icfes saber 5° y 9°

INFORMES

SABER 5o. y 9o. 2009
Informe técnico de resultados históricos

Elaboración del documento
Víctor H. Cervantes Botero
Carolina Lopera Oquendo
Luis Adrián Quintero Sarmiento

Revisor de estilo
Fernando Carretero Socha

Diagramación
Alejandra Guzmán Escobar

ISBN de la versión electrónica: 978-958-11-0577-9

Bogotá, D.C., diciembre de 2011

Advertencia

Con el fin de evitar la sobrecarga gráfica que supondría utilizar en español “o/a” para denotar uno u otro género, el ICFES opta por emplear el masculino genérico en el que todas las menciones de este se refieren siempre a hombres y mujeres.

Contenido

Lista de gráficos y tablas

Tablas

Gráficos

Presentación

SABER 5o. y 9o. es una evaluación nacional de carácter externo que se aplica periódicamente a estudiantes de educación básica de todo el país con el fin de conocer el desarrollo de sus competencias básicas en lenguaje, matemáticas y ciencias naturales. Sus resultados se han utilizado para orientar la definición de políticas y programas de mejoramiento.

La primera aplicación se realizó en 1991 a una muestra de estudiantes de trece departamentos. En 1993-1995, la aplicación tuvo representatividad nacional y regional y en 1997-1999 se llevó a cabo a una muestra nacional con representatividad departamental y de algunos municipios. Aunque estas produjeron información agregada sobre el rendimiento de los estudiantes, su carácter muestral no permitía obtener resultados para un nivel crucial de la gestión y la toma de decisiones del sistema: los establecimientos educativos.

La Ley 715 de 2001 estableció que esta evaluación tiene carácter obligatorio y censal, y debe realizarse cada tres años. Desde entonces se han efectuado tres aplicaciones, de acuerdo con los calendarios académicos (A y B) vigentes en las entidades territoriales: la primera entre 2002 y 2003, la segunda entre 2005 y 2006, y la tercera en 2009.

Como parte fundamental de la estrategia de mejoramiento de la calidad de la educación en el país, uno de los propósitos principales de la evaluación es generar información de alta calidad que permita establecer si se está avanzando en el logro educativo en el país. Sin embargo, a pesar de la continuidad de la evaluación censal de SABER 5o. y 9o., las características metodológicas y de aplicación con las que se desarrollaron estas evaluaciones no permiten realizar una comparación directa de los resultados reportados.

Por esta razón y con el fin de obtener resultados históricos comparables, el Instituto Colombiano para la Evaluación de la Educación (ICFES) realizó un estudio de equiparación que reporta los resultados de las evaluaciones en 2002-2003, 2005-2006 y 2009 en una única escala de calificación. Esta información permite establecer la evolución de los resultados en lenguaje y matemáticas para quinto y noveno grado, entre 2002 y 2009.

Este informe tiene como propósito mostrar los procedimientos técnicos llevados a cabo para la determinación de los resultados históricos de las aplicaciones censales de SABER 5o. y 9o. y presentar las limitaciones asociadas con la obtención de estos resultados.

El documento se compone de ocho secciones. La primera contiene las características generales de las aplicaciones censales de SABER 5o. y 9o.; la segunda presenta el diseño de la muestra para el estudio de equiparación; la tercera da cuenta de los cambios administrativos derivados de la integración institucional, que limitan la obtención de resultados históricos para todas las instituciones educativas evaluadas en cada aplicación; la cuarta detalla el procedimiento de detección de copia utilizado; la quinta expone las estadísticas utilizadas para evaluar la estabilidad de los ítems entre las aplicaciones originales y el estudio de equiparación; en la sexta se presentan los aspectos relacionados con la definición de la escala de resultados históricos; en la séptima se describe el procedimiento de equiparación, mediante el cual se derivan los resultados históricos. Finalmente, en la octava se presentan los tipos de resultados reportados.

La preparación de este informe estuvo a cargo de las subdirecciones de Estadística y de Análisis y Divulgación del ICFES. Adicionalmente, contó con el apoyo de los siguientes profesionales, cuya colaboración en el desarrollo del estudio de equiparación fue invaluable: Yanneth Castelblanco, John Jairo Rivera, Walter Manuel García, Pedro César del Campo, Wilmer Martínez y María Nelcy Rodríguez.

1. Caracterización de las aplicaciones censales de SABER 5o. y 9o.

A continuación se presentan las características generales que tuvieron las evaluaciones de SABER en las aplicaciones de los años 2002-2003, 2005-2006 y 2009.

La evaluación SABER 5o. y 9o. en 2002-2003, se desarrolló en cuatro momentos diferentes que se resumen en la tabla 1. Cada una de estas aplicaciones tiene características propias en cuanto a la estructura de los cuadernillos y a la población objetivo.

Tabla 1. Aplicaciones realizadas en la evaluación SABER 2002-2003

Aplicación	Descripción	Fecha Aplicación	Institución encargada aplicación
NS20021	Nuevo sistema escolar	Marzo 2002	MEN Convenio SEICORP
CN20021	• ICFES • PER* • 7 municipios de Cundinamarca • Nuevo sistema escolar - Costa Caribe • PER* Modelos educativos	Octubre 2002	• ICFES • MEN CONVENIO SEI-CORP • Secretaría de Educación de Cundinamarca
CN20022	Bogotá	Octubre 2002	Secretaría de Educación de Bogotá
CN20031	SABER	Abril 2003	Secretarías Departamentales y municipales

Elaboración: Dirección de Evaluación, ICFES
*PER: Programa de Educación Rural.

La primera se realizó en marzo de 2002, bajo el programa Nuevo Sistema Escolar y dirigida por el Ministerio de Educación. En esta aplicación se evaluó un cuadernillo para las áreas de lenguaje, matemáticas y ciencias naturales en quinto y noveno grado. Las pruebas para todas las áreas evaluadas en quinto estaban compuestas por 31 ítems. En noveno grado, la prueba de lenguaje constaba de 41 ítems; las de ciencias naturales y matemáticas tenían 36 ítems. En todos los casos, las preguntas fueron de selección múltiple con única respuesta, excepto una pregunta abierta.

La segunda aplicación se realizó en octubre de 2002 a cargo del ICFES y en esta se evaluaron las áreas de lenguaje y matemáticas en los grados tercero, quinto y noveno. En esta aplicación se contaba con un cuadernillo por área y grado. Los cuadernillos para quinto grado contenían 25 ítems de selección múltiple; los de noveno grado, 35 ítems. La tercera aplicación, también

realizada en octubre de 2002, estuvo a cargo de la Secretaría de Educación de Bogotá. En aquella se evaluaron las áreas de matemáticas y lenguaje, en los grados quinto y noveno, utilizando los mismos cuadernillos de prueba de la primera aplicación por el programa de Nuevo Sistema Escolar.

Finalmente, en abril de 2003, se evaluaron los grados quinto, séptimo y noveno. En esta aplicación se evaluaron las áreas de ciencias naturales y competencias ciudadanas; en séptimo, se evaluaron adicionalmente las áreas de matemáticas y lenguaje. Los cuadernillos de estas pruebas tienen la misma estructura, en términos de número de ítems a las aplicadas en octubre de 2002 por el ICFES. En la tabla 2 se recoge el número de estudiantes evaluados en cada una de las aplicaciones llevadas a cabo en la evaluación de 2002-2003.

Tabla 2. Estudiantes evaluados por aplicación de SABER 2002-2003

Aplicación	Grado	
	Quinto	Noveno
CN20021	123.447	79.450
CN20021	332.293	162.361
CN20022	2.380	
CN20031	223.377	117.787
Total	**681.497**	**359.598**

Elaboración: Dirección de Evaluación, ICFES.

En 2005-2006 se evaluaron las áreas de matemáticas, lenguaje, ciencias naturales, ciencias sociales y competencias ciudadanas en quinto y noveno grado. Esta aplicación constó de cinco (5) cuadernillos para cada área y grado. Cada uno de estos cuadernillos incluyó 12 ítems de selección múltiple de las áreas de lenguaje, matemáticas, ciencias naturales y sociales y tres (3) preguntas abiertas, para un total de 51[1]. En 2005 se emplearon tres de estos cinco cuadernillos y se denominaron formas 1, 2 y 3; en 2006 se aplicaron las formas 1, 4 y 5, correspondientes a los otros dos cuadernillos y el mismo cuadernillo de la forma 1 de 2005.

[1] El área de competencias ciudadanas fue evaluada en un cuadernillo aparte y contenía 110 preguntas.

La tabla 3 presenta el total de estudiantes evaluados en cada una de las formas aplicadas en 2005-2006.

Tabla 3. Estudiantes evaluados por aplicación de SABER 2005-2006

Tipo de cuadernillo (forma)	Grado	
	Quinto	Noveno
1	30.296	25.596
2	308.950	203.749
3	271.700	175.608
4	52.090	36.673
5	51.569	37.053
Total	**714.605**	**478.679**

Elaboración: Dirección de Evaluación, ICFES.

En 2009 se aplicaron pruebas en las áreas de lenguaje, matemáticas y ciencias naturales en los grados quinto y noveno, para un total de seis pruebas. Las preguntas se organizaron en subconjuntos y estos, a su vez, se estructuraron en cuadernillos –cuatro subconjuntos de preguntas por cuadernillo–, con un diseño balanceado que aseguró que todas las áreas fueran proporcionalmente evaluadas dentro de la población. A cada estudiante le correspondió contestar pruebas de dos de las tres áreas evaluadas. De esta forma, los estudiantes de quinto respondieron 84 o 96 preguntas y los de noveno, 108, en una sesión de 180 minutos de duración[2].

Todas las preguntas utilizadas en las pruebas eran de selección múltiple con única respuesta. El informe técnico de la aplicación censal de las pruebas SABER 5o. y 9o. en 2009 (Cervantes & Lopera, 2011) presenta los detalles de esta aplicación en profundidad.

[2] En quinto, el cuadernillo de lenguaje se organizó en torno a 36 preguntas o ítems, mientras que los de matemáticas y ciencias constaron de 48 ítems. En noveno, los cuadernillos de las tres áreas tuvieron 54 ítems. Los cuadernillos liberados de las pruebas están publicados y disponibles para consulta y descarga en la sección "Documentos" del sitio web http://www2.icfes.gov.co/saber59

2. Diseño de la muestra de equiparación

La selección de los estudiantes que presentaron la prueba se obtuvo de forma aleatoria, buscando la imparcialidad y confiabilidad de las pruebas de enlace. A continuación se presenta la descripción del marco muestral y del diseño de la muestra.

2.1 Marco muestral

El marco muestral para la aplicación de equiparación se conformó según los establecimientos educativos que participaron en las pruebas SABER 5o. y 9o. en 2009, y que adicionalmente cumplieron las siguientes condiciones:

- Ubicados en alguna de las siguientes ciudades: Bogotá, Medellín, Cali, Barranquilla y Cartagena[3].
- Contaban con más de 25 estudiantes en quinto y 25 estudiantes en noveno.
- Ofrecieran jornada mañana o completa.

2.2 Unidades de muestreo y método de selección

De acuerdo con la naturaleza jerárquica de la población de estudio se planteó un diseño en dos etapas de selección. Estas etapas fueron: selección de establecimientos y estudiantes.

2.2.1 Primera etapa

La primera etapa consistió en un muestreo proporcional al tamaño sin reemplazamiento estratificado de establecimientos educativos. Según la Ley 115 de febrero 8 de 1994, "se entiende por establecimiento educativo o institución educativa, toda institución de carácter estatal, privada o de economía solidaria organizada con el fin de prestar el servicio público educativo y la persona responsable de este es el rector".

Se utilizó como criterio de estratificación la distribución por cuartiles del promedio ponderado de los puntajes estandarizados en matemáticas y lenguaje de grado quinto y noveno dentro de cada establecimiento. La ponderación está dada por el número de estudiantes de cada grado. La distribución de la estratificación para los establecimientos del marco se presenta en la tabla 4. El estrato 1 corresponde a los establecimientos de menor desempeño y el 4, a los de mejor desempeño en el promedio ponderado.

[3] Estas ciudades presentan el mayor número de estudiantes evaluados en SABER 5o. y 9o. 2009.

Tabla 4. Distribución de la estratificación de primera etapa de selección

Estrato	Establecimientos	Estudiantes quinto	Estudiantes noveno	Total de estudiantes
1	190	25.422	19.550	44.972
2	311	35.328	29.619	64.947
3	511	54.226	51.409	105.635
4	540	67.430	69.519	136.949
Total	**1.552**	**182.406**	**170.097**	**352.503**

Elaboración: Dirección de Evaluación, ICFES.

2.2.2 Segunda etapa

En la segunda etapa se identificaron los estudiantes dentro de los establecimientos educativos seleccionados que cumplen las características requeridas para presentar la prueba. Luego se les asignó aleatoriamente una de las dos áreas.

2.3 Tamaños de muestra

Con base en la información de referencia proveniente de la pruebas SABER 5o. y 9o. 2009, y a fin de establecer los tamaños de muestra requeridos para obtener estimaciones de los parámetros de los ítems incluidos en la pruebas utilizadas en la aplicación de equiparación con determinado nivel de precisión, se buscó obtener un mínimo de 9.000 estudiantes por grado para la aplicación. Las tablas 5 y 6 registran los tamaños de muestra seleccionados por estrato y municipio.

Tabla 5. Distribución de la muestra por estratificación de primera etapa de selección

Estrato	Establecimientos	Estudiantes quinto	Estudiantes noveno
1	13	2.735	2.300
2	16	2.016	1.640
3	26	2.790	3.094
4	21	2.256	2.051
Total	**76**	**9.797**	**9.085**

Elaboración: Dirección de Evaluación, ICFES.

Tabla 6. Distribución de la muestra por municipio

Municipio	Establecimientos	Estudiantes quinto	Estudiantes noveno
Barranquilla	8	945	890
Bogotá D.C.	33	3.945	3.528
Cali	12	1.739	1.665
Cartagena	6	1.177	1.110
Medellín	17	1.991	1.892
Total	**76**	**9.797**	**9.085**

Elaboración: Dirección de Evaluación, ICFES.

3. Identificación de las instituciones de acuerdo con la organización administrativa 2009

Desde su implementación, el proceso de integración institucional en el país (Ley 715 de 2001, capítulo III) ha generado cambios progresivos en la estructura administrativa de las instituciones educativas del país. Por esta razón, para la producción de resultados históricos fue necesario identificar las instituciones que en las aplicaciones anteriores (2002-2003 y 2005-2006) constituían un establecimiento educativo diferente al registrado en la estructura administrativa de 2009; es decir, identificar qué instituciones permanecen como sede principal de un establecimiento educativo en 2009 y cuáles han pasado a ser sedes de un establecimiento educativo diferente.

El cruce de información de instituciones se hizo por medio de los códigos DANE (sede e institución) entre SABER 2002-2003, y SABER 5o. y 9o. 2009, y el código DANE de institución entre SABER 2005-2006, y SABER 5o. y 9o. 2009. En el caso de las aplicaciones realizadas en 2002 y 2003 fue necesario, además, identificar y validar sus códigos DANE, puesto que a partir de 2002 se implementaron cambios en ese código[4].

En las aplicaciones de las pruebas SABER 5o. y 9o., la unidad de aplicación es la sede – jornada (también denominado como sitio de aplicación). En total, en 2002-2003 se presentaron pruebas en 30.833 sedes jornada; en 2005-2006 fueron 47.140 sedes jornada y en 2009, 42.973. La tabla 7 muestra la cantidad y proporción de sedes jornadas de las aplicaciones 2002-2003 y 2005-2006 identificados según la estructura administrativa de 2009. La tabla 8 presenta la distribución de las sedes jornadas identificadas por grado. Finalmente, a partir de este ejercicio se identificaron 11.154 instituciones entre las aplicaciones de SABER 2002-2003 y 2009, y 14.836 instituciones entre 2005-2006 y 2009.

[4] El código DANE de las instituciones y sedes cambió de forma que los códigos que empezaban por "0" fueran remplazados. En la codificación actual ninguno de los códigos empieza por este carácter.

Tabla 7. Cantidad de sedes jornadas identificadas en la estructura 2009 por aplicación

Sedes-jornadas	2002-2003		2005	
	Frecuencia	Porcentaje (%)	Frecuencia	Porcentaje (%)
Identificadas	24.384	79,08	36.166	76,72
No identificadas	6.449	20,92	10.974	23,28
Total	**30.833**	**100**	**47.140**	**100**

Elaboración: Dirección de Evaluación, ICFES.

Tabla 8. Distribución de las sedes jornadas identificadas por grado

Grado ofrecido en la sede-jornada	2002-2003		2005	
	Frecuencia	Porcentaje (%)	Frecuencia	Porcentaje (%)
Quinto	6.673	27,37	10.174	28,13
Noveno	161	0,66	393	1,09
Quinto y noveno	17.550	71,97	25.599	70,78
Total	**24.384**	**100**	**36.166**	**100**

Elaboración: Dirección de Evaluación, ICFES.

4. Procedimiento de detección de indicios de copia en los resultados históricos

Para el procesamiento de los resultados históricos de las aplicaciones de SABER 5o. y 9o. en 2002-2003 y 2005-2006, se empleó una metodología similar a la utilizada en los análisis realizados en 2009 para la identificación de indicios de copia en estas aplicaciones (Martínez & Cervantes, 2010). La metodología se modificó buscando incorporar dos características en la misma. En primer lugar, se buscó que la estimación de la probabilidad $\hat{p}$, definida en ese documento como la proporción observada de evaluados en una muestra aleatoria de la población cuyas respuestas son diferentes en tantas o más como el evaluado más similar en su mismo sitio de presentación, tuviera en cuenta que en el análisis está incluyéndose únicamente un grupo de estudiantes que son más similares entre sí en el conjunto total de evaluados, y en segundo lugar, que el ajuste por respuestas correctas fuera más conservador en la decisión de considerar indicios de copia para un evaluado, debido a que al interior de un establecimiento educativo los estudiantes son más similares entre sí que respecto al conjunto total de evaluados.

Con este fin se ejecutaron las siguientes modificaciones a la metodología de copia implementada en SABER 5o. y 9o. 2009:

1. La muestra con la que se compara el estudiante se extrae de la siguiente manera:

   - Se identifican los colegios en los que hay al menos un estudiante con entre $rc - 3 \; y \; rc + 3$, donde rc es la cantidad de respuestas correctas del estudiante.
   - De estos colegios, se extrae una muestra aleatoria de tamaño 3.000.
   - De cada uno de los colegios seleccionados en el paso anterior, se selecciona uno de los estudiantes que se encuentra en el rango de respuestas correctas de manera aleatoria.

2. A partir de estos estudiantes se calcula la probabilidad $\hat{p}$ y el índice de copia I_S siguiendo el procedimiento señalado en (Martínez y Cervantes, 2010).

3. El procedimiento de ajuste por respuestas correctas fue remplazado por:

   - $C_S = I_S - redondear\left(6 \frac{rc - \overline{rc}}{S_{rc}}\right)$. Donde $\overline{rc}$ es el promedio de respuestas correctas de los estudiantes del salón y S_{rc} es la desviación estándar del número de respuestas correctas de estos.

- El estudiante se considera sospechoso de copia si tiene un índice C_S mayor o igual que 6, y se dice que el salón presenta copia masiva si más del 60% de sus estudiantes son sospechosos de copia.

En los establecimientos educativos que se presentan indicios de copia masiva en las áreas y grados evaluados en alguna de las aplicaciones anteriores de SABER 5o. y 9o., al no tener resultados confiables, no se produjeron resultados históricos comparables para esa aplicación en el área y grado donde se evidencia la copia.

5. Revisión de las estadísticas de los ítems

En el proceso de revisión de estadísticas de los ítems se tuvieron en cuenta estadísticas derivadas de la Teoría clásica de los test (TCT) y de la Teoría de respuesta al ítem (TRI). Las estadísticas empleadas en la revisión de los ítems se describen en el capítulo 5 del *Informe técnico de la aplicación de las pruebas SABER 5o. y 9o. 2009* (Cervantes & Lopera, 2011). A continuación se presentan las estadísticas de los ítems adicionales utilizadas en el análisis de los ítems en la aplicación de equiparación y en el de calibración de los ítems en las aplicaciones anteriores. Estas estadísticas apuntan a la evaluación de la conservación de las propiedades métricas de los ítems entre las diferentes versiones empleadas y se derivan de los modelos de la TRI.

El primer conjunto de medidas evaluadas se derivan del marco propuesto por Raju, van der Linden y Fleer (1995) para el análisis del funcionamiento diferencial de los ítems (DIF, por su sigla en inglés) y de las pruebas. Para cada ítem se calcularon las estadísticas de DIF no compensatorio definidas en este marco como

$$NCDIF_i = \int_{-\infty}^{\infty} \left[P_{iF}(\theta) - P_{iR}(\theta)\right]^2 f_F(\theta)\, d\,\theta \qquad (1)$$

y la medida de funcionamiento diferencial de la prueba, definida como

$$DTF = \int_{-\infty}^{\infty} \left[\sum_{i=1}^{k} \left(P_{iF}(\theta) - P_{iR}(\theta)\right)\right]^2 f_F(\theta)\, d\,\theta \qquad (2)$$

Donde:

$P_{iF}(\theta)$ representa la probabilidad de acertar al ítem dados los parámetros estimados en el grupo que se va a equiparar, que en términos de la literatura sobre funcionamiento diferencial de los ítems suele denominarse como grupo focal.

$P_{iR}(\theta)$ representa la probabilidad de acertar al ítem dados los parámetros estimados en el grupo original, que en términos de la literatura sobre funcionamiento diferencial de los ítems suele denominarse como grupo de referencia.

$P_F(\theta)$ representa la distribución de las habilidades en el grupo que se va a equiparar.

En particular, el modelo utilizado en las pruebas SABER 5o. y 9o. 2009, fue el modelo logístico de dos parámetros. Este modelo incorpora dos parámetros para describir cada ítem y se expresa como

$$P(U_i = 1 \mid \theta_j, a_i, b_i) = \frac{1}{1 + \exp(-1.702 a_i (\theta_j - b_i))} \qquad (3)$$

Donde:

U_i es la respuesta al *i-ésimo* ítem, con 1 que representa el caso en que la respuesta fue correcta y 0 si fue incorrecta.

θ_j es la habilidad del estudiante j en la prueba.

a_i representa la capacidad discriminativa del ítem.

b_i representa la dificultad del ítem.

El cálculo de las anteriores medidas asume que los parámetros estimados en ambos grupos, focal y referencia, se trasformaron a una escala común a partir de una trasformación afín, pues tal como se indica en la sección 5.1.1 del *Informe técnico SABER 5o. y 9o. 2009* (Cervantes & Lopera, 2011), "los modelos de respuesta al ítem […] no son identificables con respecto a la escala dado que los valores que puede tomar la habilidad, representada por θ_j, se pueden remplazar por $\theta^*_j = m\theta_j + s$, y dadas trasformaciones análogas de los parámetros de los ítems, la probabilidad calculada por la ecuación que representa el modelo es la misma (Hambleton, Swaminathan & Rogers, 1991)". Con el fin de encontrar la trasformación que lleva los parámetros estimados en el grupo focal a la misma escala de los parámetros en el grupo de referencia se utilizó el procedimiento propuesto por Haebara (1980) e implementado por Weeks (2010a, 2010b) en *R.* Este procedimiento estima las constantes denotadas por m y s minimizando $Q = Q_F + Q_R$, donde:

$$Q_F = \sum_{i=1}^{I} \left(\int_{-\infty}^{\infty} \left[P_{iF}(\theta) - P^*_{iR}(\theta) \right]^2 f_F(\theta)\, d\,\theta \right) \qquad (4)$$

$$Q_R = \sum_{i=1}^{I} \left(\int_{-\infty}^{\infty} \left[P_{iR}(\theta) - P^*_{iF}(\theta) \right]^2 f_R(\theta)\, d\,\theta \right) \qquad (5)$$

$P^{*}_{iG}(\theta)$ denota la probabilidad de responder correctamente el ítem dados los parámetros estimados para el grupo G trasformados de acuerdo con las constantes m y s; es decir, minimizando las diferencias entre las probabilidades de acertar cada ítem dados los parámetros de uno y otro grupo, tras la trasformación estimada, para ambos grupos.

5.1 Comparación del ajuste de los ítems

Para la calibración de los parámetros de los ítems que utilizara los valores de los parámetros obtenidos en la escala la correspondiente prueba SABER 5o. y 9o. 2009, se evaluó el ajuste de estos valores en las respuestas del grupo de evaluados analizado. Este ajuste se evaluó empleando las estadísticas INFIT y OUTFIT (Linacre, 2002; Linacre & Wright, 1994). Estas se definen como

$$INFIT_i = \left[\frac{\sum_{j=1}^{N} \left(U_{ij} - P_i(\theta_j)\right)^2}{\sum_{j=1}^{N} P_i(\theta_j)\,(1\text{-}P_i\ (\theta_j))} \right] \qquad (6)$$

$$OUTFIT_i = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{\left(U_{ij} - P_i(\theta_j)\right)^2}{P_i(\theta_j)\,(1 - P_i\ (\theta_j))} \right] \qquad (7)$$

Donde:

U_{ij} es la respuesta al *i-ésimo* ítem dada por el estudiante j, con 1 que representa el caso en que la respuesta fue correcta y 0 si fue incorrecta.

θ_j es la habilidad del estudiante j en la prueba.

$P_i(\theta_j)$ es la probabilidad que el estudiante j responda correctamente el *i-ésimo* ítem dados los parámetros estimados para el mismo.

Tanto para el INFIT como para el OUTFIT, valores cercanos a 1 indican un buen ajuste entre el modelo estimado y las proporciones empíricas de evaluados que aciertan el ítem, según su nivel de habilidad. Valores elevados indican un desajuste al modelo en el cual hay información de las respuestas que el modelo no recoge; valores pequeños de los índices de ajuste indican que las observaciones son demasiado predecibles a partir del modelo, por lo cual el modelo puede estar sobreajustado a la muestra particular (Linacre, 2002).

5.2 Advertencias

Además de los valores y gráficos de las estadísticas de los ítems calculados, se generaron algunas advertencias relacionadas con los valores encontrados para cada ítem. La inclusión de una advertencia en uno de los ítems a partir de las estadísticas relacionadas con la comparación de los ítems en dos aplicaciones diferentes buscó señalar posibles problemas con el ítem en cuanto a la no conservación de sus propiedades métricas. Las advertencias generadas son las siguientes:

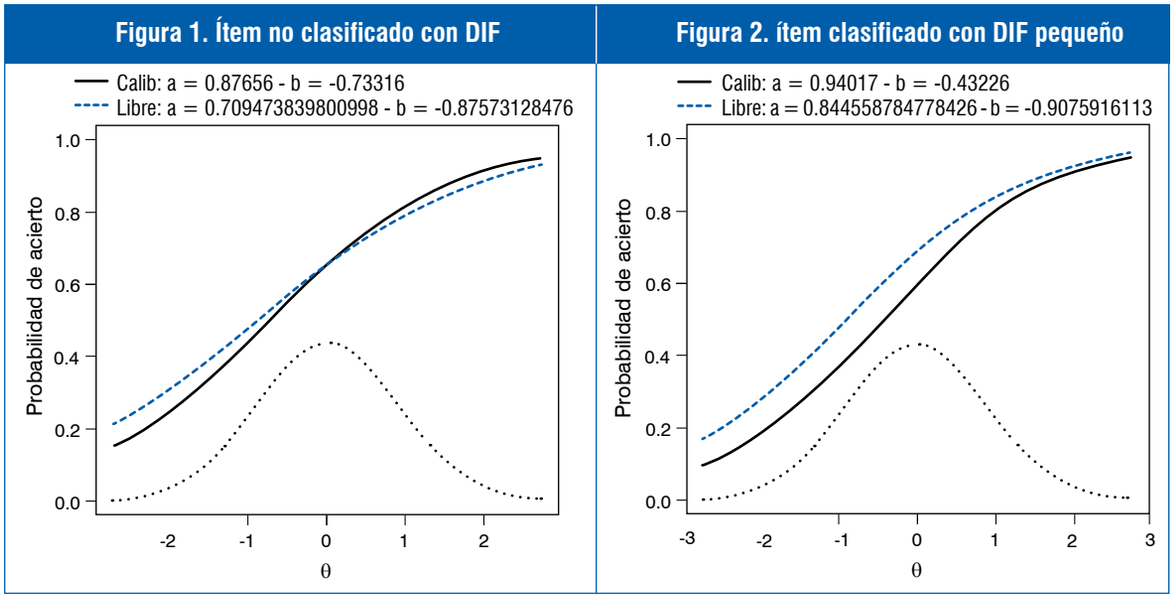
- El valor de NCDIF del ítem fue superior a 0,00467 (DIF pequeño).
- El valor de NCDIF del ítem fue superior a 0,02607 (DIF grande).
- El valor de DTF del ítem fue superior a 0,02607 veces el número de ítems.
- El valor de INFIT del ítem fue superior a 1,2.
- El valor de OUTFIT del ítem fue superior a 1,2.

Para las medidas de NCDIF, los puntos de corte se hallaron a partir de una simulación de Monte Carlo siguiendo las indicaciones de Meade, Lautenschlager y Johnson (2006). En esta simulación se emplearon tamaños de muestra en los grupos focal y referencia similares a los de la muestra de equiparación y a las aplicaciones de SABER 5o. y 9o. y se definieron diferencias en la dificultad de los ítems de 0,4 y 0,8 entre el grupo focal y de referencias como valores de DIF pequeño y grande, respectivamente. El valor del punto de corte para la medida de DIF se estableció de acuerdo con Flowers, Oshima y Raju (1999, citado por Morales et al., 2006).
Empleando estas advertencias, se consideró que un ítem señalado con DIF grande no tenía las mismas propiedades métricas en las dos aplicaciones consideradas y fue estimado como un ítem diferente en cada una de ellas. Esta acción también se tomó en los casos en que el DIF fue señalado como pequeño y el valor de OUTFIT fue mayor que 1,2.

5.3 Ejemplos de estadísticas

En el gráfico 1 se presentan ejemplos de las curvas características de dos ítems estimados en la aplicación de equiparación de mayo de 2010 y en la aplicación controlada de SABER 5o. y 9o. 2009; el ítem de la figura 1 no presentó DIF mientras que el ítem de la 2 presentó un DIF pequeño. En estas figuras, la línea punteada representa la distribución de habilidad de los evaluados en la aplicación de equiparación.

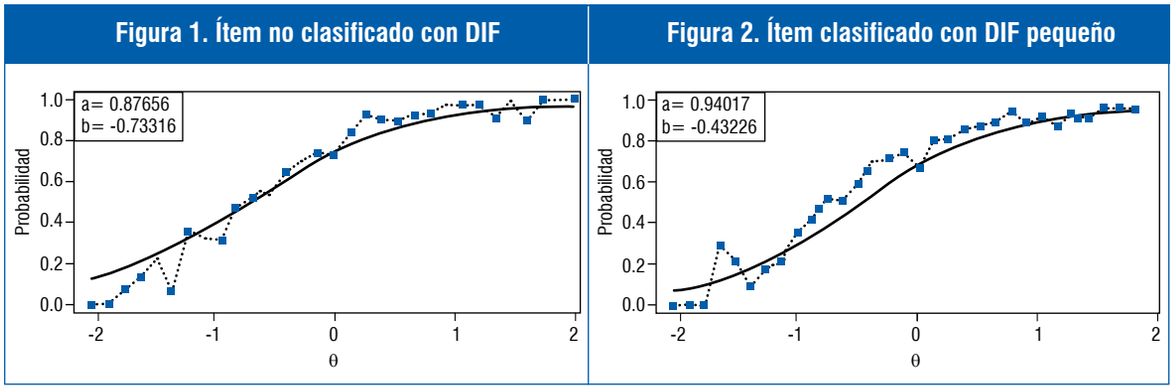
Gráfico 1. Curvas características de ítems en las aplicaciones de SABER 5o. y 9o. 2009 (línea continua) y de equiparación (línea segmentada)



Elaboración: Dirección de Evaluación, ICFES.

El gráfico 2, por su parte, presenta la curva característica de cada uno de estos ítems según la calibración en SABER 5o. y 9o. 2009, y las proporciones empíricas de acierto en 31 niveles de habilidad. La tabla 9 expone las medidas de NCDIF, INFIT y OUTFIT para los mismos ítems.

Gráfico 2. Curvas característica y empírica de los ítems



Elaboración: Dirección de Evaluación, ICFES.

Tabla 9. Valores de NCDIF, INFIT y OUTFIT de dos ítems en la aplicación de equiparación

Ítem	NCDIF	INFIT	OUTFIT
No clasificado con DIF	0,000729	0,903	0,868
No clasificado con DIF pequeño	0,006684	0,900	0,848

Elaboración: Dirección de Evaluación, ICFES.

5.4 Resultados de la revisión

A partir de la revisión de las estadísticas de los ítems presentadas en el capítulo 4 del *Informe técnico SABER 5o. y 9o. 2009* (Cervantes & Lopera, 2011), de los 426 ítems que se emplearon en la aplicación de equiparación de mayo de 2010, un total de cinco ítems fueron anulados de la calibración conjunta que se describe en la sección 7 de este informe.

De los 345 ítems de las aplicaciones anteriores, incluidos en el estudio de equiparación, 28 fueron anulados para la calificación de las pruebas de su respectiva aplicación. Por su parte, a partir de la revisión de las estadísticas de DIF y de ajuste presentadas, un total de 24 ítems fueron liberados de la calibración obtenida en la equiparación en la calificación de las aplicaciones anteriores de SABER 5o. y 9o. y ninguno fue liberado en los análisis en la muestra de equiparación de mayo de 2010.

6. Definición de las escalas en SABER 5o. y 9o. 2009

Las escalas de calificación en las cuales se reportan los puntajes de las pruebas SABER 5o. y 9o. se construyeron empleando el modelo logístico de dos parámetros de la Teoría de respuesta al ítem. Además, se empleó la metodología de valores plausibles para obtener los puntajes que se utilizaron en la realización de las estimaciones a niveles agregados de resultados relacionados con el desempeño en cada una de las pruebas. A continuación se presentan el modelo psicométrico utilizado para definir la escala y el procedimiento de valores plausibles para la generación de los puntajes.

6.1 Modelo de respuesta al ítem logístico de dos parámetros

Los modelos de respuesta al ítem (TRI) son modelos psicométricos que datan de la década de los años 1950 (verbigracia, Lord, 1950) y que se han desarrollado e implementado especialmente desde la década de los años 1970. Estos modelos representan la habilidad de un evaluado como una variable latente y modelan la probabilidad de que el evaluado proporcione determinada respuesta a partir del valor no observado de la habilidad del estudiante y de un conjunto de características o parámetros de los ítems (Lord & Novick, 1968). En particular, el modelo utilizado en las pruebas SABER 5o. y 9o. 2009, fue el modelo logístico de dos parámetros (ecuación 3). Los modelos de respuesta al ítem, como el empleado en SABER 5o. y 9o. 2009, no son identificables respecto a la escala. Esto implica que la escala en que se expresan los puntajes, por lo menos en la primera ocasión en que se aplica la prueba, se definen de forma arbitraria. Para cada una de las pruebas de SABER 5o. y 9o. 2009, se definió que la escala se expresaría de forma que el promedio de los evaluados fuera 300 y su desviación estándar, 80.

6.2 Trasformación a la escala de reporte

Como se indica en el capítulo 5 del *Informe técnico SABER 5o. y 9o. 2009* (Cervantes & Lopera, 2011), la escala para el reporte se definió con un promedio, de los evaluados en la aplicación censal, igual a 300 y desviación estándar 80. En la tabla 10 se presentan las constantes de las trasformaciones obtenidas para cada grado y área, en las que la trasformación de cada habilidad a la escala está dada por:

$$Puntaje\ en\ la\ escala\ de\ reporte = m\theta_j + s \qquad (8)$$

Las constantes (m y s) se usan para el reporte de los resultados en próximas ediciones de las pruebas SABER 5o. y 9o., así como para la realizada en el estudio de equiparación de resultados históricos.

Tabla 10. Constantes para la transformación a la escala de reporte

Grado	Área	s	m
Quinto	Lenguaje	290,365	77,6548247
	Matemáticas	281,916493	75,794025
Noveno	Lenguaje	292,848609	80,8045757
	Matemáticas	288,95652	80,6376679

Elaboración: Dirección de Evaluación, ICFES.

7. Procedimiento de equiparación de los resultados

Con el fin de observar la evolución de los resultados en las tres aplicaciones, se realizó un estudio de equiparación que permitió obtener resultados históricos comparables. En esta sección se presentan el procedimiento definido para efectuar la equiparación de los resultados, los aspectos particulares sobre este procedimiento en la muestra de equiparación y en las aplicaciones anteriores.

7.1 Procedimiento de calibración mediante ítems fijos

El procedimiento de calibración empleando valores fijos de los parámetros de los ítems se basa en la TRI para equiparar dos formas de prueba cuando se utiliza un diseño en el que existen ítems comunes entre las formas e ítems únicos a cada una de ellas. Este procedimiento consiste en utilizar como valores fijos las estimaciones de los ítems comunes obtenidas en la forma antigua o aquella que determina la escala fija, y en estimar los ítems no comunes de la forma nueva o aquella que quiere ser llevada a la escala fija (Kang & Petersen, 2009).

Además, permite tener estimaciones de ítems en formas nuevas en la misma escala definida de forma sencilla pero con dos inconvenientes: (1) si algunos ítems comunes empleados en la forma nueva no tienen un comportamiento similar en la nueva forma, las calibraciones de los ítems únicos a la forma nueva pueden resultar distorsionadas; (2) el procedimiento de estimación por máxima-verosimilitud marginal empleada por la mayoría de paquetes comerciales produce estimaciones sesgadas de habilidad, si los grupos que presentan cada forma no son equivalentes respecto a la distribución de la habilidad (Paek & Young, 2005; Kim, 2008; Kang & Petersen, 2009). En la utilización de este procedimiento se previeron los siguientes mecanismos para abordar estos inconvenientes:

a. Evaluación de la estabilidad de los ítems entre las diferentes aplicaciones de SABER 5o. y 9o. y la muestra de equiparación. Esta evaluación se llevó a cabo utilizando los procedimientos de detección del funcionamiento diferencial de los ítems señalados en la sección 5 de este documento.

b. Evaluación de la diferencia entre el promedio y la desviación estándar entre las diferentes aplicaciones de SABER 5o. y 9o. y la muestra de equiparación, y posible implementación del método de actualización de la distribución a priori por trasformación simple (STPU por su sigla en inglés) propuesto por Kim (2008).

Teniendo en cuenta que Paek y Young (2005) encontraron que cuando el procedimiento por defecto registra valores menores que 0,1 logits en la diferencia entre los promedios de los grupos que presentan cada forma, el sesgo incurrido por el mismo es suficientemente pequeño para no considerarse "pragmáticamente significativo", y que cuando la desviación estándar en el grupo que registra la forma nueva es igual o menor que la del grupo que presenta la forma de referencia el efecto del sesgo es menor que cuando tienen una desviación estándar mayor, no se contempló implementar el método STPU en los casos en que la diferencia entre los promedios fuese inferior a 0,1 en valor absoluto.

7.2 Equiparación en la muestra

Para el operativo controlado por el ICFES en el que se utilizaron cuadernillos de todas las aplicaciones censales de SABER 5o. y 9o., se eligieron: un cuadernillo aplicado en el 2002 (CN20021) en grado 5o, uno aplicado en el 2003 (CN20031) en grado 5o. y dos de grado 9o.; dos cuadernillos del 2005 (formas 2 y 3) en cada grado, y dos bloques de preguntas de matemáticas y lenguaje de 2009 para cada grado. En la selección de ítems se tuvo en cuenta que los cuadernillos elegidos en 2002-2003 y 2005-2006 tuvieran la mayor participación de población, que los ítems elegidos de 2009 no hubiesen sido liberados y que los ítems incluidos de las aplicaciones 2002-2003 y 2005-2006 no se hubiesen eliminado de las aplicaciones por aspectos psicométricos.

A partir de los cuadernillos seleccionados, se definieron ocho (8) cuadernillos por grado, cuatro (4) de lenguaje y cuatro (4) de matemáticas, donde cada uno de estos cuadernillos constó de un bloque de preguntas de 2009 y un bloque de preguntas de alguna de las aplicaciones anteriores (*véase* tabla 11)

Tabla 11. Estructura de los cuadernillos utilizados con los cuestionarios por grado, área y origen de las preguntas incluidas

Grado	Área	Forma	Aplicación de origen			
			2002	2003	2005	2009
Quinto	Lenguaje	A			●	●
		B			●	●
		C		●		●
		D		●		●
	Matemáticas	E	●		●	●
		F	●		●	●
		G	●			●
		H		●		●
Noveno	Lenguaje	A		●	●	●
		B		●	●	●
		C		●		●
		D		●		●
	Matemáticas	E		●	●	●
		F		●	●	●
		G		●		●
		H		●		●

Elaboración: Dirección de Evaluación, ICFES.

En la sección 5 de este informe se reportaron los resultados del análisis de ítems en términos de ítems anulados para la equiparación; para ningún ítem de 2009 se identificó funcionamiento diferencial con la muestra de equiparación. La evaluación de los promedios en las formas empleadas en la muestra de equiparación mostró que en la misma ningún promedio tuvo una diferencia superior a 0,025 logits respecto a la aplicación de SABER 5o. y 9o. 2009, por lo cual no se implementó el método STPU.

7.3 Equiparación en las aplicaciones anteriores

Una vez obtenida la calibración del conjunto de ítems seleccionados para el estudio de equiparación en una escala única, se procedió a recalificar las pruebas de todos los estudiantes que originalmente participaron en la evaluación y que presentaron alguno de los cuadernillos incluidos en la prueba de equiparación. En la sección 5 de este informe se reportaron los resultados del análisis de ítems respecto a los ítems anulados y de aquellos que no fueron considerados como equivalentes entre la aplicación original y la muestra de equiparación. En cuanto a la diferencia en logits, solamente uno de los cuadernillos previos mostró una diferencia que llegó a 0,2 logits; los demás no superaron el umbral definido, por lo cual no se llevó a cabo el procedimiento STPU.

Como en las aplicaciones previas no se incluyeron todos los aspectos tenidos en cuenta en el operativo de 2009 –en particular no se administraron cuestionarios sociodemográficos que permitieran el levantamiento de información de contexto de los estudiantes–, no fue posible seguir la metodología de calificación utilizada en 2009[5] . Lo anterior, unido al hecho de que solamente se utilizaron algunos cuadernillos de las aplicaciones previas y a los cambios administrativos ocurridos en las instituciones educativas derivados del proceso de integración institucional[6], implica que no fue posible obtener resultados para la totalidad de establecimientos educativos que participaron en las aplicaciones anteriores.

[5] Para la estimación de los puntajes promedio de la aplicación 2009 se utilizó, de forma independiente para cada grado y área, un modelo de la Teoría de respuesta al ítem (TRI) de dos parámetros y la metodología de valores plausibles.

[6] El proceso de integración institucional es un proceso de organización y reordenamiento de la oferta educativa de las entidades territoriales. Básicamente, consiste en tomar pequeñas instituciones educativas que ofrecen parcialmente algunos grados de la educación básica o media para aglutinarlas en torno a un solo núcleo educativo, conocido con el nombre de *institución educativa.*

8. Tipos de resultados producidos

La información histórica de SABER 5o. y 9o. produce dos tipos de resultados para cada área y grado —promedios y desviaciones estándar—, en los siguientes niveles de agregación[7]:

- Instituciones educativas.
- Departamentos, entidades territoriales, municipios certificados y no certificados, las cuales en este documento se denominaran como *entidades.*

Las instituciones educativas cuentan con resultados históricos para un área grado determinada, si cumple estos requisitos:

1. Es un establecimiento educativo que con base en la estructura institucional vigente de 2009 pudo ser identificada en los registros administrativos[8] utilizados para establecer el marco de la evaluación en 2002-2003 o 2005-2006.

2. Los estudiantes del establecimiento educativo presentaron en el área-grado las pruebas de 2002-2003 y 2005-2006 con alguno de los cuadernillos seleccionados en el estudio de equiparación.

3. El establecimiento no tiene indicios de copia masiva en el año en cuestión para el área-grado evaluados.

Adicionalmente, un *estudiante,* para efecto de los resultados históricos, se entiende como aquel que presentó alguno de los cuadernillos seleccionados para el estudio de equiparación y que pertenecían a una institución educativa (establecimiento o sede) cuya correspondencia con un establecimiento educativo de la organización administrativa vigente en 2009 pudo determinarse.

Por otro lado, en 2009 existe una estructura administrativa de las instituciones educativas en el país derivada del proceso de integración institucional (Ley 715 de 2001, capítulo III). Con base en esta estructura, los resultados que se presentan para los municipios, departamentos y entidades territoriales corresponden a los de todos los estudiantes que pertenecen a los establecimientos educativos para los que fue posible identificar resultados para 2002-2003 o 2005-2006, es decir, cumplen los criterios mencionados anteriormente.

[7] La publicación de los resultados históricos y la guía para la lectura e interpretación de estos se encuentran disponibles en la siguiente dirección electrónica: http://www.icfes.gov.co/saber59/.

[8] Se entiende por registros administrativos la información derivada del Ministerio de Educación Nacional y del DANE sobre las instituciones educativas del país.

En esta medida, para las entidades se presentan los resultados (promedios y desviación estándar) en tres grupos de información:

1. Establecimientos educativos que fue posible identificar en las tres aplicaciones simultáneamente.
2. Establecimientos educativos que fue posible identificar solo en 2002-2003 y 2009.
3. Establecimientos educativos que fue posible identificar solo en 2005-2006 y 2009.

Finalmente, teniendo en cuenta que el estudio de equiparación cuenta con varias limitaciones, relacionadas principalmente con la selección de un número reducido de las pruebas originalmente aplicadas y los cambios administrativos derivados del proceso de integración institucional, que no permitieron generar resultados para la totalidad de establecimientos educativos que participaron en las aplicaciones anteriores, los resultados históricos no pueden generalizarse para toda la población (instituciones educativas y estudiantes) que participaron originalmente en las pruebas censales de 2002-2003 y 2005-2006, es decir, los resultados históricos solo dan cuenta del conjunto de evaluados que pudieron equipararse.

Adicionalmente, las características metodológicas de las evaluaciones anteriores, en especial la variación de la estructura de prueba en cada aplicación, también generan limitaciones para la comparación de los resultados obtenidos mediante la equiparación.

Bibliografía

- **Cervantes V.H. & Lopera, C. (Eds.).** (2011). Informe técnico SABER 5o. y 9o. Instituto Colombiano para la Evaluación de la Educación, ICFES. Bogotá: ICFES.

- **Congreso de la República** (2001, 21 de diciembre). Ley 715 de 2001 "por la cual se dictan normas orgánicas en materia de recursos y competencias de conformidad con los artículos 151, 288, 356 y 357 (Acto Legislativo 01 de 2001) de la Constitución política y se dictan otras disposiciones para organizar la prestación de los servicios de educación y salud, entre otros". *Diario Oficial No 44.654*

- **Flowers, C.P., Oshima T.C. & Raju, N.S.** (1999). A description and demostration of the polytomous-DFIT framework, *Applied Psychological Measurement,* 23, 309-326

- **Haebara, T.** (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research,* 22(3), 144-149.

- **Hambleton, R.K., Swaminathan, H. & Rogers, J.H.** (1991). *Fundamentals of item response theory.* Sage: Newbury Park, CA, EEUU.

- **Kang, T. & Petersen, N.** (2009). Linking item parameters to a base scale. Reporte técnico RR2009-2 del American College Testing. Versión digital disponible en http://www.act.org/research/researchers/reports/pdf/ACT_RR2009-2.pdf

- **Kim, Seonghoon.** (2008). IRT fixed parameter calibration and other approaches to maintaining item parameters on a common scale. Trabajo presentado en Measured Progress. Versión digital disponible en http://www.measuredprogress.com/WorkArea/DownloadAsset.aspx?id=650

- **Linacre, J.M.** (2002). What do INFIT and OUTFIT, mean-square and standardized mean? *Rasch Measurement Transactions,* 16(2), 878.

- **Linacre J.M. & Wright B.D.** (1994) Chi-Square Fit Statistics. *Rasch Measurement Transactions,* 8(2), 350.

- **Martínez, W. & Cervantes, V.** (2010). Metodología utilizada para la detección de los casos de indicios de copia en SABER 5o. y 9o. 2009. Instituto Colombiano para la Evaluación de la Educación, ICFES, Bogotá. Versión digital disponible en: http://www.icfesSABER.edu.co/uploads/documentos/metodologia_deteccion_copia.pdf

- **Meade, A. W., Lautenschlager, G. J. & Jonson, E. C.** (2006). Alternate cutoff values and DFIT tests of emasurement invariante. Trabajo presentado en la 21 Annual Conference of the Society for Industrial and Organizacional Psychology, Dalla, Texas, Estados Unidos.

- **Morales, L.S, Flowers, C.P., Gutierrez, P., Kleinman, M.S. & Teresa, J.A.** (2006). Item and scale differential functioning of the Mini-Mental State Exam assessed using the Diferencial Item and Test Functioning (DFIT) framework, *Medical Care,* 44(11, sup 3), 143-151

- **Paek, I. & Young, M.J.** (2005). Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data. *Applied Measurement in Education,* 18(2), 199-215

- **Raju, N. S., van der Linden, W. J. & Fleer, P. F.** (1995). IRT-based internal measures of differential functioning of ítems and tests. *Applied Psychological Measurement,* 19(4), 353-369

- **Weeks, J. P.** (2010a). DESI. plink. [Manual y software de cómputo Versión 1.2-3]. http://cran.r-project.org/web/packages/plink/.

- **Weeks, J. P.** (2010b). plink: An R Package for Linking Mixed-Format Tests Using IRT-Based Methods. *Journal of Statistical Software,* 35(12), 1-33.

Calle 17 No. 3-40 • Teléfono:(57-1)338 7338 • Fax:(57-1)283 6778 • Bogotá - Colombia
www.icfes.gov.co