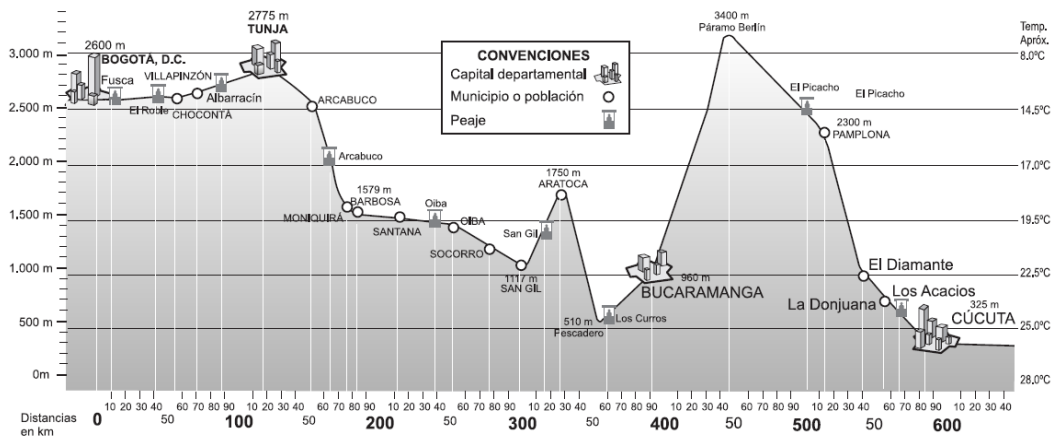
encontraron afirmaciones que estuvieran relacionadas con experiencias más frecuente en un grupo que en otro.

Sobre la forma en que se puede mejorar la pregunta, los participantes recomendaron frente al ítem de la figura 9 (respuesta correcta la C y favorece a los no indígenas) cambiar la palabra ciudad por población o municipio o dejar sólo las ciudades en la gráfica omitiendo los municipios y peajes. Respecto al ítem de fútbol, sugieren: *“que el contexto debe ser sin hacer énfasis en todo el proceso que sigue un equipo para quedar campeón”*. Finalmente, en preguntas sobre funciones matemáticas sugieren graficarlas (*“Hacerlo gráficamente con las funciones impares simétricas respecto del origen”*) y en otras sintetizar más el contexto y hacerlo más general.

Al analizar los resultados del grupo focal se encontraron las siguientes posibles fuentes de sesgo asociadas con las que se presentan en la literatura: comunalismo, creencias, epistemología de las comunidades de origen, estilos de aprendizaje, experiencias más frecuentes en un grupo que en otro, objetos poco conocidos o usados por el grupo, organización de los textos y pragmatismo. Como nuevos posibles factores de sesgo surgieron los problemas de construcción y los tecnicismos innecesarios. En general, se destacan los problemas de construcción y las experiencias más frecuentes en un grupo que en otro, porque de las clasificaciones realizadas sobre las afirmaciones, éstas presentaron un porcentaje del 25.64%.

En cuanto a los problemas de construcción se tienen afirmaciones como: *“no sé qué tan importante sea nombrar el eje Y, por que ahí dice grados y debe ser una mejor especificación para la lectura de la gráfica”*, *“puede darse dado que Pamplona no es ciudad es población, tener cuidado como se nombran esas cosas.”*, *“de pronto tiene que ver con el nivel de dificultad de la pregunta el tener mucha información.”*, *“creo que algo que mide la prueba es que el hecho de ser competente en matemáticas también implica la lectura en ese sentido.”* y *“No sé qué tan susceptible sea el hecho de colocar la información de la tumba, no sé qué sea eso para una comunidad indígena, si este tipo de cosas se considere algo malo, yo quitaría esos dos últimos renglones.”*, - aunque este ítem

favorecía a los indígenas –. Las dos primeras afirmaciones hacen referencia al ítem de la figura 9.



32. A partir de la información de la gráfica se puede afirmar que la ciudad que está a una altura mayor de 2.000m, tiene una temperatura promedio menor que 17°C y está a más de 500Km de Bogotá es

- A. Tunja.
- B. Cúcuta.
- C. Pamplona.
- D. Bucaramanga.

Figura 9. Ítem con DIF de la prueba de Matemáticas que favorece a los no indígenas.

Las experiencias más frecuentes en un grupo que en otro se evidenciaron a través de las siguientes afirmaciones “alguien que salga de paseo así sea una vez al año sabe esa diferencia, es la familiaridad con las convenciones...”, “tienen que asociar capital departamental con ciudad, para ellos debe ser algo más grande como Tunja, hay que referirse a lo grande”, refiriéndose al ítem de la figura 9, “para un indígena u otra persona es diferente que yo vea el bachillerato en un internado o en una escuelita, en otro lado o en un colegio departamental en la Guajira, para los indígenas es difícil realizar este tipo de tareas porque no está en su ámbito o contexto...”, aunque esta apreciación hacía referencia a un ítem que favorecía a los indígenas, “Los que no estudian no han visto las gráficas, en el Amazonas se quiere que los maestros también sean indígenas y estas cosas podrían dificultarse un poco.”, “a las mujeres no nos gusta el fútbol. Creo que desfavorece a los no

indígenas y las mujeres particularmente, uno le pregunta a una mujer y no sabe que hay cuatro partidos o eliminatorias, a pesar de que está presente, las niñas en el aula siempre reniegan contra eso, siento que uno se encuentra con una pregunta de esas y le molesta. Es tratar de asociar una cosa con la que uno no está familiarizado...” y “de pronto mirar si hay moda de mundial en ese año porque eso también puede favorecer. Si hay una comunidad en la que estos temas no les interese y se apartan del fútbol y teniendo en cuenta que desde el principio tienen problemas entre las mujeres no indígenas imagínese para una indígena que se dedica a tejer, se indispone.”.

Igualmente, respecto al ítem de la figura 9 también se señaló: *“yo apoyo que la gráfica es de poner atención y mirar que es lo que están preguntando, no creería que favorezca algún grupo pero sí es necesario cambiar el enunciado de la pregunta ciudad por población para evitar confusión, hay que hacer más precisa la información de la gráfica” y “Creo que el ítem está bien construido y la gráfica está bien para que lo puedan responder, tiene sus detalles... yo veo que está construido de cierta manera que no crea dificultad”.* También respecto a los ítems sobre fútbol un participante afirmó: *“a los indígenas les fascina el fútbol...y los Ticuna juegan bien al fútbol”, “aunque a mí no me guste el fútbol de todos modos puedo entender perfectamente”, “creo que como hicieron más goles en el mundial de Francia entonces ponen que ahí fue mayor el promedio, pero no tienen en cuenta el promedio de los goles sino que solo escogen el más alto, puede haber una mala interpretación.”,* *“no hicieron el procedimiento matemático, al final pusieron los goles sumados.” y “Yo creería que el problema está en asociar a porcentajes y diagramas de torta.”.*

En cuanto a la pregunta de cómo se podrían mejorar los ítems teniendo en cuenta los posibles favorecimientos, los participantes manifestaron: *“la función impar que no tenga números impares o el denominador diferente”, “yo colocaría todas las opciones con coordenadas y con puntos en el plano...”,* *“Colocar en el eje X rangos o preguntar por grupos de edad.”,* *“si dicen que el incremento fue mayor en el rango tendría que especificar en cuál rango...”,* *“es más fácil decir grupos de edades.”,* *“si dicen que el incremento fue mayor en el rango tendrá que especificar en cuál rango, porque podrían*

mirar es por grupos.” y “yo quitaría el énfasis en el proceso que tiene un equipo para ser campeón, porque no se necesita para responder la pregunta”.

En términos generales, se aprecia que existen potenciales fuentes de sesgo relacionadas con las experiencias más frecuentes en un grupo que en otro que se manifiestan en aspectos como la familiaridad con las convenciones y en deportes como el fútbol. También sobresalen los problemas de construcción en los ítems, específicamente con la utilización de términos que pueden no ser claros para alguno de los grupos como por ejemplo: la relación existente entre población, ciudad y municipio, aspecto que no es relevante para lo que se pretende evaluar en la prueba de Matemáticas. Igualmente, dichos factores se expresan en aspectos de los ítems como presentar mucha información, la relación de la lectura con lo que se pretende medir e información que puede molestar a alguno de los grupos. Sin embargo, estas potenciales fuentes de sesgo, deben verse con cuidado puesto que algunos participantes expresaron afirmaciones contradictorias con las de otros, específicamente en los ítems sobre fútbol y el ítem de la figura 9. Adicionalmente, los participantes manifestaron que los ítems que implican la interpretación de gráficas parecen ser más difíciles para los indígenas, sobre todo cuando se trata de diagramas de torta.

Sesión de la prueba de Biología. A través de las respuestas a los formatos se identificaron como posibles fuentes de sesgo que corresponden a las derivadas de la literatura: las creencias; la epistemología de las comunidades de origen; los estilos de aprendizaje; las experiencias más frecuentes en un grupo que en otro; forma en que se toman las decisiones; interferencia fonológica, ortográfica o semántica; objetos poco conocidos o usados por el grupo; organización de los textos; y pragmatismo. Como posibles factores de sesgo emergentes se evidenciaron la escuela tradicional, los problemas de construcción y los tecnicismos innecesarios para responder la pregunta. De todas ellas se describirán la epistemología de las comunidades de origen y las experiencias más frecuentes en un grupo que en otro, pues de las clasificaciones realizadas sobre las afirmaciones, obtuvieron 26.1% y 19.1%, respectivamente. Las descripciones se presentan en la tabla 15.

En cuanto a los grupos focales se encontró que de los factores que pueden introducir sesgos culturales y que son derivados de la literatura, podían observarse los siguientes: armonía, creencias, epistemología de las comunidades de origen, estilos de aprendizaje, experiencias más frecuentes en un grupo que en otro, género de los textos, objetos poco conocidos o usados por el grupo, oralidad, organización de los textos y pragmatismo. Entre los posibles factores que pueden estar asociados con sesgo cultural y que emergieron en el análisis se encuentran: colonialismo religión Estado, escuela tradicional, problemas de construcción y tecnicismos innecesarios. De todas ellas se describirán los problemas de construcción y las experiencias más frecuentes en un grupo que en otro, puesto que de las categorizaciones realizadas sobre las afirmaciones, éstas obtuvieron las frecuencias más altas, siendo ambas del 20% y 22%, respectivamente. Las afirmaciones correspondientes a estas potenciales fuentes de sesgo cultural aparecen en la tabla 15.

En términos generales, aunque se tienen tres potenciales fuentes de sesgo: epistemología de las comunidades de origen, problemas de construcción y experiencias más frecuentes en un grupo que en otro, las tres deben observarse con cuidado pues se presentaron afirmaciones similares sobre ítems sin DIF o con opiniones contrarias por parte de otros participantes. El afirmar que estos aspectos son o no factores potenciales de sesgo, siempre debe observarse a la luz de si son fuentes irrelevantes de varianza frente a lo que se pretende medir, es decir, si no se encuentran relacionados con el atributo.

Tabla 15. Afirmaciones de los participantes en la sesión de análisis de sesgo de la prueba de Biología.

Fuente potencial de sesgo o aspecto	Análisis	Afirmaciones
Epistemología de las comunidades de origen	Revisión individual de ítems	<i>“en algunas comunidades estos matrimonios son normales, en otras no.”, “Debido a que los indígenas, al igual que los beduinos, por vivir en comunidades pequeñas podrían tener mayor probabilidad de casarse con un pariente la pregunta podría ser alarmante para ellos.”, “Diferente concepto de familia en especial diferencian los “primos”...”, “Epistemología-método científico ¿cómo concibe el mundo?...” y “Quizá no es familiar el uso de productos químicos sino de otro tipo (natural).” Las tres primeras aseveraciones hacen referencia a un ítem que indaga sobre las consecuencias de casarse entre parientes cercanos a partir de una narración acerca de una comunidad de un país del medio oriente.</i>
Experiencias más frecuentes en	Revisión individual de ítems	<i>“Puede interferir las convenciones textuales, sin embargo, ellos (indígenas) están más expuestos a ver y observar los animales en su diario vivir”, “Quizá sería poco familiar la diferencia entre un perro criado en familia y uno de la calle, esa diferencia podría ser sutil”, “El indígena en su diario vivir está en contacto con los animales”, “Los indígenas podrían tener más experiencia ya que éste es un</i>

Fuente potencial de sesgo o aspecto	Análisis	Afirmaciones
un grupo que en otro	Grupo focal	<p><i>comportamiento que se evidencia también en otros animales que son más comunes para los indígenas” y “La interpretación de gráficos y los experimentos de laboratorio podrían ser experiencias más comunes para los no indígenas.”. Las cuatro primeras afirmaciones provienen de tres ítems que favorecían a los indígenas y que se referían a un contexto que habla sobre los perros, uno de esos ítems se ejemplifica en la figura 10, cuya respuesta correcta es la D. Sin embargo uno de los participantes afirmó respecto a uno de estos tres ítems que “Es familiar para indígenas y no indígenas”.</i></p> <p><i>“yo digo que para los indígenas es más familiar, si ellos son muy observadores esas fotos de perros de razas muy definidas les podrían parecer otro animal”, “yo puse que si tenía sesgo a favor de los indígenas, porque me pareció que ese comportamiento de dos animales que se huelen la cola lo podrían ver en otros animales y ser más familiar para ellos que para alguien en la ciudad.”, “los indígenas son muy observadores ellos sí miran por qué el perro se está comportando así, mientras que acá en la ciudad uno ve unos perros y sigue de largo a hacer sus cosas, uno no se pone a ver los perros a ver qué están haciendo.”, “yo puse que favorecía a los indígenas porque están más acostumbrados a matar animales”, “ellos cuando matan los animales se dan cuenta de quien tiene el intestino más largo, si el cerdo o la gallina...”, “La enfermedad del pulmón negro es común en los pueblos donde se extrae el carbón y ellos están acostumbrados a ver eso, creería que es más común para los indígenas.” -aunque estas tres últimas afirmaciones se referían a un ítem favorecía a los no indígenas-, “uno creería que a partir de eso las preguntas favorecen más a los no indígenas al ser más empíricos.” y “el análisis de información porque es entender el experimento y ponerlo en una gráfica...”. Adicionalmente, se tuvieron afirmaciones como: “yo creo que el hecho de uno verlo cotidianamente en la calle, ese comportamiento le es familiar a un indígena y a un no indígena, en las comunidades indígenas también hay perros para familiarizarse un poco con eso, por eso consideré que no hay preferencias hacia grupos no indígenas o indígenas.”, “la pregunta no exige cosas diferentes a las comunidades.”, y “yo puse que no tiene problema”, frente a los ítems que trataban sobre los perros y uno que versaba sobre un experimento para probar la acción de una enzima.</i></p>
Problemas de Construcción	Grupo Focal	<p><i>“Creo que de acuerdo al enunciado, dada la información que incluye, la respuesta correcta es la B, aunque la respuesta correcta sea la C, hay que cambiar eso, el aspecto genético en realidad sí tiene que ver y respecto del enunciado dice que no, además es la B según el enunciado, porque es la única variable que se modifica.”, “hay algunas cosas que son irrelevantes para la pregunta, como es mucho texto, lo pongo que como es para una segunda lengua esto podría dar una carga cognitiva adicional” (aunque el ítem al que se hace referencia favorece a los indígenas), “yo puse que hay mucha información irrelevante y que puede perderlos mucho ya que se habla de una cantidad de comportamientos”, “...si se dan cuenta la opción “D” tiene la misma palabra que el enunciado”, “Hay mucha información que no es relevante, con solo la pregunta del ítem se puede dar respuesta”, “hay texto que no se utiliza para la pregunta y hace perder tiempo en el examen”, “buscando dos animales similares, el omnívoro no lo aterriza mucho.”, “la que no responden los indígenas es la única que no tiene los nombres de la tabla como la “B” carbónico inferior, la más accesible es la “A”.”, “los porcentajes pueden confundir porque en la gráfica utilizan la palabra digestión pero en la tarea, entienden que 0% digestión como que digirió todo y no quedó nada, donde dicen 100% ellos asumen que está entera.” y “quería ver la construcción general de la pregunta para hacerle el análisis de manera lingüística y efectivamente el texto no tiene relación con lo que está en la pregunta”. Debe tenerse en cuenta que para uno de estos ítems un participante afirmó: “está bastante claro.”.</i></p>
Ítems sin DIF	Revisión individual de ítems	<p><i>“dependiendo de la cultura usan distintas terapias y el texto nunca especifica algo sobre esto”, “los problemas genéticos, mal formaciones o discapacidades se ven como males al espíritu y a veces son desterrados.” y “Por la frecuencia de la actividad más en un grupo que en otro”.</i></p>
Cómo mejorar los ítems	Revisión individual de ítems	<p><i>“Cambiano la forma en que se pregunta para que las opciones que tienen que ver con toxicidad y que no manejan los indígenas no afecten la respuesta”, “Se podría decir simplemente, coloco los tubos de ensayo a 36°C para evitar confusión con los términos.”, “Canis Familiaris, la mayor parte del texto no se usa en las preguntas y les puede hacer perder tiempo en el examen.”, “Utilizar un lenguaje menos técnico aunque allí se podría perder realmente lo que se quiere medir.”, “Podría brindarse mayor información acerca de los omnívoros - carnívoros.”, “Quitar la palabra secundarios y reemplazarla por otra más cercana.”, “Quitar contexto innecesario.”, “Dejar sólo al perro sin la historia.”, “Se debe quitar el contexto ciudadano”, “Pueden sustituirse los tubos por vasos.”, “Es complejo porque éste se deriva de la alfabetización científica, pero se puede simplificar más los elementos usados.”, “No usar tecnicismos”, “Contextos más generales.”, “Eliminar condiciones anaeróbicas “turberas”.” y “Simplificar enunciado y gráficos, % confunden.”.</i></p>

Fuente potencial de sesgo o aspecto	Análisis	Afirmaciones
Grupo focal		<p><i>“creo que la opción "A", que tiene la palabra rumiarse, debería estar dentro del enunciado, sería bueno entender a qué se refiere.”, “pondría omnívoros y herbívoros”, “yo creo que preguntando la diferencia del tracto digestivo en comparación con el del perro... buscando dos animales similares.”, “creo que si se quita la palabra barrios se puede cambiar por otra” y “se le puede quitar lo de la familia y se puede dejar como: uno de los perros se crió con el resto de la manada y el otro estuvo sólo. No cambia la situación de la pregunta pero se quita lo del perro que es visto como mascota”.</i></p>

32. Caifás y Pluto son dos perros que viven en barrios diferentes. Pluto se acercó a la zona de Caifás y éste reaccionó oliéndole las glándulas anales (que contienen el líquido odorífero particular de cada perro).



Este comportamiento de los perros es debido a que las glándulas anales

- A. difunden olores de los desechos de la digestión que identifican al perro.
- B. difunden hormonas con las que los perros diferencian las razas.
- C. secretan químicos característicos del medio en el cual se desarrolla el perro.
- D. secretan sustancias que determinan pautas de comportamiento.

Figura 10. Ítem con DIF de la prueba de Biología y que favorece a los indígenas.

Sesión de Ciencias Sociales. De acuerdo con las afirmaciones registradas por los participantes en los formatos individuales de revisión de ítems, se encontraron los siguientes potenciales factores de sesgo que se desprenden de la literatura: comunalismo, creencias, epistemología de las comunidades de origen, espiritismo, estilos de aprendizaje, experiencias más frecuentes en un grupo que en otro, género de los textos, objetos poco conocidos o usados por el grupo y pragmatismo. Como nuevos factores que pueden ser fuentes de sesgo se encuentran: colonialismo religión – Estado (denominado así por los participantes), juicios de valor hacia las teorías sociales y problemas de construcción. De estas posibles fuentes se describen la epistemología de las comunidades de origen y los estilos de aprendizaje, pues presentaron 21% y 18% de las clasificaciones que se hicieron sobre las afirmaciones. Las descripciones aparecen en la tabla 16.

Al revisar las afirmaciones realizadas en el grupo focal se encuentran los siguientes factores que pueden ser potenciales fuentes de sesgo y que se relacionan con los reportados en la literatura: armonía, comunalismo, creencias, epistemología de las comunidades de origen, estilos de aprendizaje, experiencias más frecuentes en un grupo que en otro, familisimo, forma en que se toman las decisiones, objetos poco conocidos o usados por el grupo, oralidad, personalismo y perspectiva social del tiempo. Como posibles nuevas fuentes de sesgo se encontraron: colonialismo religión – Estado, escuela tradicional, juicios de valor hacia teorías sociales y problemas de construcción. Debe mencionarse que hubo cuatro afirmaciones que no fue posible catalogar ni entre los sesgos propuestos por medio de la revisión en la literatura, ni como nuevos. De todos estos posibles factores que pueden estar relacionados con un posible sesgo se describirán la epistemología de las comunidades de origen y los problemas de construcción, pues presentaron porcentajes de 21.4% y 11.9% de las categorizaciones hechas sobre las afirmaciones. Estas afirmaciones también se presentan en la tabla 16.

Tabla 16. Afirmaciones de los participantes en la sesión de análisis de sesgo de la prueba de Sociales.

Fuente potencial de sesgo o aspecto	Análisis	Afirmaciones
Epistemología de las comunidades de origen	Revisión individual de ítems	<p><i>“en las comunidades indígenas el mundo se concibe diferente, al igual que el conocimiento que se adquiere diferenciado por género”, “El ítem resulta más claro para los miembros de sociedades que diferencian las explicaciones científicas y las explicaciones religiosas.”, “El ítem favorece a los miembros de una misma sociedad en la que se diferencia entre “privado” y “público”. Específicamente en la opción “C” aparece un sesgo que no proviene de las creencias y que favorece a quienes comparten tal diferenciación por la manera de ver las cosas en sus comunidades de origen.”, “El concepto de tierra no es una cosa que se puede vender y comprar. La tierra es un ser vivo “la madre”...”, “La revolución francesa es un acontecimiento importante para la cultura occidental.” (aunque esto se afirmó sobre un ítem que favorecía a los indígenas), “Como primera medida las nociones de DDHH (derechos humanos) y democracia contemporáneas no son términos y conceptos propios de las culturas indígenas, ya que estos grupos cuentan con formas de gobierno y participación diferentes a los inscritos en el Estado social de derecho.”, “establece categorías propias del conocimiento de los grupos no indígenas, que no se refieren a la cosmovisión de grupos o comunidades indígenas”, “el fracaso se presenta en los relatos míticos donde el error no siempre remite al fracaso” y “el concepto iglesia se asocia en lo occidental a religión, es posible que los indígenas no lo asocien tan directamente”. Igualmente, respecto al ítem a partir del cual se hizo una afirmación sobre los DDHH (Derechos Humanos) un participante señaló: “La pregunta está bien planteada. Hace referencia a los discursos de derechos humanos y democráticos contemporáneos y, en base a lo que estos discursos afirman, evalúa la comprensión de los mismos aplicada a una caso concreto. Puede ser que en alguna comunidad las prácticas que se describen en la pregunta se apliquen. Sin embargo, la pregunta es clara al plantear que la respuesta debe derivar de lo que afirma la teoría de los derechos humanos y la democracia contemporánea y no en la práctica del pueblo del que la persona proviene”.</i></p>

Fuente potencial de sesgo o aspecto	Análisis	Afirmaciones
Grupo focal		<p>“porque pueda que ellos le puedan hablar del concepto de libertad, no sé si ellos tengan una palabra que se equipare a la libertad en español”, “el concepto de espacios es distinto, para nosotros es más de longitud, para ellos también tiene el concepto de abajo y arriba y medio entonces ahí también hay un problema de cosmovisión... tienen un sentido de orientación buenísimo en la vida practica también como los campesinos, uno va a una chacra que es el sistema donde se cultiva uno se pierde ahí, ellos son muy buenos en las direcciones. También está en algunas culturas indígenas plasmado en las construcciones que hacen, por ejemplo en la maloca con cuatro palos y eso significa los puntos cardinales”, “preguntando qué es violencia intrafamiliar... tuve la oportunidad de ir a Sibundoy, yo cambie mi método de trabajo porque me pareció que había una violencia entre los indígenas que tenía miedo de no saberla manejar... en Sibundoy las comunidades tienen una relación más vieja con lo nuestro y encontré una violencia del hombre contra la mujer, extrema, ... Si la mujer iba detrás de un hombre eso implica que es la esposa, si el hombre va de primeras puede ser porque él lleva el machete y va a defender, hay que preguntarse eso, como nosotros vamos con nuestra cultura, opinamos sobre los hechos pero cuando uno está en una comunidad distinta hay que cambiar su punto de vista... yo descubro que en una canoa todos los hombres están alrededor de las mujeres y los niños en el interior, esa vez un poco de avispa se entró a la canoa y la barrera para las mueres era el hombre, ellos también tienen una lógica y hay que descubrir la lógica que tienen ellos para entender y ver el sentido en donde para uno no tiene sentido.”, “uno se puede preguntar qué es fracaso. Para nosotros ahora lo importante es ser popular, pero por ejemplo en los relatos míticos, que se supone lo sagrado y lo que debe ser, ellos ven que hay seres importantes que fracasan y vuelven y lo intentan, el fracaso simplemente es algo natural, un paso en la vida, por ejemplo los de la sierra, hay un personaje mítico que le dieron a escoger entre nueve mujeres, escogió la blanca y resulta que la blanca no le dio nada a la tierra y de últimas a la más negra que fue con la que el cultivo fue el mejor, eso quiere decir que fracasó ocho veces y persistió hasta obtener lo que quería. Por ejemplo, en los Tikuna hay un tipo que sabe las cosas y que las entiende, el otro está catalogado entre los indígenas como loco, pero a partir de las embarradas de él y su hermano se van construyendo otras cosas, serían las fuerzas cósmicas que dan origen a las culturas, son distintos pero en la vida se necesitan, sin esos dos no habría esa dinámica, para nosotros el fracasado no cuenta porque da pena.”, “allí se considera un mundo totalmente diferente, qué es lo que significa el fracaso, para mí como mujer indígena ¿el fracaso es qué? ¿Quedar embarazada a los 15 años? De pronto puede ser un fracaso para una no indígena. Yo estaba en una comunidad y un tipo me decía ¿usted qué? se va a morir y no tiene hijos ni nada, usted ya no sirve para nada, usted ya está vieja. Hay que entender que allá a los 15 años ya tienen su primer hijo, hay que entender la cosmovisión...”, “va hacia el concepto de libertad e igualdad que podrían tener los diferentes grupos, en la educación formal.. la libertad se define casi que indefinible”, “hombre blanco y la independencia de los Estados Unidos, lo que significa para nosotros no puede significar lo mismo para los indígenas... yo digo que instauró el sistema democrático moderno occidental, porque seguimos jugando a que lo único moderno es lo occidental, a mi si me parece que hay un sesgo aquí porque lo moderno es lo occidental, lo europeo, blanco y patriarcal.”-aunque esta afirmación se hizo sobre un ítem que favorecía a los indígenas-, “hay unas escuelas de pensamiento que son favorables a la libertad del mercado y otras que dicen que lo más importante es la igualdad...”, “esta gente piensa que el pensamiento científico es provisional y que el pensamiento de ellos si se va a mantener.”, “según las estadísticas pensaba que la “D” para los indígenas es razonable, lo que nuestro saber ancestral dice no es pasajero, pero la ciencia si pasará...”, “esa diferenciación entre privado y público en ciertas comunidades no sé cómo sería, pero creo que tiene un sesgo a favor de ciertas comunidades que manejan esos conceptos porque en las comunidades de origen algunas no diferencian entre privado y público.”, “las comunidades indígenas tienen su forma de gobierno, de tomar decisiones y de ejercer la justicia, incluso tienen la posibilidad de legislarse y resolver sus problemas, otra cosa es que se mete el concepto de derechos humanos y democracia contemporánea, eso sí es limitante como indígena, porque si mi sistema de gobierno no funciona, así pueda que lo reconozca la Constitución, pero tiene forma independiente de funcionar.” y “la concepción de tierra es diferente, la tierra para nosotros es una noción utilitarista, ellos la utilizan dentro de parámetros, no es arbitrario, para muchos como los de la Sierra Nevada de Santa Marta es la madre, un ser vivo, hay que respetarla, la tierra en el Amazonas es infértil a pesar de que es asombroso que haya esa vegetación y fauna sorprendente, eso se debe a una cooperación que hay allí, es muy frágil, había un problema porque para ellos la tierra es comunal y tienen su pedazo de tierra lo que implica es que no se puede vender simplemente se rota, entonces esto fue de alguien y sin embargo no se vende, por otro lado la agricultura de subsistencia... sin embargo ellos también siembran para los animales...”. Además sobre un ítem que favorecía a los</p>

Fuente potencial de sesgo o aspecto	Análisis	Afirmaciones
Estilos de aprendizaje	Revisión individual de ítems	indígenas un participante afirmó: “no me pareció que favorecía a uno o a otro porque es muy preciso “la revolución francesa”, la evolución de la monarquía y la republica...en principio yo creería que no hay sesgo...”.
Problemas de Construcción	Grupo Focal	“El estilo de aprendizaje de las comunidades indígenas es diferente.”, “Los estilos de aprendizaje, el fuero indígena con el cual han crecido las comunidades indígenas.”, “una dificultad en cuanto a los estilos de aprendizaje, ya que se requiere un conocimiento con relación al manejo de cuadrantes y grados” (aunque este comentario hacía referencia a un ítem que favorecía a los indígenas) y “Estilos de aprendizaje el concepto de libertad es distinto en las comunidades indígenas”.
Ítems sin DIF	Revisión individual de ítems	“la pregunta está mal planteada porque hay muchos conceptos que no tienen un significado unívoco”, “la posibilidad de respuesta es tan abierta...”, “hay información adicional que lo distrae a uno.”, “se me ocurren las habilidades de comprensión de lectura, habría que mirar este ítem comparado con todas las otras preguntas de la prueba para ver si es recurrente.” y “no sé que evalúa este ítem.”.
	Grupo focal	“Debido al estilo de aprendizaje y formas de las comunidades indígenas los estudiantes pertenecientes a estas etnias están más en conflicto con las diferentes técnicas y lenguaje utilizado.”, “los estilos de aprendizaje, la epistemología de las comunidades de antes, pues, los indígenas tienen creencias diferentes al igual que su forma de ver el mundo.”, “Estilos de aprendizaje conceptos que no manejan las comunidades indígenas, teoría marxista geografía crítica.”, “Epistemología de comunidad de origen y pragmatismo ya que la respuesta a la pregunta puede tener un tinte cultural.”, “En el ítem hay un sesgo relacionado con la visión urbana del mundo desde la que ha sido concebido. La fuente de sesgo en este caso es la visión compartida por las comunidades de origen y las creencias.”, “Epistemología de las comunidades de origen, estilos de aprendizaje.” y “Cabe anotar que la relación con el territorio para los grupos indígenas se encuentra atravesado por visiones cosmológicas, además de las relaciones que se establezcan allí por la comunidad.”.
Cómo mejorar los ítems	Revisión individual de ítems	“en la igualación que se le hace al Estado y al mercado, ponerlos en un rango de importancia equivalente, me parecía que podría haber dificultad por el papel que juegan.”, “Es la percepción que ellos tienen de esos conceptos, se van hacia el común denominador cuando no conocen algo o concepto.” y “está mal planteada la pregunta.”.
	Grupo focal	“Que la pregunta no se enfoque a opiniones personales, aunque es necesario inicialmente analizar lo que se desea medir.”, “Cambiar la opción “C” por una en que no se haga referencia a lo “privado””, “Se podría cambiar este ítem por otro que evaluara la lógica del mercado sin involucrar la noción de tierras.”, “sería mejor no involucrar la distinción entre la religión y la ciencia, pues esta es una sofisticación de aquella.”, “Omitir la referencia a los EEUU, Francia y México.”, “Planteando situaciones que incluyan diferentes grupos de la población y sean comunes a éstos. A la vez de plantear situaciones propias de estas comunidades.” y “Especificar el tipo de iglesia y su relación de los hombres occidentales y los dioses.”.
	Grupo focal	“No hay que favorecer con mapas adicionales en los ítems.”, “yo omitiría la referencia a Estados Unidos, Francia y México y quitaría la opción “A” por lo que me parece ofensiva.”, “hay que evaluar con otra cosa que no sea la tierra.” y “quitar lo de la superstición.”.

En términos generales se observa que pueden existir tres fuentes posibles de sesgo: estilos de aprendizaje, epistemología de las comunidades de origen y problemas de construcción, siendo más claras estas últimas, pues presentan menos afirmaciones en los ítems sin DIF. Un ejemplo de un ítem con DIF que favorece a los no indígenas se presenta en la figura 11, a éste hacen referencia aquellos comentarios sobre la violencia intrafamiliar y lo privado y lo público. En dicho ítem la opción correcta es la B, en donde la mayoría de indígenas se inclinan más por la C y los no indígenas por la B. De todas formas, la epistemología de las comunidades de origen y los estilos de aprendizaje como posibles

fuentes de sesgo cultural deben verse con precaución y siempre analizarlas a la luz de si están o no relacionadas con lo que pretende medir el ítem.

58. En el barrio donde yo vivo un grupo de personas ha pensado impulsar las siguientes medidas para disminuir la violencia intrafamiliar: espiar a los vecinos, aprender a los infractores a partir de rumores y flagelar a los sospechosos en la plaza. Desde el punto de vista de los derechos humanos y de las nociones de democracia contemporáneas, una posición es la siguiente

- A. se debe ser más radical ante un tema tan grave y proponer pena de muerte.
- B. la justicia privada engendrará mayor injusticia sin solucionar nada.
- C. el asunto de la violencia familiar es irresoluble porque es cosa privada.
- D. en este caso cada cual debe defenderse como pueda sin salir de casa.

Figura 11. Ítem con DIF en la prueba de Sociales que favorece a los no indígenas.

Posibles fuentes de sesgo transversales y nuevas fuentes de sesgo. De acuerdo con los resultados observados hasta aquí, existen tres factores potenciales de sesgo que podrían considerarse transversales, ellos son: las experiencias más comunes en un grupo que en otro, los problemas de construcción y epistemología de las comunidades de origen. No obstante, debe tenerse en cuenta que pueden presentarse en ítems sin DIF, razón por la cual se recomienda que si estos factores van a ser tomados como un aspecto en la revisión de los ítems, se debe siempre analizar si son lo suficientemente claros y no están relacionados con lo que se pretende medir, para que puedan ser considerados como la causa por la cual un ítem presenta DIF.

Por otra parte, aunque no fue considerado un factor potencial de sesgo, los participantes manifestaron, especialmente en los análisis de ítems de Biología y Matemáticas, que los ítems que contienen gráficas para su interpretación y posterior respuesta a la pregunta, suelen ser más difíciles para los indígenas. Sin embargo, si esto está relacionado con lo que pretende medir la prueba no debe ser relacionado con sesgo, sino con una diferencia real entre los dos grupos.

A continuación se definen o redefinen los tres aspectos considerados transversales y los nuevos factores propuestos como fuentes de sesgo; adicionalmente se mencionan dos características de los ítems que deben ser consideradas al momento de revisarlos.

Experiencias más frecuentes en un grupo que en otro. Se caracteriza por exposiciones a actividades, temas o contextos que son particularmente más propios en un grupo que en otro y que están presentes diferencialmente en el diario vivir de los evaluados. En el caso de los indígenas se destaca el contacto con animales y para los no indígenas se encuentran actividades como hacer viajes por carretera (pasear) que permiten conocer mejor las convenciones y nombres dados a ciertos lugares (capital, municipio, población, peaje), realizar experimentos y estar en contacto de forma frecuente con deportes como el fútbol. No obstante, si estos aspectos son relevantes para lo que se desea medir, no deben ser considerados como factores potenciales de sesgo.

Problemas de construcción. Se refiere a los siguientes aspectos:

- a. Extensión y complejidad de los pasajes que sirven de contextos a las preguntas: si éstos son muy largos, al considerar el español como segunda lengua, puede aumentar la dificultad para algunos grupos.
- b. Opciones de respuesta desbalanceadas: una de las opciones de respuesta es de una categoría diferente de las otras, por ejemplo: a) una de las opciones es de una categoría semántica diferente al compararse con las otras opciones, por ejemplo, sólo una de las opciones se refiere a un movimiento literario y las demás provienen del sentido común, o una de ellas se refiere a un sujeto y las demás a objetos; b) todas tienen una palabra en particular y otra no, la cual puede ser o no la respuesta correcta, sin embargo hace que los evaluados se desvíen hacia ella; y c) opciones mucho más largas o cortas que las otras.
- c. La respuesta correcta no se deriva del contexto y enunciado, sino de un conocimiento adicional, incluso, éstos pueden llevar a la selección de una respuesta errónea por comprensión de lectura, cuando no se está evaluando este atributo.

- d. Palabras que están en el enunciado, se escriben tal cual en las opciones de respuesta y eso hace que los evaluados se inclinen a escogerla, favoreciéndolos cuando esa es la respuesta correcta y desfavoreciéndolos cuando no.
- e. Conceptos que no tienen un significado unívoco y que hace que la posibilidad de respuesta sea muy abierta.
- f. Convenciones usadas en las gráficas: existen convenciones en las gráficas que exigen tener un conocimiento o habilidad adicional que no se encuentran relacionadas con lo que se está midiendo o pueden tomarse como cualquier valor extremo, por ejemplo, no es claro qué significa “0% digestión” o “100% digestión” pues el primero puede indicar que no hubo absolutamente digestión mientras que 100% señala que la hubo completamente o viceversa, es decir, el primero indica que se digirió todo y no quedó nada y el segundo que queda todo por digerir.

Epistemología de las comunidades de origen. De acuerdo con lo revisado en la literatura y lo expuesto hasta el momento, esta posible fuente de sesgo se podría presentar cuando la forma en que se realizan las preguntas y su contenido son contrarios o diferentes a cómo se concibe el mundo y el conocimiento. Específicamente se relaciona con aspectos como: costumbres que son más comunes en un grupo (p. ej. matrimonios entre parientes cercanos), diferente concepto de familia, formas de tratar las enfermedades, conceptos que no tienen una palabra equivalente en los idiomas de los grupos que se están comparando, los conceptos de orientación en el espacio, el manejo de la economía, las formas de gobierno y la forma en que se imparte justicia en diferentes grupos culturales, las relaciones de género, el manejo del fracaso, el pensamiento científico, la religión y la concepción de la naturaleza, entre otros.

Colonialismo religión – Estado. Se refiere a aspectos relacionados con la época de la colonia y la experiencia que tuvieron los diferentes grupos culturales a partir de ese momento en relación con la iglesia y el Estado. Este factor es respaldado por afirmaciones como: “a este texto le vi problemas por el contenido ya que hablan de colonialismo de los cronistas”, “qué tanto intervienen las comunidades religiosas en la generación de estos mitos, puede ser por la interferencia del catolicismo, podría ser por las ideas católicas.”,

“aunque la pregunta es clara, debido a todos los años de evangelización, ellos podrían elegir una u otra dependiendo de lo que tengan allí en cuanto a esa evangelización, si les prohibían la lengua, los obligaban a ir a misa, a las modificaciones de la cultura, puede ser que dependiendo de la personalidad de cada uno y de si está arraigado o no, si afecto a mi comunidad o no.”, “estaba pensando que efectivamente la separación (iglesia del Estado) de ésta permite el desarrollo y de alguna manera permite la libertad del mercado y estimula lo económico, no es correcta de acuerdo al marco que se le da al estudiante...”, “según entiendo, en muchos grupos culturales no se separan el poder político del religioso, eso hace que el espectro sea muy estrecho en la concepción del ítem porque supone que se ha diferenciado entre esos poderes”, “si uno observa las estadísticas parece que indicara que responden desde el punto de vista como ven al otro, parece que ellos respondieran desde lo que conocen de nuestra sociedad pero también reflejan esas tendencias que ellos responden desde las concepciones que ellos tienen, de cómo fue impuesta la religión y cómo se puede interpretar ahora, cuál es la relación con las diferentes instituciones y la sociedad. Una forma de denominar el factor del imperio romano es el colonialismo, porque las comunidades indígenas acá y en México si tienen ese hecho histórico marcado y presente y esa respuesta tiene que ver con el proceso de colonización durante el tiempo y que obedece a una condición particular de América latina y en ciertas regiones, en específico mexicanas, creo que tenderían a colocar esa.”. Debe mencionarse que estas tres últimas afirmaciones hacen referencia al ítem de la figura 12, cuya respuesta correcta si bien es la B, un 14.8% de los indígenas y un 8% de los no indígenas se inclinan por la opción C.

73. Según las interpretaciones sociales más sólidas en torno a la modernidad y el desarrollo de la democracia, la separación de la iglesia y del Estado, como la que existe en los Estados Unidos, Francia o México, puede ser considerada como un paso necesario para
- A. acabar con las ilusiones y supersticiones religiosas.
 - B. favorecer la libertad, la igualdad y los derechos humanos.
 - C. imponer una religión desde el Estado.
 - D. estimular el libre desarrollo de los mercados.

Figura 12. Ítem con DIF de la prueba de Sociales y que favorece a los no indígenas.

Tecnicismos. Se refiere al uso de palabras que requieren un conocimiento o una habilidad diferente al atributo que se está pretendiendo medir y que dificulta la respuesta a un grupo particular. Entre ellas pueden estar: rumiar, glándula sebácea, neumococo, branquial, sustancias tóxicas, rango, entre otros. Esta posible fuente de sesgo se evidencia en afirmaciones como: “*si algún termino es alejado o técnico.*”, “*Falta ver si no entienden la palabra rango.*”, “*La palabra rumiar no es un término que evalúe la pregunta y hace fácil descartarla para el que conoce el significado.*”, “*Uso de tecnicismos- neumococo o branquiales.*”, “*Lenguaje es muy técnico*”, “*el secretar sustancias y difundir hormonas es un sinónimo en la que una palabra posee un tecnicismo y la otra no.*” y “*Yo no sé que es turberas*”. Esta posible fuente de sesgo sólo aplica siempre y cuando el tecnicismo no se considere relevante para lo que se quiere medir.

Escuela tradicional. Para algunos grupos culturales es factible que mantengan una enseñanza de tipo declarativo considerada tradicional y que hace que en ítems de ese tipo se vean favorecidos individuos particulares. Esto se encuentra respaldado por afirmaciones como: “*...si están agrupados por institución educativa podría ser que sea por la institución de enseñanza del tipo declarativo, porque las preguntas en las cuales les está yendo bien son en las que se supone debería haber escuela, pero una que se ha dejado de enseñar, es*

el conocimiento del tipo declarativo... es llamativo que les vaya mejor en un cierto tipo de preguntas porque la hipótesis que uno tienen es que la educación que reciben es más folclórica y según los visto es todo lo contrario, en ciencias se me hace que es más tradicional la educación, más teórica que práctica. Como el estilo de la práctica con los ejercicios de Baldor que uno hacia setecientos ejercicios y aprendía, parece más ese estilo de enseñanza. Cuando uno ve el contexto de la pregunta no es cotidiano, ya que “los tubos seminíferos de los testículos del perro”, no es tan común ni entre los licenciados.”, “algunos autores señalan que se aprende mejor bajo unos estilos de enseñanza, cuando mi forma de aprender es catedrática y el profesor es catedrático pues voy a aprender más rápido, pero si mi profesor es autónomo y yo soy más dependiente pues no voy a aprender muy bien...”, “uno se da cuenta que en la escuela tradicional se muestran los conceptos de la ciencia como cosas ya dadas, célula es tal cosa, etc. Una de las falencias que ha demostrado la didáctica de la ciencia es que cuando la escuela es tradicional no hay gráficas, no hay experimentos, yo veo que cuando el ítem es de gráficas o interpretación de textos como solo está contextualizado con la matemática, la interacción con ciencia es muy compleja”, “en otras que educar es promover ciertas nociones culturales sobre otras, en realidad lo que aquí hay es la imposición de ciertas nociones culturales sobre otras”, “la educación va a tender a enseñarle a uno en términos de sociedad en general y no en una sociedad específica.” y “El concepto de escuela tradicional es muy clásico y aunque esté en el texto realmente pasa que la educación es muy tradicional esta pregunta debería favorecer a los indígenas, no por la cultura sino porque el dato es tal y se responde.”. Sin embargo debe tenerse en cuenta que aunque esta variable puede introducir DIF, es probable que se deba a una diferencia real entre la educación que reciben los indígenas y los no indígenas, pues de acuerdo con la revisión teórica, existe la etnoeducación para los indígenas y, según las afirmaciones aquí expuestas, también asisten a internados dirigidos por confesiones religiosas; ambos tipos de educación pueden llegar a ser diferentes en los no indígenas.

Juicios de valor hacia teorías sociales. Se refiere a los juicios subjetivos que deben elaborar los evaluados para responder una pregunta de forma correcta sobre algunos

conceptos, como igualdad y libertad, y teorías sociales. Esta posible fuente de sesgo se sustenta en afirmaciones como *“más que evaluar una competencia argumentativa está juzgando un dilema moral, es decir, usted le da prevalencia a la igualdad o a la libertad y efectivamente hay personas que deben dar a alguno de los dos derechos, desde el punto de vista moderno uno no le daría prevalencia a alguna de las dos sino que las igualaría, en términos de que tengo el mismo derecho a ser libre como de tener las mismas oportunidades frente a los demás”* y *“Pareciera más que ésta fuera una pregunta propositiva porque la argumentativa va con causas y efectos, pero aquí está sumamente subjetiva”*.

Otros aspectos importantes en los ítems. Aunque no fueron mencionados como factores que favorecieran a un grupo en particular, sí es importante considerar algunos aspectos relacionados con términos, palabras o gráficas que pueden incomodar a algún grupo de la población. Por ejemplo el uso de palabras “gorda” y más refiriéndose a alguna profesión en particular, esto se observa en afirmaciones realizadas por los participantes como: *“aparte al final hay una frase ... es la gorda de español, un texto con esa frase no puede ir en una prueba..., en primer lugar porque se está refiriendo a un tipo de persona...y decir que los profesores de español no enseñan nada quedaría como una interpretación más allá del texto, como que sólo me enseñó adjetivos y artículos, eso sí sería un sesgo en el texto, además la forma en que se refiere a una persona.”*. También se debe tener en cuenta que las gráficas o figuras que se ponen en los ítems no sean ofensivas para grupos particulares de personas, así por ejemplo uno de los participantes afirmó que en ocasiones para los indígenas existen logotipos de compañías que denotan aspectos maléficos. Finalmente, se observaron aspectos de diagramación en los ítems que podían hacer que grupos particulares de estudiantes se dirigieran a aquellas a las que, por ejemplo, les falta un punto mientras que a las demás no.

DISCUSIÓN Y CONCLUSIONES

Este trabajo presentó los resultados de la segunda fase del proyecto “Identificación de Ítems con Sesgo Cultural en las Pruebas de los Exámenes de Estado en Colombia”, cuyo objetivo fue establecer posibles fuentes de sesgo cultural en las pruebas del examen SABER 11°, con el fin de minimizarlas a través de propuestas para la construcción de ítems libres de variables irrelevantes a nivel cultural. En la primera parte se expusieron los resultados de la aplicación de la rutina estadística producto de los trabajos de Arias (2008) y Berrío (2008), y en la segunda el análisis substantivo de los ítems que fueron detectados con DIF con el fin de identificar posibles fuentes de sesgo.

La rutina implementada mostró que ambos procedimientos estadísticos (MH y Diferencia de la dificultad) fueron de utilidad y fácilmente aplicables a los datos de las pruebas de SABER 11°; sin embargo hay dos consideraciones que se deben tener en cuenta para un uso más práctico y preciso con otras aplicaciones de este examen. La primera consideración está relacionada con la calidad de la información de la que se dispone para hacer una investigación de esta clase, específicamente con la validación de los datos proporcionados por los evaluados en el formulario de inscripción. Si bien, es necesario hacer un proceso de verificación de los datos relacionados con la etnia de los examinados, dicho proceso fue bastante dispendioso por dos razones. Primero, aunque se contaba con una base de planteles educativos proporcionada por el ICFES en la que se encontraba el código ICFES⁸ del colegio y el código DANE⁹ del mismo, su validación requirió de una gran inversión de tiempo. El ICFES maneja las bases de datos de SABER 11° con su código y el MEN con el código DANE, y no siempre coincidía el dispuesto en la base del ICFES como código DANE con el de las bases del MEN o éste no era válido o no estaba

⁸ Código asignado por el ICFES a cada EE que presenta a sus estudiantes al examen SABER 11° y que viene dado por jornada. Así, si una sede física de un EE cuenta con dos jornadas, habrá un código ICFES para cada una de ellas.

⁹ Código asignado por el Departamento Administrativo Nacional de Estadística (DANE) a los EE y sus sedes.

actualizado. Segundo, si bien el ICFES y el MEN proporcionaron para 2006 los datos con el número de identificación de los evaluados y matriculados en 11° y ciclo 6, no fue el caso para 2007, porque se trataba de información confidencial. Aunque este aspecto es por demás entendible, dificultó que los criterios, para la validación de qué examinados eran indígenas, fueran los mismos para las dos aplicaciones analizadas.

Estas dificultades pueden verse superadas para ejercicios posteriores con otras aplicaciones de SABER 11°, a través de dos formas. La primera es que el ICFES, el MEN y el DANE lleguen a un acuerdo para tener una homologación completa de los códigos DANE e ICFES en todas las bases de datos que implique el manejo de resultados de evaluaciones a gran escala y que dicha homologación sea dinámica, es decir, que cada vez que se realicen cambios en alguno de los códigos, las bases que los contengan sean actualizadas en su totalidad. La segunda posibilidad es que, dado que el ICFES y el MEN trabajan en diferentes procesos conjuntamente en términos de insumos de información, compartir datos entre ambas entidades es más fácil, por ejemplo, el número del documento de identificación de los evaluados y matriculados en 11° y ciclo seis, permitiría el cruce directamente de las bases de datos a través de estas dos variables.

La segunda consideración se refiere a la aplicación de los procedimientos de diferencia de la dificultad y MH en los datos de SABER 11°, para la cual se deben tener en cuenta los siguientes puntos:

1. En ninguna de las pruebas analizadas en este trabajo se encontró evidencia satisfactoria de unidimensionalidad de las pruebas analizadas, dado que no presentaron un alto grado de ella. Debe decirse que éste es un aspecto que es de competencia del ICFES, en cuanto a la definición de los constructos que son objeto de medida de los instrumentos que desarrolla y sus implicaciones en la construcción de ítems.
2. En cuanto al impacto, se encontraron mayores diferencias entre los grupos de indígenas y no indígenas en la segunda aplicación de 2006. Esto indica que las brechas son mayores en este caso, aunque no moderadas ni grandes.

3. Las pruebas estadísticas de ambos procedimientos detectaron más de un 34.92% de ítems en promedio en cada una de las pruebas analizadas y tomando en cuenta las 21 muestras. Sin embargo, dicho porcentaje cayó a 25.79% o menos al implementar las métricas de los procedimientos. Esto concuerda con los resultados de Arias (2008) y Berrío (2008), en donde se encontró que las métricas ofrecen un mayor control del error tipo I (falsos positivos) que las pruebas estadísticas.
4. Sin importar el procedimiento, hubo una mayor tasa de detecciones con la prueba estadística en la segunda aplicación del 2006 comparada con la primera de 2007, exceptuando en la prueba de Lenguaje. Este comportamiento podría deberse a la gran diferencia de evaluados entre las dos aplicaciones, pues la segunda aplicación de 2006 es seis veces mayor que 2007. En este sentido, Herrera (2005) encontró que con tamaños de muestra grandes la tasa de falsos positivos aumenta, por lo menos con la prueba esta estadística del MH.
5. Respecto a las métricas, el Δ_{MH} presenta un promedio mayor de porcentaje de ítems detectados a través de las muestras en la primera aplicación del 2007, exceptuando la prueba de Sociales. Esto corresponde con lo encontrado por Arias (2008), puesto que a mayores diferencias entre los tamaños de los grupos de comparación (razón de tamaños) mayor control del error tipo I y menor potencia. Respecto a la métrica D , se observa que hubo una mayor tasa de detecciones en la primera aplicación de 2006, exceptuando las pruebas de Biología y Lenguaje, lo que también concuerda con los hallazgos de Berrío (2008), en donde a diferencias más extremas entre los grupos focal y de referencia se presenta un mayor error tipo I.
6. En la mayoría de comparaciones entre las muestras sin impacto y con impacto pequeño, el porcentaje de ítems detectados tendió a ser mayor cuando hubo impacto pequeño, destacándose la prueba de Sociales aplicada en 2007. Esto también coincide con los estudios de Arias (2008) y Berrío (2008), en donde se encontró que el impacto, específicamente la diferencia en la media del atributo

entre los dos grupos, afecta la detección de ítems con DIF presentándose mayor error tipo I.

7. Debido a que los ítems que presentaron desajuste respecto al modelo de Rasch, fueron pocos, no es posible señalar si hay mayor o menor cantidad de ítems detectados con DIF en la presencia de ajuste o desajuste.
8. Respecto a los parámetros de los ítems, se encontró que aquellos sugeridos para revisar por haber sido detectados por al menos una de las métricas en alguna muestra, se caracterizaron por ser ítems de dificultad media y discriminación entre media y alta. Esto concuerda parcialmente con lo encontrado con Arias (2008), en donde los ítems con discriminaciones altas (mayores a 1.4) y fáciles (dificultad entre -1.55 y -1) mostraron mayores tasas de error tipo I utilizando el MH; sin embargo dicha autora afirma que el Δ_{MH} no se deja afectar por la dificultad o la discriminación de los ítems.
9. Llama la atención que existe un número similar de ítems que favorecen a ambos grupos, al contrario de lo que se esperaba al considerar como grupo focal a los indígenas. En este sentido, podría decirse que existe cierta compensación entre los ítems que favorecen a un grupo y a otro, haciendo que las diferencias, por ejemplo, en términos de respuestas correctas no sean muy amplias, aspecto que se evidenció en que sólo hubo impacto pequeño.
10. Aunque en términos generales, las pruebas estadísticas mostraron buenos grados de acuerdo, las métricas no. Probablemente esto se deba a que las pruebas estadísticas de ambos procedimientos detectaron más ítems que las métricas, facilitando el acuerdo entre ellas.

Para resumir y teniendo en cuenta los anteriores puntos, la rutina sugerida al ICFES para la detección de ítems con DIF entre evaluados indígenas y no indígenas tendría las siguientes fases:

1. Preparación de las bases de datos, que incluye la depuración y validación de la información registrada en el formulario de inscripción sobre la etnia de los evaluados.

2. Detección de los ítems con DIF a través de los procedimientos de diferencia de la dificultad y MH con sus pruebas estadísticas y métricas expuestas en el presente trabajo. Se sugiere que sean estos dos procedimientos y sus métricas porque, aunque pueden verse afectadas por variables como el impacto, el tamaño de los grupos, la razón de tamaños y el desajuste del modelo, las características de los datos reales a través de estas variables permiten un buen desempeño de los dos procedimientos. Así, se encontró que sólo hubo impacto pequeño, por lo que cualquier efecto de esta variable sobre la identificación positiva de ítems sin DIF (error tipo 1) se espera que no sea muy grande. También, las posibles consecuencias de tener tamaños de muestra muy grandes y razones de tamaño extremas entre los grupos, pueden verse controladas por el uso de las métricas y porque en razones menores a 250 los procedimientos tienden a presentar un desempeño adecuado (Arias, 2008; Berrío, 2008). Finalmente, dado que la gran mayoría de ítems se encuentran ajustados a un modelo de Rasch, tomando como medida el infit y el outfit, el resultado que un posible desajuste pudiera tener sobre la detección de ítems con DIF con los procedimientos aquí usados se ve minimizado. Esto último es de suma importancia para efectos prácticos, puesto que si se considera que el DIF que se presenta por excelencia bajo el modelo de un parámetro es el uniforme (Berrío, 2008) y que un modelo de Rasch se ajusta para la gran mayoría de los ítems, no habría razón para pensar que se estaría presentando DIF no uniforme. Respecto a sí este procedimiento se debe usar o no con Bootstrapping, la sugerencia es que se tenga en cuenta el tiempo que se dispone para hacer los análisis de DIF y si amerita la pena hacerlo, puesto que este tipo de procedimiento en conjunto con las rutinas del MH y de la Diferencia de la dificultad puede tomar algún tiempo. En el presente trabajo, por ser una investigación, se usó este tipo de muestreo para observar la estabilidad del procedimiento; sin embargo para aplicaciones en tiempo real, no se recomienda debido a la inversión de tiempo que se requiere y a que no existe un criterio para señalar en qué porcentaje de muestras debe ser detectado un ítem para que sea declarado como con DIF.

3. Selección de los ítems que serán revisados para identificar si el DIF que presentan se debe a diferencias reales o a variables irrelevantes que están siendo medidas por el mismo. Esta selección debe ser realizada basándose exclusivamente en la métrica de cada procedimiento por aparte, es decir, si un ítem no es detectado por la métrica de un procedimiento, pero sí por la métrica del otro, igualmente debe ser revisado. Se hace esta sugerencia teniendo en cuenta que el acuerdo entre las métricas es bajo y el alto control del error tipo I que exhiben las mismas.

Como se mencionó anteriormente, la detección de ítems con DIF suele ser considerada el primer paso para la identificación de posibles fuentes de sesgo en los mismos. Al culminar el paso de detección, se hizo necesaria la implementación de estrategias para la identificación de posibles fuentes de sesgo cultural. Dichas estrategias, suelen ser de corte cualitativo y para la segunda parte de este trabajo se escogió usar grupos focales con expertos en las poblaciones que se estaban comparando y en el atributo que mide la prueba, los cuales se llevaron a cabo después de una revisión individual de los ítems.

Aunque si bien, la revisión por expertos es el método más comúnmente usado (Ercikan et al, 2010), los grupos focales con expertos también pueden usarse para este tipo de tareas (Snijkers, 2002). La idea de esta forma de trabajo era llevar a cabo un análisis a profundidad de los ítems de la prueba, los resultados estadísticos de DIF y una evaluación con expertos en el área que permitiera identificar y corregir las posibles fuentes de sesgo, tal y como lo mencionan Hambleton et. al (1993, citados por Arias, 2008). El análisis de estos datos se llevó a través del ATLAS.ti de una manera sencilla, en la que se categorizaron las diferentes afirmaciones de los participantes en posibles fuentes o factores culturales que podrían estar relacionados con sesgo cultural y que se desprendían de la literatura, o que fueron emergentes en este estudio.

De acuerdo con los resultados del análisis substantivo, se encontraron tres posibles fuentes de sesgo que pueden considerarse transversales a las pruebas analizadas, una de ellas emergente. Así mismo, se encontraron cuatro nuevos factores que, siempre y cuando sean irrelevantes para lo que se pretende medir, pueden considerarse como fuentes de

sesgo. Teniendo en cuenta lo señalado por el ETS (2009), sobre que las fuentes de varianza irrelevante pueden ser clasificadas en: cognitivas, afectivas o físicas, en el presente estudio sólo podrían enmarcarse en las dos primeras categorías.

Las **experiencias más frecuentes** en un grupo que en otro, definidas como la exposición a actividades, temas o contextos que son particularmente más propios de un grupo y que están presentes diferencialmente en el diario vivir de los evaluados, se catalogan como una fuente cognitiva, puesto que las experiencias pueden llevar a que los evaluados adquieran conocimientos o habilidades no relacionados con lo que quiere medir la prueba pero que les facilitan el dar una respuesta correcta. Dichos conocimientos y habilidades no estarían distribuidas equitativamente entre los grupos, por ejemplo el conocimiento en deportes (ETS, 2009). Este aspecto fue evidenciado en ítems de Matemáticas que presentaban un contexto futbolístico. Es de mencionar que factores como las experiencias de vida ya habían sido señaladas como un elemento que puede originar agudas interacciones entre los antecedentes de los evaluados y el contenido de la prueba (Ross & Okabe, 2006). Frente a cómo evitar que este aspecto interfiera con la medición, los participantes sugirieron eliminar información irrelevante en los contextos de los ítems, hablar de temas más generales y definir ciertos aspectos y convenciones que se usan en los textos y gráficos, para que quien esté respondiendo las preguntas tenga todos los elementos necesarios y relevantes para hacerlo de forma correcta.

Los **problemas de construcción** fueron una posible nueva fuente de sesgo encontrada en este estudio, se caracteriza por la extensión y complejidad de los contextos, las opciones de respuesta desbalanceadas, conceptos que no tienen un significado unívoco, entre otros. Estos problemas en general, se refieren más a aspectos técnicos que suelen ser explícitos en manuales y talleres de construcción de ítems. Esta fuente, de acuerdo con el ETS (2009), se catalogaría como cognitiva en la presente investigación, pues puede exigir habilidades para responder correctamente un ítem que no están relacionadas con el atributo que se pretende medir. Básicamente como sugerencia para evitar esta potencial fuente de sesgo, se propone hacer textos más cortos y sencillos, definir conceptos que pueden no ser unívocos y balancear las categorías en las opciones de respuesta. Teniendo en cuenta que Zieky (1993)

señala que algunos de los aspectos aquí mencionados como problemas de construcción (p. ej. ambigüedad) en los ítems se deben considerar como injustos para todos, es necesario aclarar que, según los resultados de la presente investigación, estos aspectos parecen acentuar las diferencias de los comportamientos psicométricos de los ítems entre los grupos que se están comparando. También debe mencionarse que Erickson et al. (2010) afirman que cuando se toman las pruebas en una segunda lengua, se obtienen desempeños más altos en ítems con menor complejidad lingüística, porque si ésta es alta e innecesaria hay un problema que afecta la comparabilidad entre los aprendices de una segunda lengua y los no aprendices (Abedi, Lord, & Plummer, 1995 y Abedi, Leon, Wolf, & Farnsworth, 2008, citados por Erickson et al., 2010).

La **epistemología de las comunidades** de origen se presenta como fuente de varianza irrelevante cuando la forma en que se realizan las preguntas y su contenido son contrarios o diferentes a como un grupo concibe el mundo y el conocimiento. Esta fuente puede catalogarse como fuente emocional, pues algunas preguntas al ser contrarias a la cosmovisión de los grupos pueden ser ofensivas, molestas o controversiales distrayendo al evaluado y haciendo que responda de una forma más emocional que lógica (ETS, 2009). Esto se evidenció en ítems que hablaban de enfermedades o malformaciones que eran consecuencia de posibles prácticas sociales que pueden presentarse en determinados grupos y en ítems que son controversiales como hablar de medidas para disminuir la violencia intrafamiliar. Específicamente de la prueba de Sociales, llama la atención que el componente “el tiempo y las culturas” presenta la mayor cantidad de ítems que favorecen a los no indígenas. Si se tiene en cuenta su definición, de la cual hacen parte las dimensiones de la cultura (científica, tecnológicas y técnicas), las estéticas y expresivas, éticas o integrativas, trascendentes, filosóficas, religiosas o sapienciales (Ortiz, et al. 2007), son estos aspectos los que pueden introducir posibles fuentes de sesgo cultural de acuerdo con lo expresado por los participantes en los grupos focales. Adicionalmente, el ítem de la figura 11 también muestra un ejemplo claro de cómo la epistemología influye en la respuesta a la pregunta, pues para unos grupos culturales la respuesta correcta puede ser la

C, mientras que para otros la B, aspecto que ya había sido señalado por Solano & Nelson (2000).

El **colonialismo religión-Estado** se refiere a aspectos relacionados con la época de la colonia y la experiencia que tuvieron los diferentes grupos culturales a partir de ese momento en relación con la iglesia y el Estado. Esta posible fuente de sesgo puede verse como una fuente emocional, puesto que dadas las condiciones a las que fueron sometidos los indígenas desde la época de la colonia y hasta la Constitución Política de 1991, tiempo en el que estuvieron regidos por la Ley 089 de noviembre de 1890 (República de Colombia, 1890), el contenido de ítems relacionados con aspectos de esta época, por ejemplo, los que hablan de la separación de la iglesia y el Estado, puede resultar molesto u ofensivo para las comunidades indígenas dada su experiencia previa.

También, respecto al ítem de la figura 12 que ilustra el colonialismo religión-Estado como fuente de sesgo cultural, debe señalarse que en dicho caso pudo haber sucedido que los grupos hayan prestado atención a estímulos diferentes. Como Banks (2006) señala, es posible que si se incluye una opción que ilustre un aspecto de la cultura de un grupo, los miembros de éste se inclinarán por ella sin importar si es o no la respuesta correcta. El que haya sido evidenciada esta posible fuente de sesgo, también puede estar relacionada con el hecho de que durante más de 100 años (1890-1991), desde la propuesta de la Iglesia, la educación de los indígenas ayudó a reforzar modelos espacio-temporales, pautas cognitivas, valoraciones, prohibiciones y formas de producción que resquebrajaron las formas propias de su cultura (Rodríguez, 2011).

Los **tecnicismos**, definidos como el uso de palabras que requieren un conocimiento o una habilidad diferente al atributo que se está pretendiendo medir y que dificulta brindar la respuesta correcta a un grupo particular, pueden ser catalogados como una fuente cognitiva si se sigue lo señalado por el ETS (2009), pues hacen referencia a un lenguaje difícil innecesario para responder la prueba. Esta fuente se evidenció específicamente en el uso de palabras como turberas, neumococo, rumiar, etc.

La **escuela tradicional** se refiere a la conservación de un tipo declarativo de enseñanza, que hace que en ítems que demandan ese tipo de conocimiento (declarativo) un grupo se vea favorecido sobre el otro. Este factor de sesgo cultural se enmarcaría como una fuente cognitiva desde la perspectiva del ETS (2009), porque puede que el tipo de enseñanza se relacione con cierta clase de habilidades que no se encuentran distribuidas equitativamente entre los grupos. Esto se evidenció en ítems que exigían recordar elementos que suelen o solían enseñarse de memoria o a través del uso exclusivo de textos en las escuelas. Un punto llamativo en esta investigación con relación a la escuela tradicional y el conocimiento declarativo, es que en estudios como los de da Costa & Araujo (2011), ítems de “acceso y recobro” que de alguna manera están asociados con conocimiento declarativo, fueron más fáciles para el grupo focal (inmigrantes), resultado que también se apreció en la presente investigación. Esto podría llevar a pensar, con la limitación de desconocer las características puntuales de cada cultura en el estudio de da Costa & Araujo, que aspectos que suelen estar más asociados con lo que Cattell denominó inteligencia fluida (p. ej. memoria) son menos sensibles a las diferencias culturales. Esto no es de extrañar, puesto que se supone que la inteligencia fluida es el resultado de la influencia de factores biológicos en el desarrollo intelectual, mientras que la inteligencia cristalizada es la manifestación de la influencia de la experiencia, la educación y la aculturación (Horn & Cattell, 1966). Adicionalmente, da Costa & Araujo (2011) concluyen que los ítems que favorecen a los inmigrantes, están más relacionados con aspectos de la lectura que son típicamente vistos en la escuela o en los libros de texto, elemento que, según los participantes en las sesiones de análisis de sesgo es característico de la escuela tradicional. Este factor potencial de sesgo cultural debe verse con especial cuidado porque si, tal como lo señala Zieky (1993), algunos grupos particulares de evaluados tuvieron menos oportunidad de aprender lo abordado por el ítem o no era enseñado en sus escuelas, el ítem no podría verse como injusto si evaluaba algo que era importante para lo que se pretendía medir. En este último caso, existiría un sesgo pero no en la prueba sino una inequidad real en las condiciones de educación de los grupos, la cual debe ser abordada por los encargados de la política pública.

Los **juicios de valor hacia teorías sociales** se refieren a los juicios subjetivos que deben elaborar los evaluados para responder una pregunta de forma correcta sobre algunos conceptos como igualdad y libertad, y sobre teorías sociales y de mercado. De acuerdo con las fuentes de varianza irrelevantes propuestas por el ETS (2009), este factor podría enmarcarse dentro de las fuentes emocionales, pues al elaborar un juicio de valor sobre aspectos como los mencionados puede hacer sentir a los evaluados que están dando apoyo a ideologías o causas particulares. Esta fuente también podría apoyar los resultados de Costa & Araújo (2011), en donde se encontró que ítems que evaluaban un aspecto denominado “reflexionar y evaluar”, el grupo de referencia (nativos) se desempeñaban mejor, es decir, que aspectos como el evaluar alguna situación o emitir juicios está más asociado con la cultura.

Respecto a cómo mejorar las preguntas, en general se puede apreciar que para los casos en los que la fuente de sesgo es clasificada como cognitiva, existen sugerencias puntuales para mejorarlas. En contraste, cuando se trata de fuentes emocionales, estas pautas son más difíciles de establecer. Frente a este tipo de fuentes, el ETS (2009) recomienda que si estos aspectos se encuentran relacionados con la validez del constructo, se debe hacer lo posible por reducir el impacto emocional que pueda tener el ítem en el evaluado. En este sentido, dicha sugerencia podría aplicarse a aquellos ítems donde existan fuentes emocionales de varianza irrelevante.

Por otra parte, existen elementos en los ítems, que si bien no fueron identificados de forma clara como potenciales fuentes de sesgo, si se deben tener en cuenta. Todos ellos se encuentran relacionados con posibles fuentes emocionales de varianza irrelevante y son términos, palabras o gráficas que pueden incomodar a algún grupo de la población, tales como la palabra “gorda” y el uso de gráficas que pueden tener un significado ofensivo para algún grupo en particular. Otro aspecto que no surgió como relacionado directamente con sesgo, fue el hecho de que a los estudiantes indígenas parece resultarles más difíciles los ítems con gráficas. Esto también se ha encontrado en estudios en los que comunicarse matemáticamente o comparar resultados les resultaba más complejo al grupo focal (Yildirim, 2006).

Llegar a esta primera aproximación a las posibles fuentes de sesgo cultural en los ítems no es una tarea definitiva ni fácil, pues tal como lo señalan Ercikan et al. (2010) y Wu & Ercikan (2006) es un reto, teniendo en cuenta los diferentes factores que afectan la equivalencia en los ítems, que los elementos culturales son difíciles de investigar a nivel metodológico y que su poder inferencial está restringido por otros factores de confusión. Esto hace que las potenciales fuentes de sesgo expuestas a lo largo de todo el trabajo, deban observarse con precaución y siempre a la luz de si son relevantes o no frente al constructo que se pretende medir.

Respecto a las fuentes de sesgo que se usaron para la categorización de las afirmaciones de los participantes por provenir de una u otra forma de la revisión de la literatura, sólo se destacaron dos. Por lo que resultados y propuestas como las de Banks (2006) y Rhodes (1988, citado por Snetzler & Qualls, 2000), entre otras, no fueron apoyadas por los resultados de este trabajo.

Si bien este trabajo no pretendía identificar ítems con sesgo cultural, sino las posibles causas del DIF que pudieran deberse a factores culturales, éstas deben ser tenidas en cuenta al momento de la revisión de ítems, pues, siguiendo a Banks (2006), si éstas no se encuentran relacionadas con el desempeño en el atributo que está siendo medido, sino que por el contrario los ítems son sensibles a ellas y afectan la ejecución de los evaluados en la misma, los ítems estarán sesgados culturalmente.

Uno de los objetivos del presente trabajo era que al culminarlo se contara con algunas pautas para evitar sesgo cultural en los ítems con base en el análisis de las posibles fuentes de DIF, éstas se presentan en el anexo 13 y fueron puntualizadas teniendo en cuenta las sugerencias hechas por los participantes durante las sesiones de análisis de sesgo. Dichas pautas pretenden ser una orientación para el ICFES y para las personas que construyen pruebas en Colombia, de tal manera que los instrumentos que se desarrollen se acerquen más a lo que Solano & Nelson (2001) denominaron validez cultural, es decir, que las pruebas aborden las influencias socioculturales que median el pensamiento del evaluado y la forma en que comprenden los ítems y los responden; estas influencias deberían ser

tenidas en cuenta y tratar de minimizarlas si no están relacionadas con el objeto de medida. De esta manera, este trabajo apoya el cumplimiento de lo que señala los artículos primero y segundo de la Ley 1324 (República de Colombia, 2009) sobre la igualdad para todos los examinados al practicar la evaluación y producir sus resultados, y propende por la validez de las pruebas, pues se procura evitar posibles fuentes sistemáticas de varianza irrelevante que alteran el proceso de medida.

Como todo trabajo de investigación, el presente tiene limitaciones y deja nuevas preguntas para abordar en el futuro. La principal limitación es la no realización de este estudio considerando, por lo menos, como grupo focal a cada una de las familias lingüísticas de los indígenas, si se tiene en cuenta que en Colombia existen 87 pueblos indígenas identificados que hablan 64 lenguas amerindias y diferentes dialectos agrupados en 13 familias lingüísticas (Hernández et al., 2007a). El no considerar este aspecto puede estar enmascarando posibles ítems con DIF debido a fuentes culturales muy específicas de cada grupo. No obstante, los tamaños de evaluados que puede tener cada uno de estos grupos culturales en SABER 11°, haría que la razón de tamaños entre ellos y la población no indígena sea muy extrema y con ello contribuir a desestabilizar los procedimientos aquí usados, en términos de error tipo I y potencia. Adicionalmente, ya se han hecho estudios (por ejemplo, da Costa & Araújo, 2011) en continentes como el Europeo en donde se tiene gran afluencia de inmigrantes de diversos países y todos ellos se han agrupado en “inmigrantes” vs. “nativos”.

En este mismo sentido, se presenta la primera pregunta de investigación emergente, en relación con si las características de la población de la primera aplicación de 2007, ameritan que se haga un estudio específico con algunos de sus subgrupos. Por ejemplo, se encuentran los colegios UNCOLI (Unión de Colegios Internacionales), los cuales suelen presentar a sus estudiantes en la primera aplicación de cada año y se caracterizan por ser bilingües. Esto último puede ser considerado una característica que haga necesario un estudio de DIF y de sesgo cultural con este grupo de colegios. Sin embargo, este tipo de estudio tendría como limitación que no se podría tener certeza sobre si el contexto cultural de los

estudiantes varía del resto de la población dado el carácter bilingüe de los colegios a los que asisten.

Por otra parte, una segunda pregunta de investigación es que, teniendo en cuenta que la presencia de multidimensionalidad puede ser una causa de DIF (Ackerman, 1992), se sugiere la realización de un estudio posterior que tenga como objetivo identificar si ésta puede ser la razón del DIF encontrado en los ítems del presente estudio. De acuerdo con Ackerman (1992), la multidimensionalidad es considerada por muchos investigadores como la principal causa de sesgo en los ítems y está relacionada con una inadecuada especificación del atributo. Es problemática cuando ítems que miden múltiples atributos son puntuados como si midieran un único constructo y discriminan entre los diferentes niveles de habilidad de los atributos que están siendo medidos o diferentes composiciones de los mismos (Ackerman, 1992). En estos casos quienes trabajan con pruebas deben asumir un rol activo e investigar la multidimensionalidad en los datos y entender las consecuencias de reportar una única puntuación (Ackerman, 1992). En este sentido, quienes diseñan las pruebas requieren identificar aquellos ítems que son válidos para ambos grupos, pues los ítems que están influenciados fuertemente por habilidades espurias deben ser eliminados, si el objetivo es tener una prueba con un mayor nivel de consistencia interna (Ackerman, 1992).

Este asunto de la multidimensionalidad cobra más relevancia dados los resultados que a este respecto se encontraron en las pruebas de SABER 11°, pues, dado que en ningún caso fue satisfactoria, debe evaluarse la posibilidad de que ésta sea una posible fuente de DIF y demanda, adicionalmente, que se revise lo que se pretende medir y la forma de abordarlo, como se había mencionado anteriormente. Teniendo en cuenta que el ICFES sigue manejando el modelo de Rasch para los análisis de ítems y calificación de las pruebas que componen este examen, se sugiere que este problema de la multidimensionalidad sea abordado desde la construcción de los ítems, pues la unidimensionalidad es un supuesto básico que se debe garantizar en algún grado para la aplicación de este modelo.

Una tercera pregunta de investigación que queda y que sobrepasa los límites de este estudio es ver si las opiniones de los expertos sobre los ítems concuerdan con lo que se encuentra a nivel de DIF, es decir, si sus opiniones sin conocer las estadísticas son acordes a la presencia y dirección del DIF, lo cual se puede hacer con los mismos datos de esta investigación. En ese mismo sentido y siguiendo a Ercikan et al. (2010), quien señala que la revisión o trabajo con expertos no puede ser considerada como suficiente para decir que un ítem con DIF está sesgado o no, se sugiere que esa futura investigación esté acompañada de procedimientos directamente aplicados con los evaluados tales como entrevistas cognitivas o pensar en voz alta, de tal forma que se tenga mayor claridad sobre las fuentes de sesgo cultural propuestas en este estudio.

REFERENCIAS

- Abad, F., Garrido, J.; Olea, J. & Ponsoda, V. (2006). *Introducción a la psicometría: teoría clásica de los tests y teoría de respuesta al ítem*. Madrid: Universidad Autónoma de Madrid. Documento electrónico.
- Acar, T. (2012). Determination of a Differential Item Functioning Procedure Using the Hierarchical Generalized Linear Model: A Comparison Study with Logistic Regression and Likelihood Ratio Procedure. [Versión Electrónica] *SAGE Open*, 1-8.
- Acevedo, M., Montañéz, J., Huertas, C. & Pérez, M. (2007). *Fundamentación Conceptual. Área de Matemáticas*. Extraído el 18 agosto, 2012 de: http://www2.icfes.gov.co/exámenes/component/docman/cat_view/8-saber-11/20-informacion-general/44-marcos-guias-y-ejemplos-de-preguntas/67-marcos?Itemid=
- Achury, G. P. (2000). *El control fiscal que se ejerce sobre los recursos destinados de los ingresos corrientes de la Nación a los Resguardos Indígenas en Colombia*. Tesis de Pregrado. Universidad Externado de Colombia, Bogotá, Colombia.
- Ackerman, T. A. (1992). A didactic explanation of the item bias, item impact, and item validity from a multidimensional perspective. [Versión Electrónica] *Journal of Educational Measurement*, 29(1), 67-91.
- Al-Fallay, I. S. (1999). Reducing Bias in the assessment of linguistic proficiency of learners of English as a foreign language (EFL). *Dirasat. Human and Social Sciences*, 26(1), 1999, 254-273. Resumen extraído el 18 Marzo, 2008, de la base de datos de EBSCO.
- American Educational Research Association (AERA), American Psychological Association (APA) & el National Council on Measurement in Education (NCME) (1999) *Standards for educational and psychological testing*. Washington, DC: AERA Publications.
- Anastasi, A. (1967). *Tests Psicológicos*. Madrid: Aguilar.
- Arias, E. (2008). *Detección de DIF con Estadísticos Basados en Tablas de Contingencia: El Mantel-Haenszel*. Tesis de Maestría para optar por el título de Magíster de Psicología. Departamento de Psicología, Universidad Nacional de Colombia, Bogotá Colombia.
- Atlas.ti: Visual Qualitative Data Analysis [Computer Program]. Versión 5.2.0. Berlín (Alemania): Scientific Software Development GmbH; 1993 - 2012.
- Baker, S. L. (2001) Short Report. The 60-Item Boston Naming Tests: Cultural bias and possible adaptation for New Zealand. [Versión Electrónica] *Aphasiology*, 15(1), 85-92.
- Banks, K. (2006) A Comprehensive Framework for Evaluating Hypotheses About Cultural Bias in Educational Testing. [Versión Electrónica]. *Applied Measurement in Education*, 19(115-132).
- Barbero, M. I. & Prieto, P. (1997). Evaluación del Rendimiento en Ciencias de los Niños y Niñas de 13 años de las Distintas Comunidades Autónomas: Impacto o Sesgo. [Versión Electrónica]. *Psicothema*, 9(2), 323-332.

- Bolt, D. (2002) A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied measurement in education*, 15,113-141.
- Bradley, J. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Berrío, A. I. (2008). *Efecto de la Razón de Tamaños y Desajustes al Modelo en la Detección de Ítems con Funcionamiento Diferencial mediante Procedimientos Basados en IRT (Diferencia de Dificultad y χ^2 de Lord*. Tesis de Maestría para optar por el título de Magíster en Psicología, Departamento de Psicología, Universidad Nacional de Colombia, Bogotá, Colombia.
- Benbow, C. (1997). The utility of out-of-level testing for gifted seventh and eight graders using the SAT-M: An examination of item bias. En: *Intellectual talent: Psychometric and Social Issues*. Benbow, C. & Lubinski, D. J. (Eds.) p. 333-346. Baltimore: Johns Hopkins University Press, 1996. Resumen extraído el 18 Marzo, 2008, de la base de datos de EBSCO.
- Camilli, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. 4. United States: SAGE Publications.
- Chilisa, B. (2000). Towards Equity in Assessment: crafting gender-fair assessment. [Versión Electrónica]. *Assessment in Education*, 7(1), 61-81.
- Clauser, B. E. & Mazor, K. M. (1998) An NCME module on using statistical procedures to identify differentially functioning test items. [Versión Electrónica]. *Educational Measurement: Issues and Practice*. Spring, 31-44.
- Cervantes, V. H. (2007). Script para el Mantel Haenszel en programa R.2.6, [Script]. Proyecto de identificación de Sesgo cultural en el Examen de Estado ICFES. Grupo de métodos e instrumentos de investigación en salud. Universidad Nacional de Colombia.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.
- Cohen, J. (1992). A power primer. [Versión Electrónica]. *Psychological Bulletin*, 112(1), 155-159.
- Da Costa, P. D. & Araújo, L. (s.f). *Differential Item Functioning (DIF): What Functions Differently for Immigrant Students in PISA 2009 Reading Items?* Draft. Versión Electrónica]. Extraído el 22 septiembre, 2012 de: http://crell.jrc.ec.europa.eu/download/PISA09_DRAFT_IPSC-TechnicalReport.pdf
- Departamento Administrativo Nacional de Estadística. DANE. (2006). *Población Indígena, Rom y Afrocolombiana*. Extraído el 28 agosto, 2008 de: <http://www.dane.gov.co/files/censo2005/etnia/sys/etnias.pdf>.
- Educational Testing Service. (2009). *ETS Guidelines for fairness review of assessments*. Educational Testing Service. Versión Electrónica]. Extraído el 9 octubre, 2012 de: http://www.ets.org/Media/About_ETS/pdf/overview.pdf
- Elder, C., McNamara, T. & Congdon, P. (2003) Rasch techniques for detecting bias in performance assessments: an example comparing the performance of native and non-native speaker on a test of academic English. [Abstract] [Versión Electrónica]. *Journal of Applied Measurement*, 4(2), 181-197.

- Elosua, P. (2006). Funcionamiento diferencial del ítem en la evaluación internacional PISA. Detección y comprensión. [Versión Electrónica] *RELIEVE*, 12(2), 247-259.
- Elosua, P., López, A. & Torres, E. (2000). Desarrollos didácticos y funcionamiento diferencial de los ítems. Problemas inherentes a toda investigación empírica sobre sesgo. [Versión Electrónica]. *Psicothema*, 12(2), 198-202.
- Elosua, P., López, A., Egaña, J., Artamendi, J. A. & Yenes, F. (2000) Funcionamiento diferencial de los ítems en la aplicación de pruebas psicológicas en entornos bilingües. *Metodología de las Ciencias del Comportamiento*, 2(1), 17-33. Resumen extraído el 18 Marzo, 2008, de la base de datos de EBSCO.
- Ercikan, K., Arim, R., Law, D., Domene, J. & Lacroix, S. (2010). Application of Think Aloud Protocols for Examining and Confirming Sources of Differential Item Functioning Identified by Experts Review. *Educational Measurement; Issues and Practice*, 29(2), 24 – 35.
- Fan, X, Willson, V. & Kapes, J. T. (1996). Ethnic group representation in test construction simples and test bias: The standardization fallacy revised. [Versión Electrónica]. *Educational and Psychological Measurement*, 56(3), 365-381.
- Linacre, M. (2006). *A User's Guide to WINSTEPS*. Chicago.
- Losada, J. L. & Arnau, J. (2000). Fiabilidad entre observadores con datos categóricos mediante el Anova. *Psicothema*, 12(2), 335-339.
- Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, 73(1), 1-43. Resumen extraído el 18 Marzo, 2008, de la base de datos de EBSCO.
- Freedle, R. & Kostin, I. (1997). Predicting black and white differential item functioning in verbal analogy performance. [Versión Electrónica]. *Intelligence*, 24(3), 417-444.
- Fidalgo, Á. M. (1996). Funcionamiento Diferencial de los Ítems. En J. Muñiz (Ed.), *Psicometría*, pp. 371-455. Madrid: Editorial Universitas, S.A.
- Fox, J. (2010). polycor: Polychoric and Polyserial Correlations. R package version 0.7-8. Disponible en: <http://CRAN.R-project.org/package=polycor>
- Gallart, D. J. & Moore, J. L. (2008) Assessing Ethnicity. Equity for first-grade male students on a curriculum-embedded performance assessment. [Versión Electrónica]. *Urban Education*, 43(2), 172-188.
- Gierl, M. T. & Nyla, S. (2001). Identifying Sources of Differential Item and Bundle Functioning on Translated Achievement Tests: A Confirmatory Analysis. *Journal of Educational Measurement*, 38(2), 164 – 187.
- Gómez, J. & Hidalgo, M. D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: una revisión metodológica. *Anuario de Psicología*, 74, 3-32.
- Gómez, J., Hidalgo, M. D., Guilera, G., & Moreno, M. (2005). A bibliometric study of differential item functioning. [Versión Electrónica] *Scientometrics*, 64(1), 3-16.
- Gómez, J. & Navas, M. J. (1998) Impacto y Funcionamiento Diferencial de los Ítems Respecto al Género en una Prueba de Aptitud Numérica. [Versión Electrónica] *Psicothema*, 10(3), 686-696.
- Gómez, P. (2000). Diseño en el Objeto Indígena. En Ordoñez, A. (Coord.), *Variación Biológica y Cultural de Colombia*, pp. 133-154. Bogotá: Instituto Colombiano de Cultura Hispánica

- Grupo de Evaluación de la Educación Básica y Media. (1999). *Antecedentes y Marco Legal del Examen de Estado*. [Versión Electrónica]. Colombia: Instituto Colombiano para el Fomento de la Educación Superior.
- Grupo de Evaluación de la Educación Básica y Media. (2005). *Nueva Estructura de los Exámenes de Estado para Ingreso a la Educación Superior y validación del Bachillerato Académico en un solo Examen*. [Versión Electrónica]. Colombia: Instituto Colombiano para el Fomento de la Educación Superior.
- Helms, M. & Van de Vijver, F. (1995). Cognitive assessment in education in a multicultural society. *European Journal of Psychological Assessment*, 11(3). Resumen extraído el 18 Marzo, 2008, de la base de datos de EBSCO.
- Hernández, A., Salamanca, L. M. & Ruiz, F. A. (2007a). Los grupos étnicos en la Colombia de Hoy. [Versión Electrónica]. En Departamento Administrativo Nacional de Estadística (DANE). *Colombia: Una Nación Multicultural. Su Diversidad Étnica*.
- Hernández, A., Salamanca, L. M. & Ruiz, F. A. (2007b). Los grupos étnicos en los censos de población. [Versión Electrónica]. En Departamento Administrativo Nacional de Estadística (DANE). *Colombia: Una Nación Multicultural. Su Diversidad Étnica*.
- Hernández, A., Salamanca, L. M. & Ruiz, F. A. (2007c). La Población Étnica y el Censo General 2005. [Versión Electrónica]. En Departamento Administrativo Nacional de Estadística (DANE). *Colombia: Una Nación Multicultural. Su Diversidad Étnica*.
- Hernández, A. (s.f.) *La Visibilización Estadística de los Grupos Étnicos Colombianos*. Extraído el 12 Octubre, 2012 de http://www.dane.gov.co/files/censo2005/etnia/sys/visibilidad_estadistica_etnicos.pdf
- Herrera, A. N. (2005). *Efecto del tamaño de muestra y la razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems*. Tesis de Doctorado, Facultad de Psicología, Universidad de Barcelona, Barcelona, España.
- Herrera, A. N., Gómez, J., Muñiz, M. D. (2007). Detección del funcionamiento diferencial de los ítems en el marco de la teoría de respuesta al ítem. *Avances en Medición*, 5, 27-46.
- Herrera, A. N., Gómez, J., Hiidalgo, M. D. (2005). Detección de sesgo en los ítems mediante análisis de tablas de contingencia. *Avances en Medición*, 3, 29-52.
- Herrera, A. N., Gómez, J., Quintero, C., Arias, E. M., Berrío, A. I. & Cervantes, V. H. (2007). *Identificación de Ítems con Sesgo Cultural en las Pruebas de los Exámenes de Estado en Colombia. Proyecto de Investigación*. Manuscrito no publicado. Universidad Nacional de Colombia, Bogotá, Colombia.
- Holland, P. W. & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. En H. Wainer & H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, N.J.: Erlbaum.
- Huang, J. & Han, T. (2012). Revisiting Differential Item Functioning: Implications for Fairness Investigation. *International Journal of Education*, 4(2), 74 – 86.
- Hunter, J. E. & Schmidt, F. L. (2000). Racial and Gender Bias in Ability and Achievement Tests. Resolving the Apparent Paradox. [Versión Electrónica]. *Psychology, Public Policy, and Law*, 6(1), 151-158.

- Instituto Colombiano para la Evaluación de la Educación (ICFES). (s.f.). *Estructura del Examen*. [Versión Electrónica] Obtenido de: <http://www.icfes.gov.co/>
- Instituto Colombiano para la Evaluación de la Educación (ICFES). (s.f.b). *¿Qué se evalúa?*. [Versión Electrónica] Extraído el 18 Agosto, 2012 de: <http://www2.icfes.gov.co/exámenes/saber-11o/informacion-general/que-se-evalua>
- Iraurgi, I. (2008). Escala de calidad de vida en usuarios de drogas inyectadas (IDUQoL): valoración psicométrica de la versión española. *Adicciones*, 20(3), 281-294.
- Kamata, A. & Vaughn B. K. (2004). An introduction to Differential Item Functioning Analysis. *Learning Disabilities: A Contemporary Journal*, 2(2), 49 – 69.
- Kellaghan, T., Greaney, V. & Murray, S. (2009). *Using the Results of a National Assessment of Educational Achievement*. World Bank. Extraído el 28 Agosto, 2011 de <https://openknowledge.worldbank.org/handle/10986/2667>
- Kline, T. J. (2004). Gender and Language Differences on the Test of Workplace Essential Skills: Using Overall Mean Scores and Item-Level Differential Item Functioning Analyzes. [Versión Electrónica]. *Educational and Psychological Measurement*, 64(3), 549-559.
- Kramer, B. J. (2009) The art and science of interviewing groups: focus group fundamentals. Madison, Wisconsin: University of Wisconsin-Madison. School of Social Work.
- Linacre, M. (2006). *A User's Guide to WINSTEPS*. Chicago.
- Luykx, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and home language influences on children's responses to science. *Teachers College Record*, 109(4), 897-926.
- Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*. 7, 105 – 118.
- Medellín, J. A. & Fajardo, D. (2006). *Diccionario de Colombia*. Norma: Colombia.
- Millsap, R. E & Everson, H. T. (1993) Methodology Review: Statistical Approaches for assessing measurement bias. [Versión Electrónica] *Applied Psychological Measurement*, 17(4), 297-334.
- Ministerio del Medio Ambiente (2008). *Biodiversidad*. Extraído el 21 Marzo, 2008 de: http://web.minambiente.gov.co/biogeomenu/biodiversidad/culturas/tabla_1.htm.
- Ministerio de Educación Nacional (2001). Etnoeducación. Una política para la diversidad. *Al Tablero*, 3. Extraído el 11 Octubre, 2012 de: <http://www.mineducacion.gov.co/1621/article-87223.html>.
- Ministerio de Educación Nacional (2001). Etnoeducación. Una política para la diversidad. *Al Tablero*, 3. Extraído el 11 Octubre, 2012 de: <http://www.mineducacion.gov.co/1621/article-87223.html>.
- Ministerio de Educación Nacional (s.f.). *Relación Modelo Educativo Grado*. Extraído el 13 Octubre, 2012 de: http://www.mineducacion.gov.co/1621/articulos-255690_archivo_pdf_modelo_grado.pdf.
- Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Ediciones Pirámide.
- Muñoz, J. (2005). *Análisis cualitativo de datos textuales con ATLAS.ti 5*. Barcelona: Universitat Autònoma de Barcelona.

- Nagle, B. & Williams, B. Methodology brief: introduction to focus groups. Center for assessment, planning & accountability. Extraído el 28 Agosto, 2012 de: <http://www.uncfsp.org/projects/userfiles/File/FocusGroupBrief.pdf>.
- Ortiz, J., Ayala, C., Chaparro, J., Sarmiento, J. & Restrepo, G. (2007). *Fundamentación Conceptual. Área de Ciencias Sociales*. Extraído el 18 agosto, 2012 de: http://www2.icfes.gov.co/exámenes/component/docman/cat_view/8-saber-11/20-informacion-general/44-marcos-guias-y-ejemplos-de-preguntas/67-marcos?Itemid=
- Oshima, T., Raju, N., Flowers, P. & Slinde, J. (1998) Differential Bundle Functioning Using the Dfit Framework : Procedures for Identifying Possible Sources of Differential Functioning”. *Applied Measurement in Education*, 11(4), 353-369.
- Oshima, T., Raju, N. & Nanda, A. (2006) A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of educational measurement*, 43, 1-17.
- Padilla, J. L. García, A. S. & Gómez, J. (2007). Evaluación de cuestionarios mediante procedimientos cognitivos. *Avances en medición*, 5(1), 115-126.
- Pomplun, M. & Omar, M. H. (2001). Do Reading passages About War Provide Factorially Invariant Scores for Men and Women? [Versión Electrónica]. *Applied Measurement in Education*, 14(2), 171-189.
- Price, L. R. & Oshima, T. C. (1997). Differential Item Functioning and language translation: A cross-national study with a test developed for certification. *Dissertation Abstracts International Section A. Humanities and Social Sciences*, 58(5-A), 1673. [Versión Electrónica]. Extraído el 5 Abril, 2008 de http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/a3/d3.pdf.
- Quintero, M. (2006). *Aspirantes Especiales*. [Versión Electrónica] Universidad del Tolima.
- Quiñónez, C. (2000). El tejido en las tribus indígenas de Colombia: Unidad y Diversidad. En Ordoñez, A. (Coord.), *Variación Biológica y Cultural de Colombia*, pp. 109-116. Bogotá: Instituto Colombiano de Cultura Hispánica.
- R: a language and environment for statistical computing [Computer Program]. Versión 2.12.1. Vienna (Austria): R Development Core team; 2011. Disponible en: [URL:http://www.r-project.org](http://www.r-project.org).
- R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponible en: URL <http://www.R-project.org/>.
- Package Stats [Computer Program]: R Development Core team; 2011. Disponible en: [URL:http://www.r-project.org](http://www.r-project.org)
- Raju, N. S., van der Linden, W. J. & Fleer, P. F. (1995). IRT-Based Internal Measures of Differential Functioning of Items and Tests. [Versión Electrónica]. *Applied Psychological Measurement*, 19(4), 353-368.
- República de Colombia. (1890). Ley 089 de 1890. [Versión Electrónica]. Extraído el 11 octubre, 2012 de: <http://www.mij.gov.co/econtent/library/documents/DocNewsNo2140DocumentNo1902.PDF>

- República de Colombia. (1980). Decreto 2342 de 1980. [Versión Electrónica]. Extraído el 11 octubre, 2012 de: http://www.mineduccion.gov.co/1621/articulos-103244_archivo_pdf.pdf.
- República de Colombia. (1992). Ley 30 de 1992. [Versión Electrónica]. Extraído el 11 octubre, 2008 de: http://www.mineduccion.gov.co/1621/articulos-86437_Archivo_pdf.pdf
- República de Colombia. (1994). Ley 115 de 1994. [Versión Electrónica]. Obtenido de: http://www.secretariassenado.gov.co/leyes/L0115_94.HTM
- República de Colombia. (1995). Decreto 804 de 1995. [Versión Electrónica]. Extraído el 11 Octubre, 2012 de: http://www.mineduccion.gov.co/1621/articulos-103494_archivo_pdf.pdf
- República de Colombia. (2009). Ley 1324 de 2009. [Versión Electrónica]. Extraído el 10 octubre, 2012 de: http://www.mineduccion.gov.co/1621/articulos-210697_archivo_pdf_ley_1324.pdf
- República de Colombia, Misión Nacional para la Modernización de la Universidad Pública. (1995). *Informe Final*. Colombia: Presencia.
- República de Colombia. (2010). Decreto 869 de 2010. [Versión Electrónica]. Extraído el 17 octubre, 2011 de: http://www.mineduccion.gov.co/1621/articulos-221588_archivo_pdf_decreto_869.pdf
- Revelle, W. (2010) psych: Procedures for Personality and Psychological Research Northwestern University, Evanston, Disponible en: <http://personality-project.org/r/psych.manual.pdf>, 1.0-93
- Revelle, W. (2012). *Package 'psych'*. [Versión Electrónica]. Extraído el 21 Febrero, 2012 de <http://cran.r-project.org/web/packages/psych/psych.pdf>.
- Rocha, M. & Pardo, C. (s.f). *Nuevo Examen de Estado para el Ingreso a la Educación Superior. Cambios para el Siglo XXI. Admisión a la Educación Superior*. Bogotá: ICFES.
- Rodríguez, E. (2007). Colombia, un espacio de vida y encuentro pluricultural. [Versión Electrónica]. En Departamento Administrativo Nacional de Estadística (DANE). *Colombia: Una Nación Multicultural. Su Diversidad Étnica*.
- Rodríguez, S. (2011). *La política educativa (etnoeducación) para pueblos indígenas en Colombia a partir de la Constitución de 1991*. Trabajo de grado como requisito parcial para optar al título de Magister en Antropología. Departamento de Antropología, Universidad Nacional de Colombia.
- Rogers, H. J. & Swaminathan, H. (1993). A Comparison of Logistic Regression and Mantel – Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, 17(2), 105 – 116.
- Rojas, F. (2001). *Proyecto de Ley para la Creación de la Universidad Indígena de Colombia*. [Versión Electrónica] Extraído el 18 Marzo, 2008 de: http://www.etniasdecolombia.org/grupos_etno_universidad2.asp?cid=652&did=1012.

- Ross, J. D. & Okabe, J. (2006). The Subjective and Objective Interface of Bias Detection on Language Tests. [Versión Electrónica] *International Journal of Testing*, 6(3), 229-253.
- Rubio, G. (2000). Arquitectura Indígena en Colombia. En Ordoñez, A. (Coord.), *Variación Biológica y Cultural de Colombia*, pp. 65-68. Bogotá: Instituto Colombiano de Cultura Hispánica.
- Ryan, K. E. & Chiu, S. (2001). An Examination of Item Context Effects, DIF, and Gender DIF. [Versión Electrónica]. *Applied Measurement in Education*, 14(1), 73-90.
- Santana, A. C. (2009) *Efecto de la razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems a través del procedimiento de regresión logística*. Tesis de maestría para optar por el título de Magíster en Psicología. Departamento de Psicología, Universidad Nacional de Colombia.
- Seymour, A. (2004). Focus groups. An important tool for strategic planning. Washington, D.C.: Justice Solutions.
- Sireci, S. G. & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, 20(2), 148 – 166.
- Smith, R.M, Schumacker, R. E. & Bush M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2(1), 66-78.
- Snetzler, S. & Qualls, A. L. (2000). Examination of Differential Item Functioning in a Standardized Achievement Battery with Limited English Proficient Students. [Versión Electrónica]. *Educational and Psychological Measurement*, 60(4), 564-577.
- Snijkers, G. J. (2002). Cognitive laboratory experiences: on pre-testing computerised questionnaires and data quality. Tesis doctoral, Universiteit Utrecht, Utrecht, Netherlands.
- Solano, G. & Nelson, S. (2001). On the Cultural Validity of Science Assessments. *Journal of Research in Science Teaching*, 38(5), 553 – 573.
- Stewart, D., Shamdasani, P., & Rook, D. (2007). *Focus Groups. Theory and Practice*. Thousand Oaks, California: Sage Publications.
- Stricker, L. J. & Emmerich, W. (1999) Possible Determinants of Differential Item Functioning: Familiarity, Interest and Emotional Reaction. [Versión Electrónica]. *Journal of Educational Measurement*, 36(4), 347-366.
- Tellegen, P. J. & Laros, J. A. (2004) Cultural Bias in the SON-R Test: Comparative Study of Brazilian and Dutch Children. [Versión Electrónica] *Psicología: Teoría e Pesquisa*, 20(2), 103-111.
- Toro, J., Reyes, B. & Martínez, R. (2007). Prueba de Biología. Núcleo Común y Profundización. En Toro, J., Reyes, C., Martínez, R., Castelblanco, Y., Cárdenas, F., Granés, J. & Hernández, C., *Fundamentación Conceptual. Área de Ciencias Naturales*. pp. 48-65. Bogotá: ICFES. Extraído el 18 agosto, 2012 de: http://www2.icfes.gov.co/examenes/component/docman/cat_view/8-saber-11/20-informacion-general/44-marcos-guias-y-ejemplos-de-preguntas/67-marcos?Itemid=
- Toro, J., Reyes, C., Martínez, R., Castelblanco, Y., Cárdenas, F., Granés, J. & Hernández, C. (2007). Introducción. En Toro, J., Reyes, C., Martínez, R., Castelblanco, Y.,

- Cárdenas, F., Granés, J. & Hernández, C., *Fundamentación Conceptual. Área de Ciencias Naturales*. pp. 48-65. Bogotá: ICFES. Extraído el 18 agosto, 2012 de: http://www2.icfes.gov.co/examenes/component/docman/cat_view/8-saber-11/20-informacion-general/44-marcos-guias-y-ejemplos-de-preguntas/67-marcos?Itemid=
- Torrado, M. C. (1998). *De la Evaluación de Aptitudes a la Evaluación de Competencias*. Bogotá: División de Procesos Editoriales del ICFES.
- Uiterwijk, H. & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch test. *Language Testing*, 22(2), 221-234. Resumen extraído el 18 Marzo, 2008, de la base de datos de EBSCO.
- Veale, J. R. & Foreman, D. I. (1983). Assessing cultural bias using foil response data: cultural variation. [Versión Electrónica]. *Journal of Educational Measurement*, 20(3), 249-258.
- Wang, N. & Lane, S. (1996). Detection of Gender-Related Differential Item Functioning in a Mathematics Performance Assessment. [Versión Electrónica]. *Applied Measurement in Education*, 9(2), 175-199.
- Walstad, W. B. & Robson, D. (1997) Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics. [Versión Electrónica] *Journal of Economic Education*, Spring, 155-171.
- WINSTEPS MINISTEP: Rasch-Model Computer Programs. [Computer Program]. Versión 3.63.0. Chicago (Estados Unidos): John M. Linacre; 2006.
- Wu, A. D. & Ercikan, K. (2006). Using Multiple-Variable Matching to Identify Cultural Sources of Differential Item Functioning. [Versión Electrónica]. *International Journal of testing*, 6(3), 287-300.
- Yildirim, H. H. (2006) *The Differential Item Functioning (DIF) Analysis of Mathematics Items in the International Assessment Program*. Tesis de doctorado en cumplimiento parcial de los requerimientos para el grado de Doctor en Filosofía. The Graduate School of Natural and Applied Sciences of Middle East Technical University.
- Zieky, M. (1992). Evaluating Items for Fairness. *CLEAR Exam Review*, 3(2), 26 – 31.

Anexo 1. Distribución de la población indígena según etnias por zonas territoriales del DANE y departamentos

Territoriales DANE y departamentos	Pueblos Indígenas o etnias
Norte	
Atlántico	Mokana
Cesar	Arhuaco, Kogui, Wiwa, Yuko, kankuamo
La Guajira	Arhuaco, Kogui, Wayuu, Wiwa
Magdalena	Arhuaco, Chimila, Kogui, Wiwa
Sucre	Senú
Noroccidental	
Antioquia	Embera, Embera Chamí, Embera Katio, Senú, Tule
Córdoba	Embera Katio, Senú
Chocó	Embera, Embera Chamí, Embera Katio, Tule, Waunan
Nororiental	
Arauca	Betoye, Chiricoa, Hitnu, Kuiba, Piapoco, Sikuani, U´wa
Norte de Santander	Barí, U´wa
Santander	(U´wa), Guanés
Central	
Boyacá	U´wa, Muisca
Caquetá	Andoke, Coreguaje, Coyaima, Embera, Embera katio, Inga, Makaguaje, Nasa, Uitoto
Casanare	Amorúa, Kuiba, Masiguare, Sáliba, Sikuani, Tsiripu, Yaruros, U´wa
Cundinamarca	Muisca
Huila	Coyaima, Dujos, Nasa, Yanacona
Meta	Achagua, Guayabero, Nasa, Piapoco, Sikuani
Amazonas	Andoke, barasana, Bora, Cocama, Inga, Karijona, Kawiyarí, Kubeo, Letuama, Makuna, Matapí, Miraña, Nonuya, Ocaina, Tanimuka, Tariano, Tikuna, Uitoto, Yagua, Yauna, Yukuna, Yuri
Guainía	Kurripako, Piapoco, Puinave, Sikuani, Yeral
Guaviare	Desano, Guayabero, Karijona, Kubeo, Kurripako, Nukak, Piaroa, Piratapuyo, Puinave, Sikuani, Tucano, Wanano
Vaupés	Bara, Barasana, Carapana, Desano, Kawiyarí, Kubeo, Kurripako, Makuna, Nukak, Piratapuyo, Pisamira, Siriano, Taiwano, Tariano, Tatuyo, Tucano, Tuyuka, Wanano, Yurutí
Vichada	Kurripako, Piapoco, Piaroa, Puinave, Sáliba, Sikuane
Centroccidental	
Caldas	Cañamomo, Embera, Embera Chamí, Embera Katio
Risaralda	Embera, Embera Chamí
Tolima	Coyaima, Nasa
Suroccidental	
Cauca	Coconuco, Embera, Eperara Siapidara, Guambiano, Guanaca, Inga, Nasa, Totoró, Yanacona
Nariño	Awa, Embera, Eperara Siapidara, Inga, Kofán, Pasto
Putumayo	Awa, Coreguaje, Embera, Embera Katio, Inga, Kamëntsa, Kofán, Nasa, Siona, Uitoto
Valle del Cauca	Embera, Embera Chamí, Nasa, Waunan

Nota: Tomado de Hernández et al. (2007a).

Anexo 2. Resguardos Indígenas por Zonas Territoriales del DANE y Departamento

Territoriales DANE y departamentos	No. de Departamentos con resguardos	No. de municipios con resguardos	No. de resguardos
Norte	4	24	34
Cesar		5	10
La Guajira		11	20
Magdalena		5	3
Sucre		3	1
Noroccidental	3	47	160
Antioquia		19	42
Córdoba		3	3
Chocó		25	115
Nororiental	3	14	30
Arauca		6	26
Norte de Santander		6	3
Santander		2	1
Central	10	60	200
Boyacá		2	1
Caquetá		10	45
Casanare		4	10
Huila		10	14
Meta		6	20
Amazonas		10	26
Guainía		6	25
Guaviare		4	24
Vaupés		4	3
Vichada		4	32
Centroccidental	3	14	77
Caldas		5	6
Risaralda		3	5
Tolima		6	66
Suroccidental	4	69	221
Cauca		26	83
Nariño		17	60
Putumayo		13	55
Valle del Cauca		13	23
Totales	27	228	710

Nota: Tomado de Hernández et al. (2007a).

Anexo 3. Organización Lingüística de las Comunidades Colombianas

MACROFAMILIA	FAMILIA	GRUPO LINGÜISTICO	LENGUA
ANDINO – ECUATORIAL	ANDINO	Quichua	Inga, Yanacona
		Guahibo	Amorua, Guayabero, Cuiba, Sikwani, Hitne?
	ECUATORIAL	Arawac Amazonico	Achagua, Kurripaco, Piapoko, Saliba, Piaroa, Yucuna
		Arawac Caribeño	Wayuunaiki
		Tupi	Kokama
	KOFAN	Kofan	Kofan
	MACRO TUKANO	Maku	Nukak
		Puinave	Pinave
		Tikuna	Tikuna
		Tukano occidental	Koreguaje, Siona
Tukano oriental		Tanimuka, Kubeo, Tukano	
GE - PANO – CARIBE	MACRO CARIBE	Caribe	Chimila, Pijao, Yukpa-Yuko, Zenu
		Peba Yagua	Yagua
		Witoto	Andoke-Nonuya, Muinane, Witoto
MACRO CHIBCHA	BARBACOA	Guambiano	Guambiano
		Awa Kwaiker	Awa Kwaiker
		Pasto Quillasinga	No Hablantes
	CHIBCHA PROPIO	Arhuaco	Damana, Bintukua, Kaggaba
		Bari-Dobokubi	Bari
		Kuna-Kueba	Kuna
		Muisca	No Hablantes
		Tunebo	Unkasia, Tegria-Cobaria
	CHOKO	Embera	Embera, Embera Chami, Embera Catio, Embera Saija
	KAMNTXA	Waunana	Noanama
Kamntxa		Kamntxa	
PAEZ COCONUCO	Paez Coconuco	Kokonuko, Paez, Totoro	

Nota: Tomado de Ministerio del Medio Ambiente (2008).

Anexo 4. Distribución porcentual por etnias de los estudiantes que presentaron SABER 11° en el segundo semestre de 2006 y que fueron considerados como indígenas para el análisis

Etnia	Porcentaje	Etnia	Porcentaje
Achagua	,06%	Nonuya	,06%
Amorua	,06%	Ocaina	,03%
Andoque	,03%	Paéz	15,34%
Arhuaco	1,11%	Pastos	1,96%
Awa	,12%	Piapoco	,03%
Barazana	,06%	Piratapuyo	,09%
Betoye	1,18%	Puinave	,15%
Bora	,03%	Sáliba	,33%
Cancuamo	,27%	Sikuani	3,28%
Cocama	,42%	Siona	,03%
Coconuco	,33%	Siriano	,03%
Cofán	,09%	TanimuKa	,06%
Coyaima-Natagaima	9,07%	Tatuyo	,06%
Cubeo	1,24%	Tikunas	1,36%
Curripaco	,12%	Tucano	1,08%
Desano	,24%	Tule	,03%
Embera Catio	5,33%	Tuyuca	,06%
Embera Chami	,99%	U´wa	,12%
Guambiano	2,59%	Wanano	,06%
Inga	3,13%	Wayuu	13,38%
Kamsa	,66%	Witoto	2,32%
Kogui	,03%	Wiwua	,12%
Koreguaje	,24%	Wounaan	,09%
Matapi	,03%	Yagua	,06%
Miraña	,15%	Yanacona	1,11%
Muinane	,12%	Yucuna	,06%
Muisca	1,87%	Yuko	,42%
		Zenú	28,68%

Anexo 5. Distribución porcentual por etnias de los estudiantes que presentaron SABER 11° en el primer semestre de 2007 y que fueron considerados como indígenas para el análisis

Etnia	Porcentaje
Paez	5,53%
Arhuaco	,26%
Wayúu	,13%
Pastos	92,75%
Inga	1,32%

Anexo 6. Protocolo para el análisis de sesgo cultural con expertos y constructores de ítems

Este protocolo tiene como objetivo brindar una guía para el desarrollo de sesiones de análisis de sesgo cultural en ítems de selección múltiple de pruebas objetivas, con el fin de identificar la fuente del posible comportamiento psicométrico diferencial de los ítems. Básicamente, se trabaja en la búsqueda de características de los ítems que pudieran hacer que un grupo u otro (por ejemplo, indígenas versus no indígenas) tuvieran una mayor probabilidad de contestarlos correctamente.

Antes de presentar el protocolo como tal, se debe señalar que es recomendable realizar un análisis estadístico previo a la realización de uno de tipo cualitativo. Específicamente, este análisis se refiere al funcionamiento diferencial del ítem (DIF, por sus siglas en inglés), el cual permite establecer si los ítems tienen propiedades psicométricas diferentes para los grupos que se están comparando. Esta recomendación se hace porque que las técnicas cualitativas en general requieren un trabajo extenso de campo y la participación de diferentes personas que pueden contribuir al análisis, las cuales pueden no tener siempre el tiempo necesario para hacer parte de la investigación. Adicionalmente, el análisis de DIF permite tener un criterio previo para conocer si los ítems efectivamente tienen un comportamiento psicométrico diferente para los grupos que se están comparando y así reducir la cantidad de preguntas a revisar, es decir, no es imprescindible revisar todas las que componen una prueba sino aquellas que hayan presentado DIF.

El protocolo se divide en dos partes, aunque se recomienda que la primera sólo sea realizada cuando se tiene un número elevado de ítems detectados con DIF, por ejemplo más de ocho, si se tiene en cuenta que se sugiere que la duración de los grupos focales (segunda parte) sea entre una y dos horas (Padilla, García y Gómez, 2007) y se estima que la discusión sobre cada ítem tome 15 minutos.

La primera parte consiste en el diligenciamiento de un formato, la cual podría enmarcarse dentro de la primera fase de lo que en la literatura se conoce como juicio de

expertos (Padilla et al., 2007) y que tomará aproximadamente una hora. En la segunda parte, se realizan grupos focales, los cuales durarán aproximadamente dos horas.

Se sugiere que antes de iniciar las sesiones de análisis de sesgo, se realice una capacitación a los participantes, en la que se informe el objetivo del análisis, las fases que lo componen, el objeto de medida de las pruebas, su papel en el estudio y conceptos básicos necesarios para el desarrollo adecuado de la actividad. A continuación se describen los participantes y sus características, las instrucciones y los instrumentos de recolección de información.

Participantes

Los participantes en el análisis de sesgo deben ser, en primera instancia e idealmente aquellos quienes construyeron los ítems o que conozcan las especificaciones de la prueba y el objetivo e intención de las preguntas. Adicionalmente, es importante contar con expertos que conozcan los grupos que se están comparando, por ejemplo indígenas y no indígenas. También es conveniente contar con la participación de un moderador que conozca de procedimientos psicométricos de evaluación de sesgo y cómo puede manifestarse en las preguntas que se están revisando.

Así las cosas, el número mínimo recomendado de participantes en el análisis de sesgo es de cinco personas. En este sentido, Padilla et al. (2007) recomiendan que sean entre cinco y diez personas las que participen en los grupos focales. Se sugiere que los cinco participantes estén divididos entre constructores de preguntas y expertos que conozcan los grupos que se están comparando.

Procedimiento e instrucciones

En este apartado se describen en *itálicas* las instrucciones que se espera sean dichas a los participantes en el análisis de sesgo y el procedimiento general de las sesiones.

Sesión de información

La sesión de capacitación debe iniciar informándoles a los participantes el contexto del análisis, si por ejemplo es una investigación, y el objetivo del mismo. Posteriormente, deben ser informados sobre conceptos como validez, sesgo y fuentes culturales de favorabilidad, o potenciales fuentes, reportadas en la literatura, las cuales pueden ser ejemplificados a través de ítems.

También debe dárseles a conocer aspectos directamente relacionados con los ítems que se van a analizar y el instrumento al que pertenecen, entre ellos se encuentran la población objetivo, el objeto de medida y el propósito de la prueba. Finalmente, debe informárseles el objetivo de su participación en las sesiones de análisis y sus funciones en las mismas.

Sesión de Análisis de Sesgo

Se recomienda que el facilitador de la sesión sea una persona con conocimientos en psicometría, DIF y sesgo. Igualmente, es ideal contar con una persona que acompañe al facilitador en la toma de apuntes durante la sesión.

Las instrucciones aquí presentadas se encuentran basadas en Kramer (2009), Nagle y Williams (2012) y Snijkers (2002). Debe tenerse en cuenta que para los grupos focales existen algunas recomendaciones de tipo logístico y de trato con los participantes que permitirán una conversación fluida y una conclusión exitosa de la actividad, las cuales no son expuestas aquí; sin embargo pueden ser consultadas en las referencias mencionadas anteriormente.

La sesión se debe iniciar dando la bienvenida a los participantes e informándoles por qué están allí y por qué fueron seleccionados. Es probable que esta información ya les haya sido suministrada en la sesión previa de información, por lo que si es así no será necesario volverla a mencionar.

Se debe procurar que las sesiones inicien y terminen a tiempo. En caso de que algún participante no haya llegado a la hora acordada, se recomienda preguntarle a los demás si están de acuerdo con esperar un momento a quien hace falta (Seymour, 2004).

"Buenos días. Bienvenidos a la sesión de análisis de los ítems de la prueba XXXX. Gracias por disponer parte de su tiempo para participar en esta investigación. Mi nombre es XXXX y seré el facilitador de esta sesión. Represento a la XX (entidad a la que se representa) y he estado trabajando en este estudio XX años. Me acompaña XXXX quien estará tomando apuntes durante toda la sesión y él viene de la XX (entidad de la que proviene el asistente).

Hoy haremos una revisión de algunos ítems que componen la prueba y tendremos una conversación con el fin de identificar si existen posibles fuentes culturales de favorabilidad. Ustedes fueron seleccionados para participar en esta sesión porque algunos han sido constructores de ítems, expertos revisores de las pruebas o son expertos en las poblaciones que se quieren comparar en esta investigación y porque queremos que nos compartan sus percepciones e ideas frente a los ítems que revisaremos. Nosotros confiamos en que su experiencia y conocimiento nos ayudará a comprender mejor la potencial existencia de favorabilidad cultural en los ítems de las pruebas que en general son aplicadas en Colombia".

El facilitador debe hacer que los participantes se presenten y solicitarles que firmen el acuerdo de confidencialidad y consentimiento informado (Anexo 7). Se recomienda que los nombres de cada persona se encuentren en escarapelas con el fin de que los participantes se llamen por sus nombres directamente. El facilitador debe informarles que la sesión se compone de dos partes y que toda la información es confidencial, es decir, que no será factible identificar a las personas con las afirmaciones que hagan durante la actividad. Así mismo, es recomendable que el facilitador introduzca al grupo en el tema a través de preguntas generales.

"Simplemente para comenzar, cada uno diga su nombre y alguna otra información que quiera añadir, por ejemplo, si han trabajado en proyectos de investigación similares o con XXXX (la entidad a cargo del proyecto). Por favor empiece usted...."

"Ahora les entregaré un documento que comprende el consentimiento informado para participar en esta investigación y el acuerdo de confidencialidad sobre la información que se tratará en la sesión para que sea firmado por ustedes. Tengan en cuenta que toda la información que brinden es confidencial y por tanto no se asociará ninguna de las afirmaciones que hagan con alguno de ustedes en particular, por lo que los datos siempre aparecerán consignados de forma anónima."

"En este momento daremos inicio a la sesión, la cual durará aproximadamente tres horas. Los celulares y demás dispositivos electrónicos deben estar apagados. La sesión se compone de dos partes, en la primera se realizará una revisión individual de los ítems y en la segunda analizaremos entre todos algunos de ellos a través de un grupo focal."

El facilitador debe entregar a cada participante el cuestionario de la prueba correspondiente y los factores culturales que han sido reportados en la literatura como características que pueden ser fuente de sesgo, e iniciar con la presentación de los ítems a través de Vídeo Beam. También les pueden ser entregados directamente los cuadernillos, sin embargo en ocasiones cuando se trabaja en lugares con seguridad como los bancos de ítems de las diferentes entidades que realizan pruebas, probablemente no se disponga de varios cuadernillos, por lo que tal vez sea más fácil presentarlos a todos los participantes de forma digital.

El cuestionario que deben responder los participantes se encuentra en el Anexo 8 y la hoja con las fuentes derivadas de la literatura en el Anexo 9. Los ítems que se van a revisar son aquellos que fueron detectados con DIF favoreciendo a uno de los dos grupos de comparación; sin embargo, para controlar que los participantes no tiendan a clasificar todos los ítems como sesgados sin que realmente los consideren así, se incluirán ítems que no mostraron DIF.

“Durante la primera parte se les entregará un cuestionario que está diseñado para la revisión de XX preguntas el cual deben responder. Los ítems serán ilustrados en la pantalla uno a uno para que todos hagan la evaluación del mismo ítem al mismo tiempo. Antes de iniciar a responder cada aspecto del cuestionario deben responder la pregunta ustedes mismos.

El cuestionario inicia con una pregunta sobre si consideran que el ítem favorece a un grupo particular, en este caso, indígenas y no indígenas, si la respuesta es afirmativa deben señalar a cuál grupo. Después de señalar a cuál grupo favorece, deberán indicar qué aspecto o aspectos del ítem hacen que un grupo particular se vea favorecido y tratarlos de definir. Para ello pueden acudir a las descripciones de factores culturales que se han reportado en la literatura y que se encuentran en la hoja adjunta al cuestionario, aunque no tienen que ser los mismos, pueden ser aspectos totalmente diferentes.

Si tienen alguna duda por favor planteen sus preguntas ahora. Traten de no tomarse más de cinco minutos en la revisión de cada pregunta.”.

"A continuación empezaré por mostrarles el primer ítem.... este es el segundo ítem... etc."

Quando todos los participantes hayan evaluado el primer ítem se debe pasar al siguiente y así sucesivamente. Se espera que la revisión de cada pregunta no tome más de cinco minutos. Al terminar de evaluar todos los ítems, los participantes pueden tomar un descanso, mientras que el facilitador debe recoger los cuestionarios y consignar en el formato los resultados preliminares de esta fase (Anexo 10).

En este momento, el facilitador se debe fijar principalmente en que haya total acuerdo sobre si el ítem presenta o no favorabilidad hacia algún grupo y el grupo al que favorece, lo que dará un máximo total de posibles dos acuerdos o desacuerdos totales, los cuales deberá registrar en el formato.

El facilitador debe tener siempre en cuenta que aunque el ítem presente DIF, no necesariamente presenta sesgo, por lo que si llega a existir acuerdo en que el ítem no favorece a alguno de los grupos, éste se considerará como no sesgado y no hará parte del análisis en los grupos focales.

Después del descanso, el facilitador debe iniciar con los grupos focales, los cuales serán grabados, y explicar cómo será usada la información que brinden los participantes. Los ítems que primero se deben revisar son los que presentaron más desacuerdos, si en el grupo focal queda tiempo se pueden revisar algunos en los que hubo total acuerdo en que favorecían a algún grupo particular.

El primer paso en la revisión del ítem es identificar su objetivo, su clasificación taxonómica (si responde a una clasificación particular a partir de una estructura de prueba) y una posible justificación de las opciones (claves y opciones no válidas), con el fin de abordar aspectos como lo sugeridos por Zieky (1992): relación entre el atributo medido y las especificaciones de la prueba, la importancia del ítem frente al mismo, si la causa del DIF no está relacionada con los objetivos del ítem y si los examinados podrían sentirse excluidos con la pregunta. Debe señalarse que aspectos como el objetivo de la pregunta, su clasificación taxonómica o la justificación de las opciones suele venir en las fichas técnicas de los ítems.

Posteriormente, para cada uno de los ítems a revisar se mostrará los porcentajes de elección de las opciones de respuesta para cada grupo, cuando se disponga de ellos. A partir de estos datos se señalará a qué grupo favorece el ítem.

Luego se indagará por aspectos relacionados con las características por las cuales los participantes consideraron que algún o algunos ítems favorecen a un grupo particular y por las que señalaron ciertas fuentes como posibles causas de la favorabilidad. Debe tenerse en cuenta que en la revisión de las preguntas en la segunda parte de la sesión, no se requerirá que se llegue a un acuerdo general sobre si el ítem presenta sesgo o no, el grupo al que favorece o la fuente del sesgo, porque el objetivo de esta fase es conocer diferentes puntos de vista.

"Ahora tomaremos un descanso de 15 minutos y volveremos para iniciar con el grupo focal."

"En este momento daremos inicio al grupo focal y éste será grabado. La información que quede registrada será analizada XXXX (por ejemplo a través de programas de análisis de texto) ¿Alguien tiene alguna objeción con la grabación de esta parte de la sesión?"

Aunque se espera que los participantes no tengan objeción con la grabación de la sesión, puede suceder que se presente; por lo cual se recomienda que se les informe sobre esto en la capacitación inicial. En caso de que alguien presente objeción, a pesar de haber sido informado previamente, no se debe grabar la sesión. Después de cada pregunta se debe resumir las opiniones, con lo cual se espera ayudar a mantener la conversación.

"En esta parte de la sesión sólo revisaremos aquellos ítems en los que no hubo total acuerdo respecto a si presentaban favorabilidad hacia algún grupo o el grupo al que favorecían, tomando los datos que ustedes brindaron en la primera parte. Aunque si disponemos de tiempo, también revisaremos aquellos en los que hubo total acuerdo. Deben tener en cuenta que no hay respuestas correctas o incorrectas, simplemente se busca conocer diferentes puntos de vista. Siéntanse libres de compartir su opinión, así sea diferente a la expresada por otros participantes."

Como debo tomar notas y posteriormente escuchar la grabación, agradezco que hable una sola persona a la vez, que levanten la mano cuando deseen intervenir y que no interrumpen a otras personas cuando estén hablando. También es importante que respondan a cada pregunta en su orden, es decir, si bien después de analizado el primer ítem ya conocerán las preguntas que voy a realizar para todos los ítems, por favor respondan a cada una de ellas cuando yo la haga. Si llego a interrumpir a alguno durante su intervención, no es por descortesía, simplemente estoy tratando de asegurar que cada uno de ustedes pueda participar, que el grupo esté centrado en la tarea y que abordemos todos los temas. Las conclusiones de este grupo focal serán resumidas junto con otros grupos que estamos llevando a cabo."

Empezaremos por el ítem XX, mencionando algunos aspectos que nos pueden ayudar en la evaluación del ítem. Esta pregunta pertenece a la categoría (competencia, por ejemplo) y a la categoría (componente, por ejemplo) (o leer el de la ficha técnica, si se tiene).

En caso de que hubiese desacuerdo sobre si el ítem favorece a un grupo particular o no, debe realizarse la siguiente pregunta, de lo contrario se debe pasar a la próxima:

"Algunos de ustedes señalaron que el ítem favorecía a un grupo, otros dijeron que no. ¿Qué aspectos del ítem hacen que favorezca o no a un grupo particular? Por ejemplo, información que es más fácil de recordar para un grupo que para otro; sintaxis compleja; supuestos implícitos; términos técnicos, vagos, ambiguos o no definidos; temas sensibles para uno de los grupos; diseño poco claro; demasiados o pocos detalles; opciones de respuestas que se solapan entre ellas; vocabulario difícil; preguntas muy largas; requiere mucha información para responderlo; implica emitir un juicio; alta dificultad (cálculos o estimaciones complejas); deseabilidad social; la pregunta no es suficientemente realista (datos hipotéticos, ficticios, inexistentes o inaccesibles); adivinación; etc. ...

¿Podrían darme un ejemplo?

"Ahora veamos el porcentaje de elección para cada una de las opciones de respuesta (cuando se disponga de estos datos). El ítem favorece al grupo XXXX. ¿Qué indicios nos brindan los porcentajes por opción de respuesta sobre la razón por la cual este ítem favorece al grupo XXXX?"

"¿Cómo podríamos definir esos aspectos que hacen que un ítem sea más favorable para un grupo que para otro?"

De acuerdo con lo que ustedes han señalado, estos son los aspectos que hacen que el ítem favorezca a un grupo particular... ¿Hemos dejado algo importante sin mencionar?"

Cuando se presente desacuerdo sobre el grupo que se ve favorecido por el ítem se debe realizar la siguiente pregunta, en caso contrario se debe pasar a la próxima:

En este caso el facilitador debe asegurarse de que la definición brindada por los participantes sobre la nueva fuente de sesgo sea similar en su estructura a las reportadas en la literatura como posibles fuentes de sesgo cultural. Finalmente, es necesario indagar sobre cómo los participantes piensan que se puede mejorar la pregunta.

"¿Cómo podemos mejorar esta pregunta para evitar que presente favorabilidad hacia algún grupo?... ¿Podrían darme un ejemplo?".

Se espera que el análisis de cada ítem no tome más de 15 minutos. Ahora bien, si ya ha pasado más de una hora y 45 minutos y aún no se han terminado de revisar todos los ítems, se debe finalizar el grupo focal resumiendo las principales conclusiones con los participantes y agradeciéndoles su participación en la actividad. También es recomendable que en el cierre de la actividad se explique nuevamente el uso que se hará de la información que los participantes han suministrado y se reafirmará la confidencialidad de los datos recolectados. Esta forma de finalizar los grupos focales, también se aplica cuando se han terminado de revisar todos los ítems.

"Ahora que hemos concluido el grupo focal, les pido que por favor hagamos un resumen de las principales conclusiones de esta discusión. Empezaremos por..."

Teniendo en cuenta el objetivo del estudio, ¿Hay alguna pregunta adicional?...

Recuerden que la información que han suministrado no será posible asociarla a alguno de ustedes en particular y que será consignada de forma anónima. Así mismo los datos serán analizados mediante programas de análisis textual para establecer frecuencias que permitan la identificación de posibles fuentes de favorabilidad en los ítems. Muchas gracias por su participación."

Anexo 7. Acuerdo de confidencialidad y consentimiento informado

En el marco de la investigación “Sesgo cultural en las pruebas del examen de Estado SABER 11° en Colombia”, yo _____, identificado(a) con documento de identidad No. _____ participaré en la revisión y análisis de preguntas de dicho examen. En el desarrollo de la actividad tendré acceso a información y datos confidenciales, por lo que me comprometo a no revelarlos bajo circunstancia alguna.

Adicionalmente, manifiesto que se me han dado a conocer y he leído y comprendido los objetivos, alcances y limitaciones de la investigación, así como mi papel y funciones en ella y específicamente en la sesión de análisis de sesgo. También declaro que se me ha indicado que la información que brinde en el desempeño de mis funciones en la investigación sólo será utilizada en ella y siempre aparecerá consignada de forma anónima. Por lo anterior, acepto voluntariamente participar en la mencionada investigación y que mis intervenciones sean grabadas a través de radiograbadora, comprendiendo que me puedo retirar en cualquier etapa del proyecto sin ser penalizado de forma alguna.

Firma del participante en la investigación.
Documento de Identidad No:

Anexo 8. Cuestionario sobre los ítems de las pruebas del examen de Estado SABER 11°

Prueba: _____

Formación de pregrado del participante: _____

Formación de postgrado del participante: _____

A continuación encontrará una serie de preguntas sobre ítems que serán expuestos en una pantalla, después de responderlos debe completar cada una de las siguientes casillas. Tenga en cuenta que en la primera de ellas debe marcar el número de la pregunta que aparecerá en pantalla.

<p>Ítem: _____</p>	<p>¿El ítem favorece a un grupo en particular?</p> <p>Sí _____ No _____</p> <p>Si su respuesta a la anterior pregunta fue afirmativa, por favor indique a qué grupo favorece:</p> <p>Indígenas _____ No indígenas _____</p>	<p>En caso de que considere que el ítem presenta favorabilidad hacia un grupo, describa y defina cuáles aspectos en el ítem hacen que se presente esa favorabilidad. Puede acudir a la hoja que contiene las definiciones de factores culturales encontradas en la literatura, aunque no necesariamente deben ser los mismos aspectos.</p>	<p>En caso de que considere que el ítem presenta favorabilidad hacia un grupo, describa cómo podría evitarse:</p>
-------------------------------	---	---	--

Anexo 9. Definiciones de fuentes de sesgo o factores culturales reportados en la literatura

Familisimo

“Enfatiza el valor de poner las necesidades familiares antes de los intereses personales. Los valores familiares acentúan la codependencia o interdependencia, la cooperación, el compartir la solución a los problemas y recursos materiales” (pág. 118, Banks, 2006).

Presentisimo

“Enfatiza la importancia del aquí y ahora. Los valores se centran en las necesidades actuales y en alcanzar metas a corto plazo antes que embarcarse en temas más amplios” (pág. 118, Banks, 2006).

Personalisimo

“Es la noción de la buena comunicación interpersonal entre individuos. La comunicación implica compartir sentimientos, aportarle al otro y validar el valor propio del otro” (pág. 118, Banks, 2006).

Espiritisimo

“Se destacan las interconexiones entre los seres humanos y lo natural y súper natural. Más que controlar ciertos aspectos del ambiente, se cree que los individuos están pensados para estar en armonía con ellos” (pág. 118, Banks, 2006).

Comunalisimo

“Hace referencia a la interconexión de las personas, por ejemplo se cree que los afroamericanos están más orientados hacia las personas que hacia los objetos. Esta característica acentúa la importancia del deber de cada persona hacia un grupo social, la

cooperación, la interdependencia, la cohesión e identidad de grupo, la construcción de estrategias colaborativas y la puesta en común de los recursos” (pág. 119, Banks, 2006).

Perspectiva social del tiempo

“Se refiere a la noción de que el tiempo debe ser considerado en términos de las actividades sociales en las que se está comprometido, más que en el sentido estricto de cumplir con horarios rígidos. El énfasis está centrado en las tradiciones y costumbres del pasado que guían eventos futuros” (pág. 119, Banks, 2006).

Oralidad

De acuerdo con Banks (2006) se refiere al “valor dado al conocimiento que es obtenido o transmitido a través de medios de comunicación oral. Las formas de comunicación oral son consideradas más apropiadas porque se piensa que son mucho mejores para transmitir significados que las palabras escritas” (pág. 119).

Armonía

“Enfatiza la importancia de dar homenaje a un ser supremo que controla todos los aspectos del mundo. Hay una creencia de que los humanos y lo natural y lo súper natural están interconectados. Más que controlar ciertos aspectos del medio ambiente, los individuos son pensados para ajustarse al universo adoptando creencias metafísicas tales como la suerte o el destino para explicar las circunstancias de la vida” (pág. 119, Banks, 2006).

Objetos poco conocidos o usados por el grupo

Por ejemplo, de acuerdo con Tellegen & Laros (2004) un ítem puede considerarse sesgado si uno de los grupos muestra problemas con el reconocimiento de una determinada foto o gráfico mientras que el otro grupo no. También puede considerarse si se usan diseños viejos de los objetos o que no tienen un uso amplio, el uso de figuras que son difíciles de reconocer u objetos que presentan diseños ampliamente diferentes entre las dos culturas.

Forma en la que se toman las decisiones

En algunas investigaciones realizadas con nativos norteamericanos (Rhodes, 1988, citado por Snetzler & Qualls, 2000) se ha encontrado que pruebas que implican respuestas rápidas, adivinación, tomar riesgos y la eliminación de opciones en preguntas de selección múltiple contradicen el proceso de toma de decisiones de estas personas puesto que en su cultura las decisiones se toman de forma lenta y segura y están basadas en combinar elementos a partir de diferentes opciones posibles en una respuesta apropiada.

Experiencias más frecuentes en un grupo que en otro

La frecuencia de las experiencias o actividades a las que están expuestas las personas en su vida diaria dependen del grupo al que pertenecen.

Estilos de aprendizaje

La forma en que aprenden las personas puede depender del grupo cultural al que pertenecen. Por ejemplo, “los adultos y los niños pueden compartir la misma actividad por períodos largos de tiempo en el cual hay interacciones verbales pero no están centrados en el proceso de aprendizaje. Este estilo puede no ser óptimo en un sistema escolar occidental tradicional que organiza la enseñanza alrededor de períodos de clase cortos y frecuentes en los cuales se espera que los estudiantes escuchen pasivamente a los profesores, sigan instrucciones y brinden largas respuestas verbales a las preguntas (Lipka, 1998)” (pág. 556, Solano y Nelson, 2001).

Epistemología de las comunidades de origen

La forma en la que se realizan las preguntas puede ser contraria o diferente a la forma como se concibe el mundo y el conocimiento. También puede asociarse con las experiencias vividas por el evaluado que hacen que la respuesta que da sea correcta en su contexto cultural.

Creencias

Las creencias derivadas del hogar o comunidades de los evaluados implican supuestos culturales subyacentes a sus respuestas.

Interferencia fonológica, ortográfica o semántica

Existen aspectos fonológicos, ortográficos o semánticos de la primera lengua de los evaluados que causan interferencia en sus respuestas.

Género de los textos

El género de los textos puede llevar a que no se brinde la respuesta esperada a una pregunta, por ejemplo si se pregunta por la conclusión de un problema basado en una historia, la respuesta que brinda el evaluado puede ser simplemente cómo terminaría la historia más no la solución al problema planteado. Por ejemplo, el texto es de tipo científico, pero es considerado por el evaluado como literario.

Pragmatismo

Los evaluados “necesitan reconocer la estructura pragmática bajo la cual los ítems deben ser interpretados en lugar de simplemente tomar las palabras por su valor nominal” (pág. 914, Luykx, Lee, Mahotiere, Lester, Hart & Deaktor, 2007).

Organización de los textos

Las convenciones textuales y gráficas pueden interferir con las respuestas de los evaluados y no estar relacionadas con lo que se quiere medir. También pueden interferir aspectos como textos en negrita, agrupación jerárquica de preguntas relacionadas o formato de las preguntas, etc. (Luykx et al, 2007).

Anexo 11. Número de veces que cada ítem fue detectado con DIF a través de las muestras por las pruebas estadísticas y las métricas de los procedimientos utilizados.

Prueba e Ítem	II Aplicación de 2006					I Aplicación de 2007						
	Prueba estadística* de MH	Categoría B** del ΔMH	Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	Prueba estadística* de MH	Categoría B y C del ΔMH		Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	
							B	C				
Biología												
1	21	0	21	17	0	0	0	0	0	0	0	0
2	21	0	21	0	0	3	0	0	16	3	0	0
3	2	0	19	0	0	1	0	0	2	0	0	0
4	21	0	21	3	0	20	1	0	21	20	1	1
5	21	0	21	0	0	1	0	0	1	0	0	0
6	18	0	21	0	0	9	0	0	5	0	0	0
7	21	0	21	0	0	20	4	0	21	19	4	4
8	18	0	11	0	0	20	6	0	21	11	6	6
9	1	0	1	0	0	13	0	0	7	7	0	0
10	0	0	0	0	0	21	5	0	21	21	5	5
11	16	0	8	3	0	2	0	0	7	0	0	0
12	21	0	20	0	0	3	0	0	3	0	0	0
13	21	0	21	0	0	21	2	0	21	13	2	2
14	20	0	20	0	0	0	0	0	0	0	0	0
15	21	0	21	5	0	1	0	0	3	0	0	0
16	21	0	21	0	0	16	0	0	20	13	0	0
17	12	0	9	0	0	1	0	0	1	0	0	0
18	0	0	0	0	0	4	0	0	13	0	0	0
19	7	0	21	0	0	6	0	0	6	0	0	0

Prueba e ítem	II Aplicación de 2006					I Aplicación de 2007					
	Prueba estadística* de MH	Categoría B** del ΔMH	Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	Prueba estadística* de MH	Categoría B y C del ΔMH		Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas
							B	C			
20	19	0	21	0	0	5	0	0	2	0	0
21	21	0	21	0	0	2	0	0	3	0	0
22	5	0	4	0	0	5	0	0	11	0	0
23	10	0	2	0	0	11	0	0	12	1	0
24	0	0	0	0	0	21	9	0	21	21	9
Filosofía											
1	3	0	5	0	0	4	0	0	8	0	0
2	21	0	21	17	0	5	0	0	5	0	0
3	21	0	21	3	0	11	0	0	11	0	0
4	11	0	1	0	0	19	0	0	21	0	0
5	21	0	21	2	0	19	1	0	20	6	1
6	0	0	6	0	0	2	0	0	2	0	0
7	20	0	20	0	0	15	0	0	14	1	0
8	21	0	21	0	0	3	0	0	2	0	0
9	1	0	3	0	0	11	0	0	10	0	0
10	16	0	21	0	0	17	1	0	17	2	1
11	0	0	16	0	0	21	6	0	21	7	5
12	19	0	20	0	0	19	0	0	19	0	0
13	14	0	17	0	0	2	0	0	2	0	0
14	6	0	18	0	0	2	0	0	3	0	0
15	2	0	1	0	0	6	0	0	10	0	0
16	2	0	4	0	0	11	0	0	11	0	0
17	21	0	21	0	0	11	0	0	12	0	0
18	4	0	7	0	0	3	0	0	3	0	0

Prueba e Ítem	II Aplicación de 2006					I Aplicación de 2007					
	Prueba estadística* de MH	Categoría B** del ΔMH	Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	Prueba estadística* de MH	Categoría B y C del ΔMH		Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas
							B	C			
19	21	0	21	8	0	16	0	0	16	0	0
20	15	0	19	0	0	10	0	0	11	0	0
21	21	0	21	0	0	21	3	0	21	3	3
22	3	0	17	0	0	2	0	0	2	0	0
23	15	0	20	0	0	3	0	0	2	0	0
24	12	0	21	0	0	5	0	0	5	0	0
Física											
1	18	0	20	0	0	3	0	0	3	0	0
2	20	0	21	0	0	2	0	0	3	0	0
3	19	0	17	0	0	12	0	0	15	3	0
4	3	0	3	0	0	2	0	0	0	0	0
5	20	0	19	0	0	3	0	0	3	0	0
6	21	0	21	0	0	1	0	0	2	0	0
7	21	1	21	21	1	10	0	0	10	0	0
8	6	0	6	0	0	21	0	0	21	14	0
9	5	0	9	0	0	14	0	0	14	0	0
10	1	0	1	0	0	2	0	0	3	1	0
11	1	0	0	0	0	8	0	0	8	1	0
12	1	0	3	0	0	0	0	0	1	0	0
13	21	0	21	21	0	5	0	0	3	0	0
14	21	0	21	0	0	15	0	0	13	1	0
15	14	0	7	0	0	15	1	0	14	1	1
16	21	0	21	0	0	16	3	0	18	0	0
17	21	0	21	0	0	2	0	0	2	0	0

Prueba e Ítem	II Aplicación de 2006					I Aplicación de 2007					
	Prueba estadística* de MH	Categoría B** del ΔMH	Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	Prueba estadística* de MH	Categoría B y C del ΔMH		Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas
							B	C			
18	9	0	11	0	0	15	2	0	15	0	0
19	14	0	13	0	0	18	0	0	17	0	0
20	21	0	21	0	0	0	0	0	0	0	0
21	1	0	2	0	0	12	0	0	13	0	0
22	7	0	9	0	0	5	0	0	5	0	0
23	1	0	2	0	0	1	0	0	1	0	0
24	21	0	21	1	0	3	0	0	3	0	0
Lenguaje											
1	21	0	21	0	0	0	0	0	4	0	0
2	21	0	21	17	0	14	0	0	11	0	0
3	21	0	21	0	0	1	0	0	1	0	0
4	3	0	9	0	0	21	2	0	21	16	2
5	3	0	3	0	0	15	0	0	20	0	0
6	0	0	3	0	0	21	14	4	21	18	18
7	1	0	13	0	0	14	0	0	14	0	0
8	15	0	14	0	0	16	0	0	20	0	0
9	21	0	21	0	0	19	0	0	21	1	0
10	21	1	21	5	1	8	0	0	13	0	0
11	8	0	2	0	0	21	1	0	21	3	1
12	13	0	8	0	0	15	0	0	20	0	0
13	11	0	20	0	0	3	0	0	2	0	0
14	15	0	20	0	0	17	3	0	9	0	0
15	2	0	2	0	0	4	0	0	4	0	0
16	21	0	21	20	0	11	0	0	18	0	0

Prueba e Ítem	II Aplicación de 2006						I Aplicación de 2007					
	Prueba estadística* de MH	Categoría B** del ΔMH	Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	Prueba estadística* de MH	Categoría B y C del ΔMH		Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	
							B	C				
17	2	0	3	0	0	14	6	0	8	0	0	
18	2	0	4	0	0	12	0	0	0	0	0	
19	9	0	16	5	0	21	5	0	21	0	0	
20	0	0	2	0	0	21	14	0	21	12	12	
21	21	0	21	0	0	4	0	0	3	0	0	
22	19	0	1	0	0	3	0	0	7	0	0	
23	21	0	21	1	0	12	0	0	19	1	0	
24	8	0	15	0	0	13	0	0	20	1	0	
Matemáticas					0							
1	17	0	19	0	0	2	0	0	2	0	0	
2	16	0	21	0	0	21	8	3	21	13	11	
3	7	0	16	0	0	10	0	0	20	0	0	
4	21	0	21	0	0	2	0	0	1	0	0	
5	21	0	21	19	0	1	0	0	0	0	0	
6	10	0	17	0	0	3	0	0	11	0	0	
7	21	0	21	16	0	2	0	0	11	0	0	
8	21	2	21	0	0	18	0	0	18	0	0	
9	21	0	21	0	0	1	0	0	0	0	0	
10	21	0	12	0	0	2	0	0	1	0	0	
11	6	0	5	0	0	18	1	0	20	2	1	
12	12	0	2	0	0	16	0	0	20	0	0	
13	2	0	15	0	0	21	0	0	21	0	0	
14	11	0	2	0	0	1	0	0	1	0	0	
15	17	0	21	0	0	3	0	0	6	0	0	

Prueba e ítem	II Aplicación de 2006						I Aplicación de 2007					
	Prueba estadística* de MH	Categoría B** del ΔMH	Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	Prueba estadística* de MH	Categoría B y C del ΔMH		Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	
							B	C				
16	4	0	4	0	0	14	0	0	7	0	0	
17	17	0	21	0	0	1	0	0	8	0	0	
18	8	0	21	0	0	0	0	0	9	0	0	
19	13	0	21	0	0	0	0	0	0	0	0	
20	21	0	21	1	0	14	1	0	21	1	1	
21	4	0	4	0	0	1	0	0	1	0	0	
22	18	0	21	0	0	18	0	0	21	0	0	
23	19	0	16	0	0	2	0	0	2	0	0	
24	1	0	7	0	0	5	0	0	14	0	0	
Química												
1	21	0	21	0	0	3	0	0	4	0	0	
2	6	0	18	0	0	16	0	0	14	0	0	
3	21	0	21	3	0	7	0	0	11	0	0	
4	14	0	16	0	0	4	0	0	4	0	0	
5	21	0	21	13	0	7	0	0	11	1	0	
6	1	0	1	0	0	1	0	0	1	0	0	
7	21	0	21	0	0	3	0	0	3	2	0	
8	15	0	19	0	0	21	5	0	21	21	5	
9	8	0	7	0	0	8	0	0	7	0	0	
10	2	0	5	0	0	8	0	0	6	0	0	
11	1	0	4	0	0	21	0	0	21	3	0	
12	2	0	0	0	0	21	10	0	21	9	8	
13	7	0	15	0	0	1	0	0	4	0	0	
14	21	0	21	12	0	11	0	0	11	0	0	

Prueba e Ítem	II Aplicación de 2006						I Aplicación de 2007					
	Prueba estadística* de MH	Categoría B** del ΔMH	Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	Prueba estadística* de MH	Categoría B y C del ΔMH		Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	
							B	C				
15	2	0	2	0	0	12	0	0	14	0	0	
16	13	0	13	0	0	2	0	0	2	0	0	
17	3	0	2	0	0	4	0	0	4	0	0	
18	7	0	7	0	0	11	0	0	12	0	0	
19	16	0	21	0	0	2	0	0	1	0	0	
20	15	0	18	1	0	0	0	0	0	0	0	
21	21	0	21	0	0	18	2	0	18	0	0	
22	5	0	4	0	0	1	0	0	3	0	0	
23	21	0	21	21	0	1	0	0	0	0	0	
24	21	0	21	5	0	5	0	0	7	1	0	
Sociales												
1	7	0	11	0	0	10	0	0	10	0	0	
2	7	0	19	0	0	16	0	0	19	0	0	
3	13	0	1	0	0	10	0	0	13	0	0	
4	14	0	3	0	0	2	0	0	2	0	0	
5	2	0	16	0	0	19	1	0	21	0	0	
6	21	0	21	0	0	17	1	0	17	0	0	
7	21	0	21	0	0	6	0	0	4	0	0	
8	4	0	3	0	0	0	0	0	0	0	0	
9	21	21	21	19	19	11	0	0	19	0	0	
10	21	1	21	1	1	13	0	0	19	0	0	
11	21	0	18	0	0	8	0	0	9	0	0	
12	21	0	21	0	0	8	0	0	6	0	0	
13	17	0	21	0	0	17	0	0	17	0	0	

Prueba e Ítem	II Aplicación de 2006					I Aplicación de 2007					
	Prueba estadística* de MH	Categoría B** del ΔMH	Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas	Prueba estadística* de MH	Categoría B y C del ΔMH		Prueba estadística de diferencia de la dificultad	Métrica de diferencia de la dificultad	Ambas métricas
							B	C			
14	0	0	1	0	0	4	0	0	17	0	0
15	21	0	21	0	0	10	0	0	16	0	0
16	20	0	20	0	0	2	0	0	2	0	0
17	21	0	21	0	0	5	0	0	3	0	0
18	1	0	8	0	0	17	0	0	11	0	0
19	7	0	13	0	0	10	0	0	7	0	0
20	15	0	16	0	0	20	5	0	19	0	0
21	3	0	17	0	0	21	6	0	18	5	5
22	18	0	21	0	0	11	0	0	17	0	0
23	21	0	21	0	0	10	0	0	19	0	0
24	21	0	21	0	0	12	0	0	12	0	0
25	1	0	7	0	0	19	1	0	18	1	1
26	21	0	21	0	0	17	0	0	20	1	0
27	16	0	5	0	0	4	0	0	2	0	0
28	0	0	1	0	0	3	0	0	7	0	0
29	21	0	21	0	0	21	3	0	21	3	3
30	1	0	10	0	0	1	0	0	1	0	0

Nota: * $p < 0.05$, ** No hubo ítems detectados en la Categoría C del ΔMH.

Anexo 12. Promedio de las dificultades de los ítems con DIF en las muestras en las que fueron detectados por alguna métrica, total y para cada uno de los grupos.

Aplicación y Prueba	Ítem	Dificultad grupo focal	Dificultad del grupo de referencia	Diferencia de las dificultades	Dificultad Total	Discriminación*
II Aplicación de 2006						
Biología	1	-0.92	-1.29	0.37	-1.29	1.25
	4	0.27	0.72	-0.44	0.71	0.9
	11	-1.15	-1.01	-0.15	-1.01	1.13
	15	0.22	0.66	-0.44	0.66	0.77
Filosofía	2	-0.48	-0.10	-0.37	-0.10	0.74
	3	-2.07	-2.30	0.24	-2.31	1.03
	5	-0.19	-0.56	0.37	-0.56	1.54
	19	-0.15	0.26	-0.41	0.26	0.93
Física	7	-1.02	-0.55	-0.47	-0.55	0.49
	13	-0.99	-0.65	-0.34	-0.65	0.76
	24	-0.29	-0.62	0.33	-0.62	1.18
Lenguaje	2	-1.12	-1.51	0.40	-1.51	1.18
	10	0.25	-0.15	0.40	-0.15	1.01
	16	-1.37	-1.12	-0.25	-1.12	0.78
	19	-1.69	-1.57	-0.12	-1.57	0.99
	23	-0.45	-0.83	0.38	-0.83	1.38
Matemáticas	5	0.94	0.52	0.42	0.52	1.13
	7	0.10	-0.31	0.40	-0.31	1.34
	8	-1.65	-2.12	0.47	-2.11	1.1
	20	0.29	0.86	-0.56	0.85	0.93
Química	3	-0.67	-0.41	-0.26	-0.41	0.96
	5	-0.03	0.28	-0.31	0.28	0.95
	14	0.21	0.53	-0.32	0.53	1.07
	20	-0.50	-0.71	0.21	-0.71	1.37
	21	-0.47	-0.17	-0.30	-0.17	0.88

Aplicación y Prueba	Ítem	Dificultad grupo focal	Dificultad del grupo de referencia	Diferencia de las dificultades	Dificultad Total	Discriminación*
	23	0.08	-0.30	0.38	-0.30	1.24
	24	-0.16	0.14	-0.30	0.14	0.87
Sociales	9	-1.02	-1.57	0.55	-1.56	1.16
	10	0.46	0.00	0.46	0.01	1.28
I Aplicación de 2007						
Biología	2	-0.65	-0.35	-0.30	-0.35	0.53
	4	-0.59	-0.21	-0.37	-0.22	0.66
	7	0.13	0.63	-0.50	0.62	0.82
	8	0.46	0.97	-0.51	0.96	0.94
	9	-1.38	-1.17	-0.21	-1.17	1.46
	10	-0.51	-0.92	0.41	-0.91	1.44
	13	0.40	0.04	0.36	0.04	0.98
	16	0.33	-0.04	0.36	-0.03	1.36
	20	-0.32	-0.04	-0.28	-0.04	1.04
	23	0.36	0.00	0.36	0.00	1.04
	24	0.28	-0.17	0.45	-0.16	1.19
Filosofía	5	-0.70	-0.30	-0.40	-0.31	0.94
	7	-0.44	-0.84	0.40	-0.84	1.42
	10	-1.11	-0.73	-0.39	-0.73	0.91
	11	0.34	0.87	-0.53	0.86	0.88
	21	-0.01	-0.44	0.43	-0.44	1.19
Física	3	0.37	0.72	-0.34	0.72	0.98
	8	-0.34	-0.64	0.31	-0.64	1.01
	10	0.48	0.83	-0.35	0.83	0.99
	11	-0.57	-0.23	-0.34	-0.23	1.04
	14	-0.87	-0.49	-0.38	-0.50	1.12
	15	-0.89	-0.48	-0.41	-0.49	1.19
	16	1.43	0.97	0.45	0.98	1
	18	1.16	0.68	0.48	0.69	1.01
Lenguaje	4	0.01	0.57	-0.56	0.56	0.58
	6	0.05	-0.51	0.56	-0.51	1.53

Aplicación y Prueba	Ítem	Dificultad grupo focal	Dificultad del grupo de referencia	Diferencia de las dificultades	Dificultad Total	Discriminación*
	9	-0.03	-0.46	0.43	-0.45	1.6
	11	-0.86	-0.37	-0.50	-0.37	0.72
	14	1.95	1.61	0.34	1.61	0.93
	17	2.01	1.64	0.38	1.64	0.95
	19	1.48	1.01	0.46	1.02	0.99
	20	-1.89	-1.33	-0.55	-1.34	0.99
	23	0.21	0.69	-0.48	0.68	0.79
	24	-0.46	-0.84	0.37	-0.83	1.28
Matemáticas	2	0.93	0.28	0.65	0.29	1.25
	11	-1.00	-0.54	-0.46	-0.54	0.84
	20	0.19	0.75	-0.56	0.74	0.89
Química	5	-0.13	-0.39	0.27	-0.39	1.34
	7	-0.46	-0.66	0.21	-0.66	0.89
	8	-0.08	-0.46	0.38	-0.45	1.3
	11	-0.62	-0.26	-0.36	-0.26	0.73
	12	-2.10	-1.60	-0.51	-1.60	1.08
	21	1.40	0.89	0.52	0.89	0.93
	24	0.52	0.83	-0.32	0.83	0.94
Sociales	5	0.78	0.32	0.45	0.33	1.04
	6	-0.13	0.34	-0.47	0.34	0.96
	20	0.00	0.34	-0.34	0.33	1.32
	21	-1.50	-1.13	-0.37	-1.13	1.21
	25	-0.51	-0.94	0.43	-0.93	0.89
	26	-0.61	-1.02	0.40	-1.01	1.29
	29	0.15	-0.34	0.48	-0.33	1.72

Nota: *La discriminación corresponde a la estimada para el ítem a partir de la población.

Anexo 13. Algunas pautas para evitar sesgo cultural en los ítems

Introducción

El sesgo se puede definir como “error sistemático que distorsiona el significado de las puntuaciones y que está causado por la intervención de habilidades espurias junto a la habilidad principal en un ítem” (Ackerman, 1992; Mellenbergh, 1989; Shealy & Stout, 1993; citados por Elosua, López, & Torres, 2000, p. 198), y dado que es un error sistemático se puede incluir su evaluación en el análisis de validez. Se cree comúnmente que el sesgo se debe a alguna característica no deliberada del ítem de la prueba (Banks, 2006) que da una ventaja injusta a un grupo de examinados sobre otro (Clauser & Mazor, 1998).

Solano & Nelson (2001) plantean el término validez cultural en el contexto de las pruebas de ciencias, el cual se define como la efectividad con la que dichas evaluaciones “abordan las influencias socioculturales que forman el pensamiento del estudiante y la forma en la que los estudiantes hallan el sentido a los ítems de ciencias y los responden” (p. 555). Una forma de abordar las influencias socioculturales en las pruebas es identificar si existen potenciales fuentes irrelevantes de funcionamiento diferencial en los ítems (DIF), es decir, sesgo. Con DIF nos referimos a que un ítem presenta propiedades psicométricas (como dificultad y discriminación) diferentes entre dos grupos que se están comparando. Dicho DIF se presenta cuando evaluados de, por lo menos, dos grupos diferentes que cuentan con la misma magnitud de atributo, tiene probabilidad distinta de responder correctamente el ítem debido a su pertenencia a un grupo en particular.

En el análisis de DIF y de sesgo, siempre hay como mínimo dos grupos, el focal y el de referencia. El primero corresponde al grupo minoritario de la población o que tradicionalmente se ha considerado como desfavorecido y el segundo al mayoritario y que se considera como el favorecido. Para Banks (2006) los ítems sesgados culturalmente tienen características que no están relacionadas con el desempeño en el constructo que está siendo medido, sino que son sensibles a grupos culturales particulares y afectan su

ejecución. En el caso de pruebas de selección múltiple, ciertos grupos culturales pueden interpretar el ítem o las opciones de respuesta en formas que no fueron anticipadas durante la construcción de la prueba y, de acuerdo con ello, ser atraídos por distractores que contienen estímulos culturales específicos (Veale & Foreman, 1983).

Sin embargo debe tenerse en cuenta que si un ítem es detectado con DIF, no quiere decir que inmediatamente esté sesgado, ello depende de un análisis substantivo del ítem en el que se busca observar si la posible fuente de DIF es relevante o irrelevante frente a lo que se pretende medir, porque en el primer caso se habla de impacto (diferencias reales entre los grupos) y en el segundo de sesgo. Para entender un poco estos conceptos, se presenta el siguiente ejemplo de Zieky (1992) teniendo en cuenta estos dos ítems:

- 1) ¿Cuánto es $5.3 \times 1,000$?
 - a. 53
 - b. 530
 - c. 5,300
 - d. 53,000
- 2) ¿Cuántos metros hay en 5.3 kilómetros?
 - a. 53
 - b. 530
 - c. 5,300
 - d. 53,000

Si estos ítems están pretendiendo medir habilidades para multiplicar y presentan DIF, en el primer caso el ítem no presenta sesgo o es justo. Sin embargo el segundo puede ser el ejemplo contrario, porque conocer el sistema métrico no es parte de lo que se pretende medir y existen grupos de personas que están menos familiarizadas con el sistema métrico presentado en el ítem 2. Dicho sesgo desaparecería en el segundo ítem, si éste pretendiera medir habilidades de conversión en el sistema métrico al que hace referencia la pregunta (Zieky, 1992).

Este documento tiene como propósito ilustrar algunas fuentes potenciales de sesgo cultural que fueron propuestas en un estudio que elaboró el grupo de investigación Métodos e Instrumentos de Investigación en Ciencias del Comportamiento de la Universidad Nacional de Colombia con base en los ítems del examen SABER 11°. En dicho estudio se compararon evaluados indígenas y no indígenas y se revisaron los ítems que habían sido detectados con DIF. Cuenta con sugerencias para evitar el sesgo cultural, está inspirado en el elaborado por el ETS en 2009 y presenta siete fuentes posibles de sesgo, a partir de las cuales se pueden revisar los ítems que presenten algún tipo de DIF. Es de reiterar que estas fuentes sólo pueden ser consideradas como posibles causas irrelevantes de DIF si lo que representan o definen no está relacionado con lo que pretende medir el instrumento. Aunque el ETS sugiere una clasificación de las fuentes de varianza irrelevante en tres grandes categorías: fuentes cognitivas, fuentes emocionales y fuentes físicas, aquí únicamente se hablará de las dos primeras.

Fuentes Cognitivas

Las fuentes cognitivas se presentan cuando para responder correctamente el ítem se requiere de conocimientos o habilidades que no están relacionados con el propósito de la prueba y que no están distribuidos equitativamente entre los grupos (ETS, 2009). Entre las fuentes cognitivas se encuentran: las experiencias más frecuentes en un grupo que en otro, problemas de construcción, tecnicismos y escuela tradicional.

Experiencias más frecuentes en un grupo que en otro

Se caracteriza por exposiciones a actividades, temas o contextos que son particularmente más propios de un grupo y que están presentes diferencialmente en el diario vivir de los evaluados. Se destacan aspectos como el contacto con animales, actividades como hacer viajes en carretera (pasear) que permiten conocer mejor las convenciones y nombres dados a ciertos lugares (capital, municipio, población, peaje), realizar experimentos y estar en contacto de forma frecuente con deportes como el fútbol. No obstante, si estos aspectos son relevantes para lo que se desea medir, no deben ser

considerados como factores potenciales de sesgo. La tabla 1 menciona ejemplos de estos elementos específicos y ofrece sugerencias de cómo tratarlos.

Tabla 1. Aspectos asociados con las experiencias más frecuentes en un grupo que en otro

Aspecto	Forma propuesta para evitar posible sesgo
Contacto con animales	Hacer referencia solamente al animal como tal, evitando historias sobre éste cuando son innecesarias. Elaborar los ítems sobre animales, con ejemplos de ellos que sean de conocimiento común entre los diversos grupos culturales.
Actividades específicas como pasear	Ser claros en usar palabras como ciudad, municipio y población, señalando si son diferentes o iguales. También aclarar lo que significan palabras como peaje, cuando es necesario para responder la pregunta y no está relacionado con el objeto de medida.
Experimentos	No usar contextos de experimentos, sino exponer la misma situación de forma más general, siempre y cuando no esté relacionado con el objeto de medida.
Deportes	Cuando se usan contextos de deportes, no hacer énfasis en aspectos que no son relevantes ni necesarios para responder la pregunta, por lo que se propone que los contextos sean más generales.

Problemas de construcción

Se refiere a los siguientes aspectos:

- a. Extensión y complejidad de los pasajes que sirven de contextos a las preguntas: si éstos son muy largos, al considerar el español como segunda lengua, puede aumentar la dificultad para algunos grupos.
- b. Opciones de respuesta desbalanceadas: una de las opciones de respuesta es de una categoría diferente a las otras, lo que se puede presentar en alguna de las siguientes formas: a) una de las opciones es de una categoría semántica diferente, por ejemplo, sólo una de ellas se refiere a un movimiento literario y las demás provienen del sentido común, o una de ellas se refiere a un sujeto y las demás a objetos; b) todas tienen una palabra en particular y otra no, la cual puede ser o no la

- respuesta correcta, sin embargo hace que los evaluados se desvíen hacia ella; y c) opciones mucho más largas o cortas que las otras.
- c. La respuesta correcta no se deriva del contexto y enunciado, sino de un conocimiento adicional, incluso, éstos pueden llevar a la selección de una respuesta errónea por comprensión de lectura cuando no se está evaluando este atributo.
 - d. Palabras que están en el enunciado se escriben de la misma forma en las opciones de respuesta y eso hace que los evaluados se sientan atraídos por alguna(s) de ellas, favoreciéndolos cuando esa es la respuesta correcta y desfavoreciéndolos cuando no.
 - e. Conceptos que no tienen un significado unívoco y que hacen que la posibilidad de respuesta sea muy abierta.
 - f. Convenciones usadas en las gráficas: existen convenciones en las gráficas que exigen tener un conocimiento o habilidad adicional que no se encuentra relacionado con lo que se está midiendo, llegando incluso a que se tome cualquier valor extremo. Por ejemplo, en el estudio se encontró que no es claro qué significa “0% digestión” o “100% digestión” pues el primero puede indicar que no hubo absolutamente digestión mientras que 100% señala que la hubo completamente o viceversa, es decir, el primero indicar que se digirió todo y no quedó nada y el segundo que queda todo por digerir.

La tabla 2 muestra estos elementos y la forma sugerida de evitarlos.

Tabla 2. Aspectos asociados con problemas de construcción

Aspecto	Forma propuesta para evitar posible sesgo
Extensión y complejidad de los contextos	Hacer los contextos con una redacción lo más sencilla posible y de corta extensión.
Opciones de respuesta desbalanceadas	Diseñar las opciones de respuesta de forma tal que todas pertenezcan a la misma categoría, o en caso de no ser posible por lo menos a dos categorías diferentes, por ejemplo, dos objetos y dos elementos abstractos. También debe considerarse la extensión de las opciones, no puede existir una extremadamente larga o corta.
Respuesta correcta que no	Revisar si la información expuesta en el ítem o contexto

Aspecto	Forma propuesta para evitar posible sesgo
se deriva del enunciado	es suficiente para responder la pregunta y observar si por sola comprensión de lectura se daría la respuesta correcta o bien llevaría a elegir una opción incorrecta.
Palabras en el enunciado que se repiten en las opciones de respuesta	Evitar que las palabras que se encuentren en el enunciado de la pregunta se repitan en las opciones de respuesta, a no ser que sea absolutamente necesario.
Conceptos equívocos	Hacer una descripción de los conceptos que tienen varios significados, indicando específicamente desde qué perspectiva se está tomando o cuál es su definición.
Convenciones	Cuando se usen convenciones en un mapa o en una gráfica, su significado y los objetos que representan deben estar claramente definidos, a no ser que el conocimiento de los mismos haga parte de lo que se quiere medir.

Tecnicismos

Se refiere al uso de palabras que requieren un conocimiento o una habilidad diferente al atributo que se está pretendiendo medir y que dificulta la respuesta a un grupo particular. Entre ellas pueden estar: rumiar, glándula sebácea, neumococo, branquial, sustancias tóxicas, rango, entre otros. Esta posible fuente de sesgo sólo aplica siempre y cuando el tecnicismo no se considere relevante para lo que se quiere medir. La tabla 3 describe algunos tecnicismos y una propuesta para abordarlos tratando de evitar posibles sesgos culturales, aunque en términos generales se sugiere hacer una descripción de las palabras que puedan ser muy técnicas y específicas para que el evaluado cuente con dicha información, siempre y cuando el conocimiento de la palabra no haga parte del objeto de medida.

Tabla 3. Aspectos asociados con tecnicismos

Palabras técnicas	Forma propuesta para evitar posible sesgo
Rango	Definir qué es rango, o reemplazarlo por su descripción o una palabra más cercana. Por ejemplo, si se habla de rangos de edad, puede decirse grupos de edad entre “x” valor y “y” valor.
Tóxico	Definir qué es tóxico.

Palabras técnicas	Forma propuesta para evitar posible sesgo
Canis Familiaris	Si no es necesario para responder las preguntas omitirlo del texto. Aunque el ejemplo aquí es sobre la taxonomía del perro, esto también aplica para los nombres científicos de cualquier organismo.
Tubos de ensayo	Se puede reemplazar por vasos, siempre y cuando el conocimiento de los tubos de ensayo no sea parte del objeto de medida.
Turberas	
Rumiar	
Neumococo	Definir el término, a no ser que su conocimiento haga parte del objeto de medida.
Branquial	
Hormonas	

Escuela tradicional

Es probable que en algunos grupos culturales se mantenga un tipo declarativo de enseñanza que es considerado tradicional y que hace que en ítems que exijan esa clase de conocimiento se vean favorecidos individuos particulares. En este caso, la sugerencia es revisar si ese conocimiento de tipo declarativo es el que se quiere evaluar y si no es así, modificar la pregunta hacia el objeto de medida. Además debe tenerse en cuenta que si un posible DIF en el ítem puede ser causado por este factor, es necesario evaluar si es debido a que el grupo no tuvo acceso al aprendizaje de su contenido o no se enseña en sus lugares de estudio, en cuyo caso si la pregunta está relacionada con lo que se quiere medir, sería un ítem justo.

Fuentes Emocionales

Las fuentes emocionales de varianza irrelevante se presentan cuando el lenguaje o las imágenes ocasionan emociones fuertes que pueden interferir con la habilidad para responder un ítem correctamente, por ejemplo contenido ofensivo, molesto o controversial, que dificulta que el evaluado se concentre en el significado de una lectura o la respuesta a un ítem, llegando incluso a contestarlo de forma emocional más que lógica (ETS, 2009).

Entre las fuentes emocionales propuestas en el estudio llevado a cabo por la Universidad Nacional de Colombia, se encuentran: la epistemología de las comunidades de origen, el colonialismo religión-Estado y los juicios de valor hacia teorías sociales o de mercado.

Epistemología de las comunidades de origen

Esta posible fuente de sesgo se podría presentar cuando la forma en que se realizan las preguntas y su contenido son contrarios o diferentes a cómo el evaluado concibe el mundo y el conocimiento. Específicamente se relaciona con aspectos como: costumbres que son más comunes en un grupo (p. ej. matrimonios entre parientes cercanos), diferente concepto de familia, formas de tratar las enfermedades, conceptos que no tienen una palabra equivalente en los idiomas de los grupos que se están comparando, los conceptos de orientación en el espacio, el manejo de la economía, las formas de gobierno y la forma en que se imparte justicia en diferentes grupos culturales, las relaciones de género, el manejo del fracaso, el pensamiento científico, la religión y la concepción de la naturaleza, entre otros. La tabla 4 describe algunos de estos aspectos y muestra sugerencias para tratar de evitar el sesgo cultural que puede introducir este factor.

Tabla 4. Aspectos relacionados con la epistemología de las comunidades

Aspecto	Forma propuesta para evitar posible sesgo
Costumbres que son más comunes en un grupo	Evitar que las preguntas o sus contextos puedan dar a entender que se están juzgando las costumbres de un grupo particular o que dichas costumbres tienen consecuencias negativas.
Formas de tratar las enfermedades	Hacer que las preguntas que contengan información de este tipo sean lo suficientemente generales para que no hagan referencia a la forma en que se trata una enfermedad en una cultura específica.
Conceptos que no tienen una palabra para designarlos en otro idioma	Si no se tiene la seguridad de que existe una palabra para designar un concepto o proceso en el lenguaje de origen de las personas que presentan la prueba, se debe definir esta palabra de forma clara en el ítem. Por ejemplo, es posible que la palabra libertad no tenga una equiparable en algunas lenguas.
Conceptos de orientación en el espacio	Se pueden usar mapas que ayuden en la orientación, siempre y cuando esto no afecte lo que se pretende

Aspecto	Forma propuesta para evitar posible sesgo
Forma en que se imparte justicia	<p>medir.</p> <p>Debe evitarse las preguntas en las que la forma en que se imparte justicia pueda ser diferente para distintos grupos culturales. En caso de que sea parte del objeto de medida, debe aclararse de qué contexto se está hablando (país, ciudad, grupo cultural, etc.) y definir la forma en que se imparte justicia.</p>
Concepción de la naturaleza	<p>Tener especial cuidado en aspectos que impliquen que se está atentando contra la naturaleza, como por ejemplo la explotación del carbón, pues estos temas pueden ser sensibles a grupos particulares.</p>

Colonialismo religión-Estado

Se refiere a aspectos relacionados con la época de la colonia y la experiencia que tuvieron los diferentes grupos culturales a partir de ese momento en relación con la iglesia y el Estado. En ese sentido, se sugiere eliminar opciones de respuesta que describan hechos o experiencias particulares de los grupos en relación al período de colonización, pues dada su historia, es probable que tiendan a elegirla así no sea la respuesta correcta. Esta sugerencia aplicaría solamente si se presenta interferencia con el objeto de medida.

Juicios de valor hacia teorías sociales y de mercados

Se refiere a los juicios subjetivos que deben elaborar los evaluados para responder una pregunta de forma correcta sobre algunos conceptos como igualdad y libertad, y teorías sociales y de mercado. Este tipo de preguntas deben evitarse porque implica hacer evaluaciones subjetivas sobre qué está bien o mal en términos de economía y prácticas sociales. Además es factible que ciertos grupos culturales hayan tenido que enfrentarse a dilemas particulares viéndose obligados a dar prevalencia a la libertad o a la igualdad, y eso hace que en su contexto la respuesta correcta pueda ser otra.

Otros aspectos a tener en cuenta en los ítems

Aunque no se presenten como elementos que pueden favorecer a un grupo en particular, si se debe tener en cuenta que algunos términos, palabras o gráficas pueden incomodar a ciertos grupos de la población. Entre estas palabras se destacan algunas que hacen referencias a características físicas como “gorda” e imágenes que pueden representar aspectos maléficos para algunas culturas. Para el primer caso se sugiere evitar el uso de estas palabras y en el segundo, tratar de tener entre los revisores de las pruebas finales personas provenientes de diferentes culturas que puedan aclarar esos aspectos.