UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Abril 2016, Vol. 1, No. 0, pp. 0–11

Modelo de recalificación para la prueba SABER 11

Requalification model for SABER 11 test

Dora Suarez[a]
dsuarez@contratista.icfes.gov.co

Cristian F. Tellez[b]
ctellez@icfes.gov.co

Resumen

Actualmente, las pruebas estandarizadas son una herramienta fundamental a la hora de evaluar la calidad de la educación. Los cambios poblacionales y el la inclusión de nuevas forma de evaluación hacen necesario el uso de metodologías que permitan comparar los resultados de las pruebas en las diferentes aplicaciones de la misma. En el presente artículo se exponen diferentes metodologías para la equiparación de puntuaciones a través de transformaciones aplicadas al caso particular de la prueba SABER 11, aplicada por el Instituto Colombiano para la Evaluación de la Calidad de la Educación (ICFES), dado que para esta prueba se presentó un cambio estructural a partir de la segunda aplicación de 2014. Para lograr la equiparación se utilizan modelos lineales generalizados y modelos de teoría clásica de los test, para hacer las comparaciones encontrando menor error cuando se utilizan modelos de regresión beta. Las metodologías aquí expuestas pueden ser aplicadas para escenarios en los que se necesite hacer equiparación de puntuaciones para pruebas estandarizadas.

***Palabras clave*:** Modelos lineales generalizados, Equiparación de puntajes , Modelos de regresión Beta, Pruebas estandarizadas .

Abstract

Nowadays the standardized tests are an essential tool when assessing the quality of education, the population changes and the inclusion of new evaluation forms would require the use of methodologies that allow the comparison between the results of the tests in different applications. This article outlines the different methodologies for equalization Rating through transformations applied to the particular case of the test SABER 11, applied by the Instituto Colombiano para la Evaluación de la

[a]Msc. Universidade Federal de pernambuco, Estadística-Subdirección de estadísticas, ICFES.
[b]Estudiante de doctorado en estadística , Universidad Nacional de Colombia , Subdirector de estadísticas, ICFES.

Calidad de la Educación (ICFES), due to the structural change in the test presented from 2014. To achieve the equalization the Generalized Linear Models and the Test Classic theory models are used, to make comparisons with less error when Beta Regression models are used. The methodologies presented can be applied to different scenarios in which equalization Rating for standardized test is required.

***Keywords*:** Generalized linear models, Rating equalization , Beta regression models, Standardized test .

1. Introducción

Los puntajes de las pruebas de estado de la educación media son usadas frecuentemente como criterio en la toma de decisiones individuales e institucionales (Chica Gómez et al. 2011).

Los cambios presentes en los modelos educativos a través del tiempo impulsan desarrollos en las pruebas estandarizadas de forma que tengan mayor capacidad de captar las diferencias cognitivas de los evaluados (Martínez Rizo 2001).

El ICFES aprobó una nueva estructura y organización del Examen de Estado para la Educación media SABER 11 a partir de la segunda aplicación del año 2014.

Este cambio estructural permite que la prueba de estado esté alineada con los otros exámenes del Sistema Nacional de Evaluación Estandarizada (SNEE): SABER 3, SABER 5, SABER 9 y SABER PRO (Bernal Velásquez 2013).

Los exámenes de estado están sujetos a ser aplicados bajo los principios de independencia, igualdad, comparabilidad, periodicidad, reserva individual, pertinencia y relevancia. De esta forma, se hace necesario asegurar la comparabilidad entre las aplicaciones tanto las hechas antes de la segunda aplicación y aplicaciones siguientes.

Como consecuencia, un cambio estructural en la prueba estandarizada, necesita un proceso de recalificación de los puntajes en las aplicaciones anteriores de la prueba con el fin de obtener uno equivalente en la escala de la prueba actual.

Este articulo pretende comparar varias alternativas disponibles para la recalificación de los puntajes de forma que al equiparar sea posible llevar los puntajes globales de la pruebas estandarizadas aplicadas después de un cambio estructural (Bernal Velásquez 2013).

2. Generalidades de la prueba

Los cambios propuestos para el examen SABER 11 para la segunda aplicación de 2014 consisten en modificar la estructura del examen con el objetivo de que los resultados obtenidos sean comparables, en algún sentido, con los de los otros exámenes del sistema nacional de evaluación estandarizada SNEE (Julián P. Ma-

riño von Hildebrand 2014). El cambio consistió en pasar de 9 pruebas, que incluían una profundización, a 5 pruebas. En el Cuadro 1 se muestra la estructura y forma de calcular los puntajes globales de la prueba SABER 11 antes y después de la re-estructuración de la prueba. Esta ponderación es realizada de forma que lectura critica, matemáticas, ciencias naturales y sociales y ciudadanas tengan el mismo peso e ingles tenga un peso menor dadas las diferencias entre estudiantes de calendario A y B, colegios oficiales y no oficiales (Julián P. Mariño von Hildebrand 2014). El indice global anterior era calculado bajo la misma esencia, agrupando cada una de las áreas evaluadas.

Prueba	Caracterítica
Antes de 2014-II	Matemáticas (M) Lenguaje (L) Biología (B) Física (F) Química (Q) Ciencias Sociales (S) Filosofía (Fi) Inglés (I)
Indice global anterior	$IG_1 = \frac{P_B + P_Q + P_F + 2P_S + P_{Fi} + 3P_L + 3P_M + P_I}{13}$
Después de 2014-II	lectura Critica (LC) Matemática (MA) Ciencias Naturales (CN) Sociales y ciudadanas (SC) Ingles (IN)
Indice global actual	$IG_1 = \frac{3P_{LC} + 3P_{MA} + 3P_{CN} + 3P_{SC} + P_{IN}}{13}$

Tabla 1: Estructura del examen antes y después de 2014-II y forma de calculo del puntaje global.

Cada uno de los puntajes esta en una escala de medición de 0 a 100, por tanto, el indice global resultante corresponde al promedio ponderado de los puntajes en las diferentes pruebas, es también una medida entre 0 y 100. El índice resultante es multiplicado por 5, que corresponde al puntaje global que se publica en los reportes individuales.

Todos los exámenes de las aplicaciones desde 2012-I hasta 2014-I, fueron recalificados utilizando teoría clásica del test y teoría de respuesta al ítem (TRI). La recalificación consistió en evaluar a lo largo de las aplicaciones de la prueba, que ítems podían utilizarse como anclas para llevar de la escala anterior a esta nueva escala (Kolen & Brennan 2004). El resultado de Estas recalificaciones son el insumo principal para poder realizar los modelos de recalificación de los puntajes globales a partir de los puntajes para cada una de las pruebas.

3. Metodología

En 2014-I, a través de metodologías basadas en la teoría clásica de los test (TCT), se realizó la recalificación de los exámenes hechos desde 2012-I a 2014-I, esta recalificación ofrece, por cada una de las pruebas, el puntaje equivalente a la nueva estructura del examen. Una regla general de equiparación de las puntuaciones globales puede ser obtenida mediante el uso de los puntajes los puntajes individuales de las pruebas aplicadas desde 2012-I a 2014-II junto con su respectiva recalificación. Para ello, se cuenta con 1'426.641 evaluados con puntuaciones en las diferentes pruebas y sus respectivas recalificaciones. La distribución del número de estudiantes por aplicación y el promedio del puntaje global es presentado en el Cuadro 2.

Periodo	2012-I	2012-II	2013-I	2013-II	2014-I
Número de evaluados	97272	577100	87378	575224	89667
Puntaje global promedio	260.73	249.92	260.39	250.14	254.81

Tabla 2: Distribución de las pruebas por año y tipo del evaluado

Para equiparar los puntajes globales, se busca explicar los puntajes de la recalificación por medio de los puntajes reales obtenidos en las diferentes pruebas. Cuando se observa la correlación entre las diferentes pruebas, encontramos valores altos, tanto entre las pruebas re-calificadas como en los valores de los puntajes originales de la prueba (Matrices (1) y (2)). Al realizar la prueba de esfericidad de Barlett (Grossman et al. 1991) se concluye que existe una correlación estadísticamente significativa entre las pruebas a un nivel de significancia del 5 %.

Esto indica entonces la presencia de una variable latente que no se ha observado directamente a través de los puntajes brutos, por lo cual se procede a reducir la dimensionalidad de la prueba por medio de un análisis de componentes principales con rotación ortogonal, normado (trabajando con la matriz de correlaciones directamente) (Peña 2002). Para buscar esta reducción de la dimensionalidad se realiza una transformación lineal de las variables de forma que se conserva la mayor cantidad de variabilidad en los datos con la menor perdida de información (Peña 2002).

	P_{MA}	P_{LC}	P_{SC}	P_{CN}	P_{IN}
P_{MA}	1.000	0.594	0.582	0.730	0.565
P_{LC}	0.594	1.000	0.758	0.676	0.586
P_{SC}	0.582	0.758	1.000	0.676	0.567
P_{CN}	0.730	0.676	0.676	1.000	0.611
P_{IN}	0.565	0.586	0.567	0.611	1.000

(1)

	P_{Biol}	P_{Soc}	P_{Filo}	$P_{Fís}$	P_{Ing}	P_{Len}	P_{Mate}	$P_{Quí}$
P_{Biol}	1.00	0.57	0.48	0.50	0.51	0.54	0.53	0.60
P_{Soc}	0.57	1.00	0.53	0.47	0.53	0.58	0.50	0.57
P_{Filo}	0.48	0.53	1.00	0.43	0.48	0.50	0.43	0.50
$P_{Fís}$	0.50	0.47	0.43	1.00	0.46	0.45	0.49	0.58
P_{Ing}	0.51	g0.53	0.48	0.46	1.00	0.53	0.52	0.55
P_{Len}	0.54	0.58	0.50	0.45	0.53	1.00	0.49	0.54
P_{Mate}	0.53	0.50	0.43	0.49	0.52	0.49	1.00	0.57
$P_{Quí}$	0.60	0.57	0.50	0.58	0.55	0.54	0.57	1.00

(2)

Para realizar el análisis de componentes principales (ACP), se incluyen como variables activas todas las 8 (ocho) pruebas. En el Cuadro 3 se pueden ver los valores propios asociados a cada uno de los ejes factoriales y la varianza que retiene cada una de las componentes del ACP. Notemos que únicamente el primer eje factorial ya retiene el 57.69 % de toda la varianza. Los vectores propios asociados a cada valor propio se muestran en el Cuadro 4.

Eje	Valor Propio	Porcentaje de Varianza
1	4.62	57.69 %
2	0.64	7.96 %
3	0.55	6.92 %
4	0.50	6.28 %
5	0.47	5.92 %
6	0.44	5.51 %
7	0.40	5.03 %
8	0.37	4.68 %

Tabla 3: Valores propios y porcentaje de varianza acogida por cada eje factorial

	Eje 1	Eje 2	Eje 3	Eje 4	Eje 5	Eje 6	Eje 7	Eje 8
Biología	-0.36	-0.08	-0.03	0.46	0.33	0.54	0.40	-0.29
Sociales	-0.37	0.28	0.01	0.39	-0.04	0.01	-0.78	-0.11
Filosofía	-0.33	0.53	0.52	-0.46	0.33	-0.03	0.11	-0.04
Física	-0.33	-0.58	0.53	-0.06	-0.39	-0.14	-0.01	-0.31
Inglés	-0.35	0.09	-0.45	-0.51	-0.46	0.44	-0.03	-0.07
Lenguaje	-0.36	0.34	-0.18	0.30	-0.38	-0.53	0.45	0.06
matemáticas	-0.35	-0.34	-0.44	-0.25	0.52	-0.44	-0.09	-0.18
Química	-0.38	-0.25	0.09	0.05	0.08	0.12	-0.02	0.87

Tabla 4: Vectores propios asociados a cada eje del análisis de compoenentes principales

Dada la gran cantidad de variabilidad que recoge el primer eje factorial respecto a los demás ejes, se ha decidido crear un índice basado en este eje que será usado

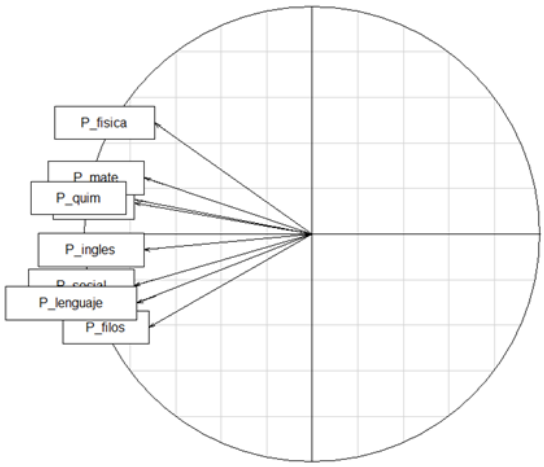

Figura 1: Plano factorial para los primeros 2 ejes factoriales, el primer eje contiene el 57 % de variabilidad

como medida de resumen, para cada individuo, del desempeño en la prueba basado en las diferentes subpruebas, que llamaremos en adelante índice global del ACP. En adelante llamaremos a este el incide global del ACP, denotado como IG_{ACP} y que calcularemos como:

$$\widehat{IG}_{\text{ACP}} = \lambda_1 \sum_{j=1}^{8} x_j \cdot a_{1j}, \tag{3}$$

donde λ_1 es el primer valor propio, x_j es el puntaje estandarizado de la prueba j en el examen y a_{1j} es la componente principal asociada al primer valor propio. Este indice creado es utilizado como variable explicativa para los modelos de regresión que serán propuestos a continuación. El primer plano factorial puede ser visto en la Figura 1 La variable dependiente será el indice global basado en la recalificación de la prueba, por conveniencia el indice global será dividido entre 100 para obtener una medida que este entre 0 y 1 y se denotará por IG_{Recal}.

3.1. Procedimientos de equiparación

Modelo lineal

En el modelo de regresión lineal se relaciona el indice global calculado a través de la recalificación y el indice resultante del análisis de componentes principales. El modelo puede ser escrito como:

$$\mathrm{IG}_{Recal} = \beta_0 + \beta_1 \mathrm{IG}_{ACP} + \epsilon \tag{4}$$

Al ajustar el modelo lineal sobre los indices de la prueba se tiene:

$$\widehat{\mathrm{IG}}_{Recal} = -3.0610 + 0.091 \cdot \mathrm{IG}_{ACP} \tag{5}$$

$$\tag{6}$$

Modelo Gamma

El modelo Gamma ayuda a modelar el valor esperado de variables que son mayores que cero, considerando que el indice global calculado a través de la recalificación (medido de 0 a 1) es un valor siempre mayor que cero, este modelo parece ser adecuado para establecer la equivalencia. El modelo quedará definido entonces como:

$$\mu_i = E(IGA_{\mathrm{recal}}) = g^{-1}(\beta_0 + \beta_1 \mathrm{IG}_{\mathrm{ACP}}) + \epsilon, \tag{7}$$

en que $g(\mu_i)$ es la función de enlace logística dada por:

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right). \tag{8}$$

En este caso se asume que μ_i tiene distribución gamma, con $\mu_i > 0$.

Modelo Beta

Para un modelo Beta, se asume que la media de la variable aleatoria esta medida entre 0 y 1. Teniendo en cuenta que IG_{Recal} es una variable aleatoria entre 0 y 1 se utiliza una regresión Beta para modelar esta variable. En este caso, se asume que para el i-ésimo individuo, $\mathrm{IG}_{Recal,i} \sim \mathrm{Beta}(\mu_{\mathrm{i}}, \phi)$, por lo tanto, el modelo de regresión estaría dado por:

$$\mu_i = g^{-1}(\beta_0 + \beta_1 \mathrm{IG}_{ACP}) + \epsilon; \quad i = 1, 2 \ldots n. \tag{9}$$

Nuevamente se utiliza la función de enlace logística como se presenta en (8).

Teoría clásica de los test TCT - Modelo equipercentil

Desde la teoría clásica de los tests, se pueden realizar dos tipos de equivalencias entre las puntuaciones ya sea lineal como no lineal. En el primer caso, se puede utilizar la función de identidad, de medias o funciones lineales. En el segundo caso, el circulo de angulo igualado o el método equipercentil (Kolen & Brennan 2004).

En este documento se hará uso del método equipercentil, ya que el puntaje global es una variable continua. En este método se define una relación no lineal entre las escalas de puntuación, estableciendo una igualdad entre las funciones de distribución acumuladas para las poblaciones que quieran ser equiparadas. Si se desean equiparar los puntajes IG_1 e IG_2, cuyas funciones de densidad son $F(IG_1)$ y $G(IG_2)$ respectivamente, entonces la relación de equiparación para el modelo equipercentil estará dada por:

$$IG_2 = G^{-1}(F(IG_1)). \tag{10}$$

Para establecer la relación entre las dos puntuaciones son utilizados métodos de pre-suavizado y post-suavizado con el objetivo de mejorar el ajuste. Este suavizamiento se realiza con el método kernel (en general se usa kernel gaussiano) de equiparación (Gasser & Müller 1979), luego se evalúa el error y la precisión de la estimación (Steinberg & Moses 2011).

Validación de los modelos
En la validación de los modelos se verifica que se cumplan los siguientes supuestos:

- Buena especificacion del modelo (Prueba de reset de Ramsey);
- Valor esperado de los residuales igual a cero;
- Homocedasticidad (Prueba de Breusch-Pagan);
- No colinealidad (Coeficiente kappa, como hay una sola variable no aplica),
- Normalidad en los errores.

Los modelos que son presentados aquí son validados a través del análisis de residuales. En cada uno de los modelos los residuales serán calculados como:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{var(y_i)}}}. \tag{11}$$

4. Resultados

Para cada uno de los modelos ajustados son calculadas medidas de ajuste del mismo sobre los datos, a saber la suma de cuadrados de los errores (SCR), el

Medida	Lineal	Gamma	Beta
SCR	9431.46	9249.24	9239.13
Akaike	-8851.85	-8841.99	-8811.66
BIC	-8835.26	-8825.40	-8795.06
Deviance	0.934	0.928	0.931

Tabla 5: Criterios de medida

Modelo	Cobertura	Amplitud
Beta	0.95	12536904
Gamma	0.95	12848854
Lineal	1.00	48609023
Equipercentil	1.00	46317587

Tabla 6: Criterios de medida

criterio de información de Akaike (AIC), el criterio de información bayesiano (BIC), el R^2 por deviance (deviance), la cobertura y la amplitud (Salibian-Barrera & Zamar 2002) de los modelos. En cuanto a la suma de cuadrados de los errores, la cobertura y la amplitud, se observa que el modelo Beta es superior a los demás modelos. El modelo gamma es superior en cuanto al criterio de información de Akaike y bayesiano y el modelo de regresión lineal con respecto al R^2 por deviance. Estos resultados pueden observarse en los Cuadros 5 y 6.

Para determinar el número de componentes a retener en la creación del $\widehat{IG}_{ACP}$, se hizo un estudio de simulación, en el que se fue variando el número de componentes retenidas para una muestra de tamaño 1864 (determinada bajo un muestreo aleatorio simple **M.A.S**) que incluía a todos los evaluados a los cuales se tuvo acceso, donde se obtuvieron los resultados que son presentados en la Figura 2 (los resultados que se muestran es el promedio de 100 simulaciones):

Se evidencia que retener una sola componente en el cálculo del $\widehat{IG}_{ACP}$ es adecuado, ya que tiene menor **SCE, AIC, BIC** y un mejor ajuste en terminos de R^2.

De acuerdo a los criterios presentados en el resumen metodológico, se decide realizar la equiparación del puntaje global de la prueba a través del modelo *beta*, puesto que presenta menor suma de cuadrados del error **(SCE)**, una cobertura del 95 % aproximadamente y una amplitud del intervalo de predicción menor que la encontrada para los demás modelos. En este modelo la variable modelada siempre estará el intervalo [0,1] lo cual asegura que ninguna de las predicciones estará por fuera lo establecido en la creación del índice actual.

De acuerdo a los parámetros encontrados para el modelo Beta y los valores para la creación del indice global a través del ACP, el indice global actual se calcularía con base a los puntajes obtenidos en cada una de las pruebas individuales de la

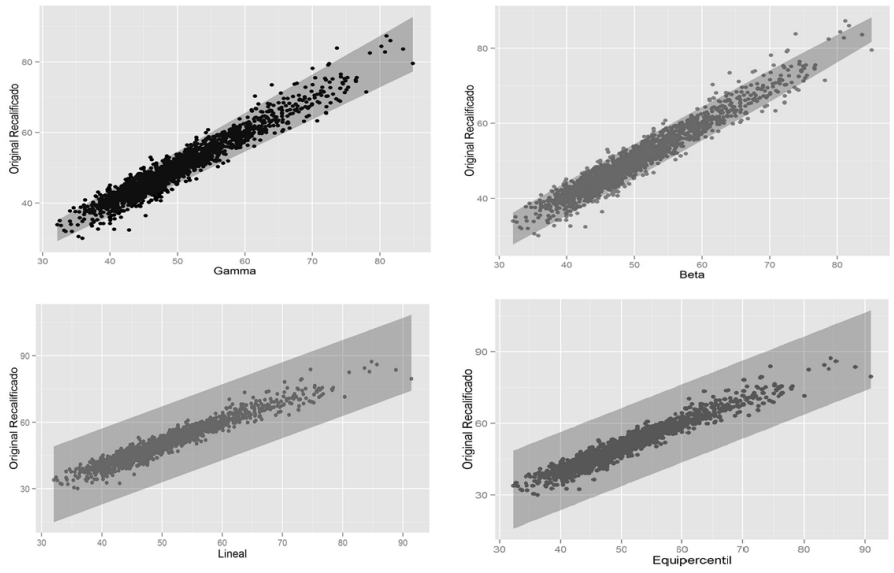

Figura 2: Amplitud y cobertura para los modelos propuestos

siguiente forma:

$$\widehat{IG}_j = \beta_1 \lambda_1 \sum_{i=1}^{8} \frac{a_i x_{ij}}{S_i} - \beta_1 \lambda_1 \sum_{i=1}^{8} \frac{a_i \bar{x}_i}{S_i} + \beta_0 \tag{12}$$

donde:

- x_{ij} es el puntaje obtenido en la prueba i para la *j-ésima* persona.
- $\bar{x}_i$ es el promedio de las puntuaciones para la *i-ésima* prueba.
- S_i es la desviación estándar estimada para la *i-ésima* prueba.
- a_i son coeficientes provenientes del análisis de componentes principales, específicamente obtenidos del vector propio asociado al primer valor propio.
- λ_i es el primer valor propio del análisis de componentes principales.
- β_k son los coeficientes de la regresión beta con $k = 0, 1$.

Al realizar las estimaciones de los parámetros presentados en (12) se obtiene la siguiente relación entre el indice global actual y los puntajes de las pruebas:

$$\widehat{IG}_j = \mathbf{a} \cdot \mathbf{M}^\top \tag{13}$$

con

$$\mathbf{a} = [-2.466826, 0.0008034867, 0.0007437763, 0.0006196794, 0.0005891047, \tag{14}$$
$$0.0005464668, 0.0007962810, 0.0005441301, 0.0008231145] \tag{15}$$

y

$$\mathbf{M} = [1, P_{\text{Biología}}, P_{Sociales}, P_{\text{Filosofía}}, P_{\text{Física}}, P_{\text{Inglés}}, P_{Lenguaje}, P_{\text{Matemáticas}}, P_{\text{Química}}] \tag{16}$$

5. Conclusiones

Este ejercicio, permitió comparar y llevar a la misma métrica los resultados de los evaluados del examen de estado SABER 11, obtenidos como un puntaje global, para distintas aplicaciones en las que se contaba con un cambio estructural de la prueba.

Los resultados muestran que con 4 métodos diferentes es posible predecir los puntajes globales de los evaluados utilizando solo los puntajes obtenidos en una aplicación anterior a 2014-2. Sin embargo, al evaluar medidas de ajuste para de las puntuaciones obtenidas la equiparación con el uso de algún modelo de regresión lineal generalizado es mas robusto que con el uso de equiparaciones basadas en los métodos convencionales de TRI y TCT.

En particular, la equiparación utilizando un modelo de regresión beta mostró un mejor ajuste en cuanto a la cobertura y la amplitud de los intervalos a ser medidos ya que los puntajes a equiparar se encuentran en una escala acotada fácilmente llevada a una escala de 0 a 1. Puntuaciones no acotadas como las de el examen SABER PRO pueden ser equiparadas mediante el uso de un modelo de regresión gamma.

Referencias

Bernal Velásquez, R. (2013), 'Sistema nacional de evaluación estandarizada de la educación', *Alineación del examen Saber* **11**.

Chica Gómez, S. M., Galvis Gutiérrez, D. M., Ramírez Hassan, A. et al. (2011), 'Determinantes del rendimiento académico en colombia: pruebas icfes saber 11º, 2009'.

Gasser, T. & Müller, H.-G. (1979), *Kernel estimation of regression functions*, Springer.

Grossman, G. D., Nickerson, D. M. & Freeman, M. C. (1991), 'Principal component analyses of assemblage structure data: utility of tests based on eigenvalues', *Ecology* pp. 341–347.

Julián P. Mariño von Hildebrand, e. (2014), *Sistema Nacional de Evaluación Estandarizada de la Educación. Alineación del examen SABER 11; Lineamientos generales 2014-2.*, ICFES.

Kolen, M. J. & Brennan, R. L. (2004), *Test equating, scaling, and linking*, Springer.

Martínez Rizo, F. (2001), 'Evaluación educativa y pruebas estandarizadas. elementos para enriquecer el debate', *Revista de la educación superior* **30**(120), 71–85.

Peña, D. (2002), *Análisis de datos multivariantes*, Vol. 24, McGraw-Hill Madrid.

Salibian-Barrera, M. & Zamar, R. H. (2002), 'Bootstrapping robust estimates of regression', *Annals of Statistics* pp. 556–582.

Steinberg, J. & Moses, T. (2011), 'Smoothing scaled score distributions from a standardized test using proc genmod', *SESUG 2011: The 19th annual proceedings of the SouthEast SAS Users Group, Arlington, VA* .