UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Statistics on the evaluation of education
June 2016, Vol. 9, No. 1, pp. 39–49

Requalification model for Saber 11 test[1]

Modelo de recalificación para la prueba Saber 11

William Acero R.[a]
wacero@contratista.icfes.gov.co

Jesús Fernando Sánchez[b]
jsanchez@contratista.icfes.gov.co

Dora Suárez[c]
dsuarez@contratista.icfes.gov.co

Cristian F. Téllez[d]
ctellez@icfes.gov.co

Abstract

Standardized tests are nowadays an essential tool when assessing the quality of education, population changes and the inclusion of new evaluation forms would require the use of methodologies that allow comparisons between the results of the tests in its different applications. This article outlines different methodologies for the equating scores through transformations applied to the particular case of the Saber 11 test, that was implemented by the Instituto Colombiano para la Evaluación de la Calidad de la Educación (ICFES), due to the structural change made in the test since the second application of 2014. Generalized linear models and classic test theory models are used to achieve equalization; when beta regression models are used, less error is found. Presented methodologies can be applied to different scenarios in which scores equalization for a standardized test is required.

***Keywords*:** generalized linear models, equating scores, beta regression models, standardized test.

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

[1] Acero, W., Sánchez, J., Suárez, Dora., Téllez, C. (2016) Requalification model for Saber 11 test. *Comunicaciones en Estadística*, **9**(1), 39-49.

[a] Statistician, Subdivision de Stadistics, Icfes, Colombia.

[b] Statistician, Subdivision de Stadistics, Icfes, Colombia.

[c] Msc. Federal University of Pernambuco, Statistician, Subdivision de Stadistics, Icfes, Colombia.

[d] PhD Student in Statistics , National University of Colombia. Deputy Director of Statistics, Icfes, Colombia.

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.

1 Introduction

Scores of State tests score in middle education are frequently used as criteria for the making decisions of individual and institutional choices (Chica et al. 2011).

Changes in educational models over time drive developments in standardized tests so that Educational models changes over time drive developments in standardized tests so that they have a greater ability to capture cognitive differences of evaluated people(Martínez 2001).

ICFES approved a new structure and organization of state test for middle education Saber 11 since its second application in 2014. This structural change allows the test to be aligned with other exams of the National System of Standardized Evaluation (SNEE): Saber 3, Saber 5, Saber 9 and Saber Pro (Bernal 2013).

State tests are subject to be applied under the principles of independence, equality, comparability, periodicity, individual reserve, relevance and significance. In this way, it is necessary to ensure comparability between both applications, those made before the second application and the following ones.

Consequently, a structural change in the standardized test requires a process of scores re-qualifying in previous applications of the test to obtain an equivalent on the scale of the current test.

This article aims to compare several available alternatives for scores re-qualifying so that it's possible by equating to bring the global scores of standardized tests applied after a structural change. (Bernal 2013).

2 Test Generalities

Proposed changes in Saber 11 for the second application in 2014 attempt to modifying the structure of the test in order that the results obtained are comparable, in

some sense, with those of other tests of SNEE(Mariño 2014). The change consisted in reducing the number of tests from 9, which included deepening, to 5. Table 1 shows the structure and the way of calculating the global scores of Saber 11 test before and after test's restructuring. This weighting is performed so that critical reading, mathematics, natural and social sciences have the same weight and English have a lower weight, given the differences between students of schedule A and B, of official and non-official schools (Mariño 2014). The previous global index was calculated under the same essence, grouping each of the areas evaluated. .

Table 1: *Structure of the test before and after 2014-II and method for calculating the global score. Source: own elaboration.*

Test	Characteristic
Before 2014-II	Mathematics (M) Language (L) Biology (B) Physics (F) Chemistry (Q)) Social Sciences (S) Philosophy (Fi) English (I)
Global Index Previous	$IG_1 = \frac{P_B + P_Q + P_F + 2P_S + P_{Fi} + 3P_L + 3P_M + P_I}{13}$
After 2014-II	Critical Reading(LC) Mathematics (MA Natural Sciences (CN)) Social and Sciences (SC) English (IN)
Global present Index	$IG_1 = \frac{3P_{LC} + 3P_{MA} + 3P_{CN} + 3P_{SC} + P_{IN}}{13}$

Each score is in a measurement scale from 0 to 100. Therefore, the resulting global index corresponds to the weighted average of scores on different tests; it's also a measure between 0 and 100. The resulting index is multiplied by 5, which corresponds to the global score that is published on individual reports.

All the tests from 2012-I to 2014-I were re-qualify using classical test theory and item response theory (TRI).

Re-qualifying consisted of evaluating throughout the test applications, which items could be used as anchors to bring them from the previous scale to this new scale (Kolen & Brennan 2004). The result of this reclassification is the primary input to perform re-qualifying models of global scores from scores for each test.

3 Methodology

In 2014-I the requalification of tests made from 2012-I to 2012-I was conducted through methodologies based on classical test theory (TCT). This re-qualifying provides, for each of the tests, the score equivalent to the new structure of the test. A general equating rule for the global scores can be obtained using the individual scores of the tests applied from 2012-I to 2014-II along with their respective requalification. To this end, we count on 1'426.641 evaluated people with scores in the different tests and their respective re-qualifying. The distribution of the number of students per application and the global score average is presented in 2.

Table 2: *Distribution of tests by year and evaluated type. Source: own elaboration.*

Period	2012-I	2012-II	2013-I	2013-II	2014-I
Number of evaluated people	97272	577100	87378	575224	89667
Global average score	260.73	249.92	260.39	250.14	254.81

To equating global scores, one seeks to explain the scores of re-qualifying through the actual scores obtained on the various tests. When the correlation between the different tests is observed, we found high values, both between requalified tests as in the values of the original test scores (matrices (1) and (2)).When making the sphericity test of Bartlett (Grossman et al. 1991) it's concluded that there is a statistically significant correlation between tests at a significance level of 5%.

It indicates the presence of a latent variable that has not been directly observed through raw scores, so we proceed to reduce the dimensionality of the test by an analysis of principal components with orthogonal and normalized rotation, (working with the correlation matrix directly) (Peña 2002). To find this dimensionality reduction a linear transformation of variables is performed so that the greatest amount of variability in the data is retained with the least information loss(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)

To perform the analysis of principal components (PCA) all eight (8) tests are included as active variables. Table3 shows the own values associated with each of the factorial axes and the variance retained by each of the PCA components. Let's note that only the first factorial axis retains already 57.69% of the total variance. The eigenvectors associated with each eigenvalue are shown in table 4.

Table 3: *Eigenvalues and percentage of variance taken by each factorial axis. Source: own elaboration.*

Axis	Eigenvalue	Variance percentage
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 %

Table 4: *Eigenvectors associated with each axis of main components analysis. Source: own elaboration.*

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 7	Axis 8
Biology	-0.36	-0.08	-0.03	0.46	0.33	0.54	0.40	-0.29
Social Sciences	-0.37	0.28	0.01	0.39	-0.04	0.01	-0.78	-0.11
Philosophy	-0.33	0.53	0.52	-0.46	0.33	-0.03	0.11	-0.04
Physics	-0.33	-0.58	0.53	-0.06	-0.39	-0.14	-0.01	-0.31
English	-0.35	0.09	-0.45	-0.51	-0.46	0.44	-0.03	-0.07
Language	-0.36	0.34	-0.18	0.30	-0.38	-0.53	0.45	0.06
Mathematics	-0.35	-0.34	-0.44	-0.25	0.52	-0.44	-0.09	-0.18
Chemistry	-0.38	-0.25	0.09	0.05	0.08	0.12	-0.02	0.87

Given the large mass of variability of the first factorial axis regarding the other axes, it has been decided to create an index, based on this axis, that will be used as a summary measure, for each individual, of the performance in the test based

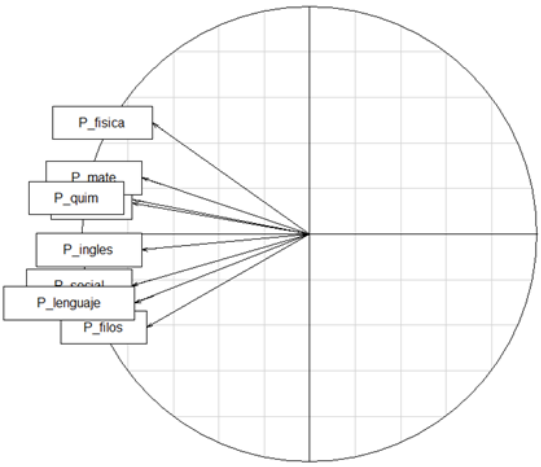

Figure 1: *Factorial plane for the first 2 factorial axis. First axis contains the 57% of variability. Source: own elaboration.*

on the different sub-tests. From now on, we call this index the ACP global index. Onward this will be called ACP global index, denoted as IGACP y IG_{ACP}. That will be calculated as:

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

Where λ_1 is the first eigenvalue, x_j is the standardized score of the test j in the exam and a1j is the main component associated to the first eigenvalue. This created index is used as an explanatory variable for regression models that will be proposed next. The first factorial plane can be seen in 1. LDependent variable will be the global index based in the requalification of the test, by convenience the global index will be divided between 100 to obtain a measure between 0 and 1 and it will be denoted by IG_{Recal}.

3.1 Equalization procedures

3.1.1 Linear Model

In the linear regression model the global index calculated through re-qualifying, is related to the index resulting of the main components analysis. Model can be written as:

$$\text{IG}_{Recal} = \beta_0 + \beta_1 \text{IG}_{ACP} + \epsilon$$

Adjusting the linear model over the indexes of the test this is obtained:

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

3.1.2 Gamma Model

Gamma model helps shaping expected value of variables that are greater than 0, considering that the global index calculated trough re-qualifying (measured from 0 to 1) is value always greater than 0, this model seems to be the right to establish the equivalence. Model will then be defined as:

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

Where $g(\mu_i)$ is the logistic link function given by:

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right). \tag{3}$$

In this case, it's assumed that μ_i has gamma distribution, with $\mu_i > 0$.

3.1.3 Beta Model

For a beta model it's assumed that the median of the random variable is measured between 0 and 1. Having into account that IG_{Recal} is a random variable between 0 and 1, a beta regression is used to shape this variable. In this case it's assumed that for the i-simal individual, $\mathrm{IG}_{Recal,i} \sim \mathrm{Beta}(\mu_{\mathrm{i}}, \phi)$. Therefore regression model would be given by:

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

Logistic link function is used again as presented in (3).

3.1.4 Classical Tests Theory TCT- Equi-percentile Model

From the classical test theory, you can perform two types of equivalence between scores: linear and nonlinear. In the first case can be used the identity function of medians or linear functions. In the second case, the circle of matched angle or the equi-percentile method (Kolen & Brennan 2004).

In this document, the equi-percentile method will be used, since the global score is a continuous variable. In this method it's defined a nonlinear relation between the score scales, establishing equality between the accumulated distribution functions for populations that want to be equated. If you desire to equate scores IG1 and IG2, whose density functions are $F(IG_1)$ y $G(IG_2)$, respectively, then the equalization relation for equi-percentile model will be given by:

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

To establish the relation between both scores, pre-smoothing and post smoothing methods are used to improve adjustment. This smoothing is made with kernel method (generally Gaussian kernel is used) of equalization (Gasser & Müller 1979), then error and accuracy of the estimate are evaluated (Steinberg & Moses 2011).

3.1.5 Model Validation

In the model validation it is verified that following assumptions are met:

A good specification of the model (reset test of Ramsey):

Expected value of residuals equal to zero;

Homoscedasticity (Breusch-Pagan test);

No collinearity (kappa coefficient, as there is only one variable it doesn't apply),

Normality in errors.

The models that are introduced here are validated through analysis of residuals. In each of the models residuals will be calculated as:

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

4 Results

Measures of adjustment are calculated for each of the adjusted models, these are: the sum of errors squares (SCR), the Akaike information criteria (AIC),), Bayesian information criteria (BIC), the R2 for deviance (deviance), coverage and amplitude (Salibian-Barrera & Zamar 2002) of models. As for the sum of errors squares, coverage, and amplitude of Beta model is superior to others. Gamma model is superior regarding the Akaike information criteria and Bayesian and linear regression model with respect to R2 by deviance. These results are shown in 5 and 6.

In order to determine the number of components to retain in the creation of IGACP, a simulation studio was conducted, in which the number of retained components was varying for a sample size 1864 (determined under a simple random sampling) that included all accessible evaluated, where results introduced in Figure 2 (results shown are the average of 100 simulations) were obtained.

It's evident that retaining a single component in the calculation $\widehat{IG}_{ACP}$ is appropriate, since it has less SCE, AIC, BIC and a better adjustment in terms of R^2.

Table 5: *Measure Criteria. Source: own elaboration.*

Measure	Linear	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

Table 6: *Measure Criteria. Source: own elaboration.*

Model	Coverage	Amplitude
Beta	0.95	12536904
Gamma	0.95	12848854
Lineal	1.00	48609023
Equipercentile	1.00	46317587

According to presented criteria in the methodological summary, it's decided to perform the equalization of the global score of the test through the beta model, since it presents a lower sum of errors squares (SCE), a coverage of approximately 95% and an amplitude of prediction interval lower than the found for other models. In this model, the shaped variable will always be in the [0,1] interval, which ensures that none of the predictions will be outside the established in the creation of the current index.

According to found parameters for beta model and the values for the creation of the global index trough ACP, global index would be calculated based on obtained scores in each one of the individual tests as follows:

$$\widehat{IG}_j = \beta_1\lambda_1 \sum_{i=1}^{8} \frac{a_i x_{ij}}{S_i} - \beta_1\lambda_1 \sum_{i=1}^{8} \frac{a_i \bar{x}_i}{S_i} + \beta_0 \tag{4}$$

Where:

x_{ij} is the score obtained in the test i for the *j-simal* person.

$\bar{x}_i$ is the average of scores for the i-simal test.

S_i is the standard deviation estimated for the i-simal test.

a_i are coefficients coming from the analysis of main components, specifically obtained of the eigenvector associated to the first eigenvalue.

λ_i is the first eigenvalue of the analysis of main components.

β_k are the coefficients of the beta regression with $k = 0, 1$.

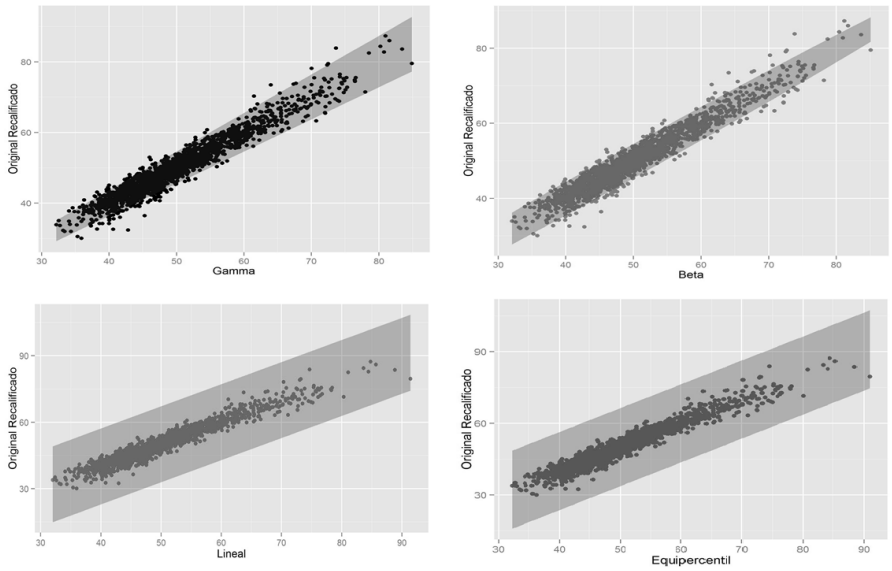

Figure 2: *Amplitude and coverage for proposed models. Source: own elaboration.*

When making estimates of presented parameters in (4) following relation between current global index and the tests scores is obtained:

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

with

$$\begin{aligned}\mathbf{a} =&[-2.466826, 0.0008034867, 0.0007437763, 0.0006196794, 0.0005891047,\\ &0.0005464668, 0.0007962810, 0.0005441301, 0.0008231145]\end{aligned}$$

and

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

5 Conclusions

This exercise allowed us to compare and bring to the same metric the results of the evaluated people of state test Saber 11, obtained as a global score for different applications in which a structural change of test was present.

Results show that with four different methods it's possible to predict the global scores of the evaluated using only the scores obtained in a previous application to

2014-2. However, in assessing adjustment measures of scores the equalization with the use of some generalized linear regression model is more robust than with the use of equating based on conventional methods of TRI and TCT.

In particular, equalization using a Beta regression model showed a better adjustment regarding coverage and amplitude of intervals, since scores to equate are on a limited scale easily carried to a scale from 0 to 1. Not bounded scores like the scores of Saber pro test can be equated through the use of a gamma regression model.

Received: February 19, 2016
Accepted: April 21, 2016

References

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

Chica, S. M., Galvis, D. M. & Ramírez, A. (2011), 'Determinantes del rendimiento académico en Colombia: pruebas Icfes Saber 11°, 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* **72**(1), 341–347.

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

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

Martínez, 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* **30**(2), 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* .