UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Statistics on the evaluation of education
June 2016, Vol. 9, No. 1, pp. 85–97

Proposal of a socio-economic index for students who present Saber Pro tests.[1]

Propuesta para la construcción de un índice socioeconómico para los estudiantes que presentan las pruebas Saber Pro

Edwin Javier Cuéllar Caicedo[a]
ecuellar@contratista.icfes.gov.co

Stalin Guerrero[b]
sguerrero@contratista.icfes.gov.co

Daniela López[c]
daniela.lopez@usantotomas.edu.co

Abstract

This paper presents a proposal for the construction of an index to measure the socioeconomic status of students of higher education, in the context of standardized tests. To that end, we used information of the Colombian Institute for the Evaluation of Education (Icfes) corresponding to 245.461 students who participated in Saber Pro tests in 2012, and also the information of socio-demographic questionnaire. A review of the literature allows defining the variables and dimensions that will shape the socioeconomic status of students. The methodology used in this study is the Principal Component Analysis (PCA), which allows us to reduce the dimensionality of information Considered in the socio-economic variables. Since most variables are categorical, the method of optimal allocation (De Leeuw & Mair 2007) is used to assign values to quantitatively analyze those variables. Results show a good performance of the index compared to some variables of interest, and a positive relation between it and scores in Saber Pro tests.

***Keywords*:** socio-economic index, principal component analysis, optimal allocation, Gifi methods..

Resumen

El artículo presenta una propuesta para la construcción de un índice de nivel socioeconómico de estudiantes de educación superior, en el contexto de pruebas

[1]Cuellar, E. J., Guerrero, S., López, D. (2016) Proposal of a socio-economic index for students who present Saber Pro tests. *Comunicaciones en Estadística*, **9**(1), 85-97.

[a]Statistician, Research Center, Icfes. Colombia.
[b]Statistician, Subdivision of Statistics, Icfes. Colombia.
[c]Student of Statistics, Santo Tomás University. Colombia.

estandarizadas. Para ello, se utiliza la información del Instituto Colombiano para la Evaluación de la Educación (Icfes), correspondiente a 245.461 estudiantes que participaron en el examen Saber Pro en 2012, junto con la información del cuestionario sociodemográfico. Una revisión de la literatura permite definir las variables y dimensiones que conformarán el nivel socioeconómico de los estudiantes. La metodología utilizada para este estudio es el análisis de componentes principales (ACP), el cual nos permite reducir la dimensionalidad de la información contenida en las variables. Dado que la mayoría de variables son categóricas, se propone usar el método de asignación óptima (De Leeuw & Mair 2007) con el fin de atribuirles valores que permitan analizar dichas variables de manera cuantitativa. Se encuentra que el índice construido se comporta bien frente a algunas variables de interés y tiene una relación positiva con los puntajes de las pruebas de las competencias genéricas de Saber Pro.

***Palabras clave*:** índice socioeconómico, asignación óptima, análisis de componentes principales (ACP), métodos de Gifi.

1 Introduction

In recent decades Colombia has suffered multiple processes of social, economic and political transformation which, among other things, have driven the higher education system (MEN 2015). Knowing the socio-economic context of university students in Colombia is a crucial factor to have a better picture of the quality of higher education in the country. Test on the Quality of Higher Education, Saber Pro, is an external standardized test to assess the quality of higher education and it's mandatory from 2009 [1]. This test evaluates five generic competencies (critical reading, quantitative reasoning, written communication, English and generic competencies) and specific common competencies to different groups of reference[2].

This research seeks to measure the socioeconomic status of university students, in order to provide a useful input for the processes of evaluation and characterization of students of higher education in Colombia.

In the literature review, it was found that there are few studies in this field in Colombia, which makes relevant the construction of an index consisting of a set of economic, social and cultural factors. These factors are known as unobservable or latent variables (Díaz & Barrios 2002)), and can be approximated through observable variables, such as monthly household income, household endowment, and the family context, among others.

The fundamental base for an index construction is the definition of the variables that comprise it. Researches in education proposing socioeconomic indexes for

[1] Law 1324 of July 13, 2009, and Decree 3963 of 2009f

[2] Reference groups are sets of academic programs built for the interpretation of the exam results and are based on the classification of knowledge areas and basic knowledge cores of SNIES (National System of Information of Higher Education), as also in the classification of formations of Unesco

students of primary and secondary education are very common in literature (Gil-Flores 2013, Paz-Navarro et al. 2009, Porcel et al. 2010, Saz 2006) and those are used as a reference in this research.

Socio-economic level of students is a critical factor in the context of education and is related to their academic performance (Acevedo & Jaramillo 2007, Armenta et al. 2008, **?**, Contreras et al. 2008). Parental occupation and educational level are variables commonly used in this context. (Vargas 2013, Saz 2006). Similarly, we included in the definition of socioeconomic level, besides the variables estimating the economic level, those that reflect the possession of a cultural capital (Jabnoun 2009, Saz 2006, Hernández & González 2011).

2 Statistical Methodology

2.1 Principal components analysis (PCA)

PCA methodology allows reducing the data dimension, preserving the maximum information as possible. Therefore, a system of q dimension can be reduced to a system of smaller dimension through the generation of new variables (components) resulting from the linear combination (weighted sum) of original variables.

In a study made over $n-$- individuals through $p-$variables $X_1, ..., X_p$, it's possible to find new variables noted by Y_k that are linear combinations of original variables X_j, , and depend on certain conditions. In this sense the firs principal component Y_1, determined, which contains the bigger amount of total variability contained in data. So, we have:

$$Y_1 = \gamma_{11}X_1 + \gamma_{12}X_2 + ... + \gamma_{1p}X_p$$

Where weightings $\gamma_{11}, ..., \gamma_{1p}$ are chosen such that maximize the reason of variance of Y1 to total variance; with the restriction: $\sum_{j=1}^{p} \gamma_{1j}^2 = 1$.

On the other hand, the suggested criteria to select the number of components is based on the variability to be maintained. Since the sum of original variances is the trace of the variances and covariances matrix $\mathbf{S}$ and it's equal to the sum of eigenvalues of $\mathbf{S}$, each principal component explains a proportion of total variability, which is calculated as follows:

$$\frac{l_k}{tra(\mathbf{S})}$$

This quotient is called the proportion of total variability explained by the k components.

The use of PCA for the index construction in different fields has shown advantages over other methodologies, due to its easy implementation and because assumptions of normality, homoscedasticity and linearity can be ignored (Estévez 2002).

2.2 HOMALS

This method is part of the denominated Gifi methods and it's also known as homogeneity analysis. It has multiple functions applicable to multivariable data analysis. However, for this study, it was considered only as a tool of optimal assignation, that allows assigning a quantitative score to categorical variables of nominal or ordinal type by an iterative procedure (De Leeuw & Mair 2009).

Following the notation of De Leeuw & Mair (2009), for $i = 1, ..., n$ individuals in $j = 1, ..., m$ categorical variables with different response levels K_j, a block of matrixes $G = [G_1, ..., G_m]$, is built, where G_j are dummy indicator matrixes of dimension $n \times k_j$. Besides binary diagonal matrix M_j of dimension $n \times n$ for each variable j. is defined. Elements of diagonal (i, i) take value 0 when individual i has a missing value in variable j, or 1 in other case.

Later unknown matrixes $\mathbf{X}$ $(n \times p)$ y Y_j $(k_j \times p)$, which contain scores of items and categories respectively, are considered. To find such values it is established the following loss function:

$$\sigma(\mathbf{X}; Y_1, ..., Y_m) = \frac{1}{m} \sum_{j=1}^{m} tr(X - G_j Y_j)' M_j (X - G_j Y_j)$$

Both matrixes must simultaneously minimize this function by the normalization of $u' M_j \mathbf{X} = 0$ y $X' M_j \mathbf{X} = I$. From a theoretical point of view, it's concluded that the loss function represents the sum of squares of $(\mathbf{X} - G_j Y_j)$, which involves the obtained scores for individuals and categories.

In programming language `R` [3], the function `homals` of the `homals` package (De Leeuw & Mair 2007), is available. This function makes this iterative process and assigns the optimal scores for each category of each variable.

3 Conceptual construction

The information supplied by ICFES was used in the construction of a socioeconomic level index (INSE), corresponding to 245.461 students from Colombia who took the test of quality of higher education, Saber Pro and the information of the socio-demographic questionnaire consisting of 64 questions about family, social and economic context. It's important to note that those students were close to finish their high education cycle, with the requirement of having approved at least the 75%of the academic credits of their respective undergraduate program in 2012.

The INSE is built from three dimensions: *home's economic potential*, *cultural capital* and *housing endowment*. For the home's economic potential dimension, three proxy variables related to the home's economic income are taken into account: the

[3] It's a free software and with integrated development environment (IDE) for statistic computing and graphics.

price of tuition, monthly family income and SISBEN [4], all of them are ordinal variables.

For the cultural dimension, four variables are taken into account: mother's and father's education and occupation. The education variables present response answers like: complete primary, incomplete primary, complete secondary, incomplete secondary, complete technician, incomplete technician, incomplete undergraduate, incomplete undergraduate, postgraduate, among others. These are considered ordinals variables and are very important if we consider that parents' education is a variable that has a high correlation with the education of their children. This fact due to the socioeconomic mobility difficulties and to intergenerational inheritance (Nuñez 2012). For parental occupation, the categories are student, retired, home, entrepreneur, independent worker among others. The above variables are considered important due to the existing relation between the cultural capital and the academic performance. (Coleman 1988, Gil-Flores 2013).

The last dimension corresponds to housing endowment and it is built from the information about the possession of some elements that could be important or make the difference between the housing socioeconomic conditions of a student and other. In similar studies, elements of endowment or availability of spaces have been used as an information about the lack of physical elements that are necessary for the students learning (Piñeros & Rodríguez 1998, Ravela 2005).

For the PCA application was necessary to use only numerical variables. However, the variables defined for each dimension in the index corresponds to categorical variables. Therefore , transformation to numerical variables is a requirement for the index analysis and construction. It's decided to apply the method of optimal assignation to all variables, in order to give a quantitative score to each category, considering that those are nominal or ordinal. This procedure was made using programming language R with help of the `homals` function.

Once categorical variables turned into numerical, the PCA is applied to each one of the dimension, to summarize such variables in components that retain the larger possible amount of information (maximum variance). In order to retain a number of necessary components a parallel analysis of Horn (1965), was made under which, for each dimension, only one component must be retained. Finally, the index corresponds to the first coordinate of the PCA made with the three scales obtained from each of the dimensions.

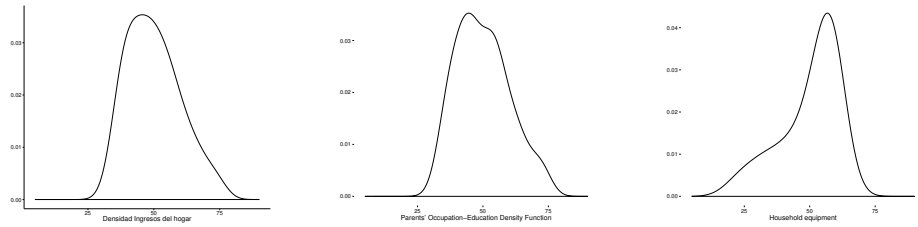
4 Results

Under the scheme presented above, the information is analyzed and the corresponding procedures performed. It's found that for each of the dimensions it's enough with the use of a single component for the construction of the scale. This indicates that variability and contained information in the first component is enough

[4]System of potential beneficiaries for social programs.

to represent the information contained in the variables involved in the construction of each one of the dimensions. To confirm the use of a single component we use the parallel analysis of Horn (Horn 1965).

Figure 1 shows the density functions obtained for each dimension: home's economic potential, cultural capital and housing endowment.



(a) (b) (c)

Figure 1: *Density function for dimensions of a) home's economic potential, b) cultural capital and c) housing endowment. Source: ICFES.*

Coordinates of variables for each component are presented in tables 1, 2 y 3.

Table 1: *Coordinates by variable for the home's economic potential dimension l. Source: own*

	Dim.1	Dim.2	Dim.3
Tuition price	0.77	-0.53	0.35
monthly family income	0.84	-0.06	-0.54
Sisben level	0.75	0.62	0.25

Table 2: *Coordenates by variable for the cultural capital dimension. Source: own*

	Dim.1	Dim.2	Dim.3	Dim.4
Education level of the mother	0.73	-0.36	-0.42	-0.41
Education level of the father	0.70	-0.52	0.29	0.39
Occupation level of the father	0.65	0.41	0.58	-0.27
Occupation level of the mother	0.65	0.55	-0.42	0.31

Table 3: *Coordinates by variable for the dimension housing endowment. Source: own*

	Dim.1	Dim.2	Dim.3	Dim.4
Internet	0.86	-0.14	-0.10	-0.49
TV service	0.62	0.78	0.06	0.07
Telephone	0.70	-0.27	0.64	0.18
Computer	0.76	-0.22	-0.51	0.33

In the obtained results for the analysis by dimension, it's found that the explained variability by the first component for each dimension i.e., home's economic potential, cultural capital and housing endowment is of 62%, 47%, and 55%, respectively. Table 4 shows the contribution of each dimension to the constructed INSE.

Table 4: Contribution of dimensions to socioeconomic index. Source: own

	Dim.1	Dim.2	Dim.3
Economic potencial	41.64	0.54	57.82
Cultural capital	26.45	60.51	13.04
Housing endowment	31.91	38.95	29.14

From table 4 we can conclude that home's economic potential is the dimension that has a bigger contribution to the INSE, followed by the housing endowment and the cultural capital. In 2, shows the graphic of the obtained INSE's density.

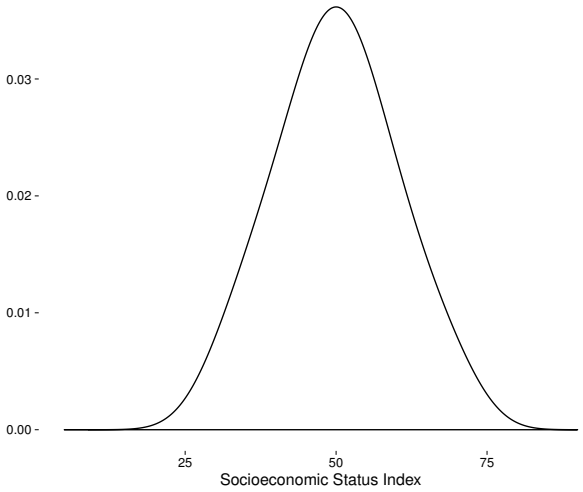


Figure 2: *Density function of the socio-economic index. Source: own elaboration.*

From figure 2 it's important to mention that there's not asymmetry, the index seems quasi-symmetric. This behavior could be due to the fact that socioeconomic conditions of students in higher education in Colombia are higher than the conditions of the total Colombian population, case in which an index with positive asymmetry would be expected. In table 5 are presented the descriptive statistics corresponding to obtained INSE.

Table 5: *Descriptive statistics INSE. Source: own elaboration.*

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
24.69	43.14	49.98	50.00	56.76	74.72

Con el fin de evaluar la validez externa del índice construido, es necesario analizar el comportamiento del INSE frente a algunas variables de interés como lo son el estrato socioeconómico y el origen de la institución educativa.

El estrato socioeconómico[5] eis a variable that is in the questionnaire but was not introduced in the index construction. Figure 3 shows the graphic of density by strata and it shows also that the high levels of stratifications[6] are associated to high INSE scores, while low levels of strata are associated to low INSE scores.

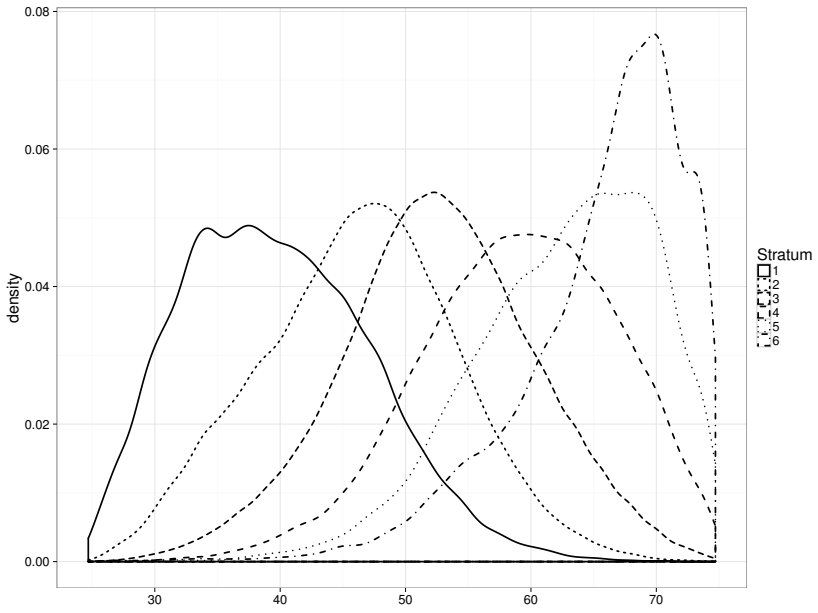


Figure 3: *INSE by socio-economic stratification. Source: own elaboration.*

Figure 4 shows the INSE behavior among students of higher education of official and non-official funding, students of non-official funding are associated with relatively higher INSE scores.

[5]Los estratos socioeconómicos son una herramienta que utiliza el Estado colombiano (Ley 142 de 1994, Artículo 102) para clasificar los inmuebles residenciales de acuerdo con los lineamientos del DANE, el cual tiene en cuenta el nivel de pobreza de los propietarios, la dotación de servicios públicos domiciliarios, la ubicación (urbana, rural), asentamientos indígenas, entre otros.

[6]Dentro del estrato, el nivel mínimo es 1 y el máximo es 6.

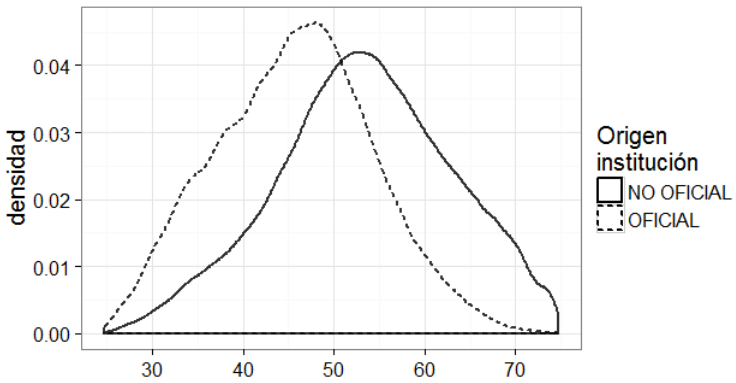


Figure 4: *INSE by origin. Source: own.*

In higher education level some students are the head of household. For this reason, the contrast in the INSE between the students who are head of the family or not is presented in figure 5

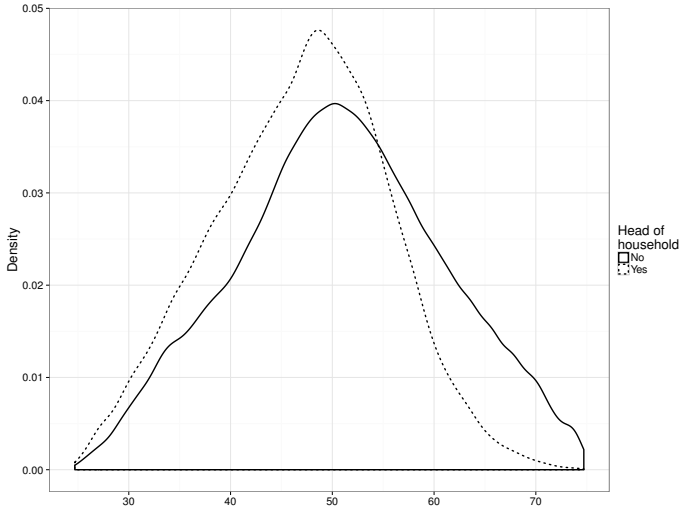


Figure 5: *INSE - Student is head of family. Source: own elaboration.*

In figure 5 it's observed that some students that are head of the family show lower INSE scores that those still depending on their parents or some other person. The presentation of the three previous results show a good index behavior compared with variables of interest and validate its construction.

Finally, it's important to understand if it exists some relationship between the

socio-economic level (NSE)[7]) and the results obtained by students. As it's shown in Table 6, the relationship between socioeconomic level and scores in the tests is similar to the different tests of generic competencies: English, critical reading, quantitative reasoning, written communication, critical communication and citizenship skills.

Table 6: *Scores English test by INSE and reference group. Source: own*

Test	Low	Medium- Low	high medium	High
English	9.570	9.798	10.027	10.933
Written communication	9.859	10.006	10.100	10.327
Critical reading	9.655	9.884	10.030	10.421
Citizenship skills	9.702	9.908	10.041	10.366
Quantitative reasoning	9.693	9.886	10.003	10.389

In table 7 are, as an illustration, the average scores for the university reference groups, obtained for the English tests by INSE. This behavior is very similar to the results found in the other generic skills.

Table 7: *English test scores by INSE and reference group. Source: own.*

	Reference group	Low	Low medium	High medium	High
1	Administration	9.568	9.725	9.937	10.904
2	Architecture	9.716	9.967	10.207	11.103
3	Fine arts	9.920	10.452	10.733	11.685
4	Agricultural Sciences	9.614	9.849	10.083	10.717
5	Military Sciences	9.936	10.273	10.181	10.557
6	Natural Sciencies	10.116	10.508	10.814	11.904
7	Social Sciences	9.580	9.749	9.979	11.174
8	Communication	9.861	10.056	10.265	11.149
9	Law	9.644	9.706	9.860	10.626
10	Economy	9.885	10.241	10.539	11.842
11	Education	9.612	9.893	10.182	10.872
12	Humanities	9.966	10.421	10.944	12.239
13	Engineering	9.980	10.217	10.424	11.296
14	Medicine	10.354	10.569	10.620	11.366
15	Recreation	9.513	9.664	9.883	10.327
16	Health	9.605	9.739	9.868	10.218

The relationship between NSE and the English test scores is direct, the higher the NSE higher the English test score. This relationship is maintained in the different university reference groups.

[7]We call the socioeconomic level to the classification of the INSE by means of its quartiles, in this way, the students who obtain an INSE that is in the first quartile will be assigned to the group with a low SES, the second quartile to the group with an SES average. And the two remaining groups are assigned in this way.

5 Conclusions

The methodology introduced in this article may be replicable in studies in which a socioeconomic index construction is required, particularly in studies about academic performance in higher education. It's expected that future works contribute to this analysis and approaches to this reality.

External validity of INSE is verified trough the index behavior about variables of interest as socioeconomic stratification, funding of educational institution and whether the student is or not the family head. Concerning these three variables INSE presents the expected behavior, according to the characteristics of subpopulations that define the variables of interest.

Observed symmetry for INSE of the students in Saber Pro contrasts with the results of socioeconomic conditions of students in Saber 11 which shows a positive asymmetry. This behavior could be explained by the coverage effect in education associated with economic barriers among different education levels (according to statistics from the Ministry of National Education[8].

Received: February 2, 2016
Accepted: April 19, 2016

References

Acevedo, S. & Jaramillo, A. (2007), 'Perfil socioeconómico de los estudiantes de pregrado EAFIT. Medellín: Universidad EAFIT'.

Armenta, N., Pacheco, C. & Pineda, E. (2008), 'Factores socioeconómicos que intervienen en el desempeño académico de los estudiantes universitarios de la Facultad de Ciencias Humanas de la Universidad Autónoma de Baja California', *Revista IIPSI* **1**(1), 153–165.

Caso-Niebla, J. & Hernández, L. (2007), 'Variables que inciden en el rendimiento académico de adolescentes mexicanos', *Revista Latinoamericana de Psicología* **39**(3), 487–501.

Coleman, J. S. (1988), 'Social capital in the creation of human capital', *The University of Chicago Press* **94**, 95–120.

Contreras, K., Caballero, C., Palacio, J. & Pérez, A. (2008), 'Factores asociados al fracaso académico en estudiantes universitarios de Barranquilla (Colombia)', *Psicología desde el Caribe* **22**, 110–135.

[8]Ver `http://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-212350_Estadisticas_de_Educacion_Superior_.pdf` the coverage rate of higher education is close to 50%).

De Leeuw, J. & Mair, P. (2007), 'Homogeneity analysis in R: The package homals', *Department of Statistics, UCLA* .

De Leeuw, J. & Mair, P. (2009), 'Gifi methods for optimal scaling in r: The package homals', *Journal of Statistical Software* **31**(4), 1–30.

Díaz, S. D. & Barrios, G. H. (2002), 'Eficiencia escolar y diferencias socioeconómicas: a propósito de los resultados de las pruebas de medición de la calidad de la educación en Chile.', *Educação e Pesquisa* **28**(2), 25–39.

Estévez, J. F. (2002), 'La construcción de un Índice cuantitativo sobre educación superior utilizando la técnica de análisis de componentes principales.', *Revista de la Educación Superior.* **31**(121), 138–153.

Gil-Flores, J. (2013), 'Medición del nivel socioeconómico familiar en el alumnado de educación primaria', *Revista de Educación* **362**, 298–322.

Hernández, E. & González, M. (2011), 'Modelo de ecuación estructural que evalúa las relaciones entre el estatus cultural y económico del estudiante y el logro educativo', *Revista Electrónica de Investigación Educativa* **13**(2), 188–203.

Horn, J. L. (1965), 'A rationale and test for the number of factors in factor analysis.', *Psychometrika* **30**(2), 85–179.

Jabnoun, N. (2009), 'Economic and cultural factors affecting university excellence', *Quality Assurance in Education* **17**(4), 416–429.

MEN (2015), Colombia un país que avanza hacia el mejoramiento de las oportunidades de acceso a la educacion superior., Technical report, Ministerio de Educación Nacional.

Nuñez, J. A. (2012), Pobreza, empleo y movilidad social: Evidencia e interpretación de los problemas sociales en Colombia, PhD thesis, Pontificia Universidad Javeriana.

Paz-Navarro, L., Roldán, R. & González, M. (2009), 'Funcionamiento familiar de alumnos con bajo rendimiento escolar y su comparación con un grupo de rendimiento promedio en una preparatoria de la universidad de guadalajara', *Revista de Educación y Desarrollo* **10**, 5–15.

Piñeros, L. & Rodríguez, A. (1998), Los insumos escolares en la educación secundaria y su efecto sobre el rendimiento académico de los estudiantes: un estudio en Colombia, Technical Report 20934, El Banco Mundial.

Porcel, E., Dapozo, G. & López, V. (2010), 'Predicción del rendimiento académico de alumnos de primer año la facena (unne) en función de su caracterización socioeducativa.', *Revista Electrónica de Investigación Educativa* **12**(2).

Ravela, P. (2005), Estudio de los factores institucionales y pedagógicos que inciden en los aprendizajes en escuelas primarias de contextos desfavorecidos en Uruguay, Technical report, ANEP/MECAEP/UMRE.

Saz, M. M. A. (2006), 'Influencia del nivel socioeconómico y cultural en el rendimiento de los estudiantes de tercero básico y graduandos del año 2006', *Dirección General de Evaluación e Invetigación Educativa* .

Vargas, G. M. G. (2013), 'Factores asociados al rendimiento académico en estudiantes universitarios desde el nivel socioeconómico: Un estudio en la Universidad de Costa Rica', *Revista Electrónica Educare* **17**(3), 57–87.