
Construction of a family socioeconomic index for the students who present the Saber 11 test ¹

Construction of a family socioeconomic index for the students who present the Saber 11 test

Laura Melissa Cruz Castro^a
lacruz@contratista.icfes.gov.co

Felipe Ortiz^b
aortiz@contratista.icfes.gov.co

Diego Fernando Lemus^c
dlemus@icfes.gov.co

Abstract

The aim of this study is to construct an index to measure the socioeconomic status of students who took the standardized tests Saber 11 ° in 2012. This study uses the information of the Colombian Institute for the Evaluation of Education (ICFES) corresponding to 424.916 students who participated in the tests and also answered also the socio demographic questionnaire, consisting of 58 questions seeking to know some aspects related to their family, social, economic and cultural context. In the first part of the study we made a review of the specialized literature, in order to define the variables and dimensions to be used in the construction of the socioeconomic status. Once the possible variables were selected for inclusion in the index, we passed to the second part in which a Principal Component Analysis (PCA) is made, summarizing the most information as possible into a single component called socioeconomic index, using the proposed methodology in Flores et al. (2008). Since most variables are categorical, we proposed to use the method of optimal allocation in order to attribute weights that allow the quantitative analysis of those categories (De Leeuw & Mair 2009).). Finally we made an analysis of the students' academic performance in the mentioned tests with the socioeconomic index, where a high correlation between the obtained score and the constructed index is presented.

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

¹Cruz, L. M., Ortiz, F., Lemus, D. (2016) Construction of a family socioeconomic index for the students who present the Saber 11 test *Comunicaciones en Estadística*, **9**(1), 73-84.

^aStatistician, Subdivision of Statistics, Icfes. Colombia.

^bConsultor, Division of Evaluation, Icfes. Colombia.

^cDeputy Director of Instrument Design, Division of Evaluation, Icfes. Colombia.

Resumen

El objetivo del presente estudio es la construcción de un índice que permita medir el nivel socioeconómico de los estudiantes que presentaron las pruebas estandarizadas Saber 11° del 2012. En este estudio, se utiliza la información del Instituto Colombiano para la Evaluación de la Educación (Icfes) correspondiente a 424.916 estudiantes que participaron en dichas pruebas y además respondieron el cuestionario sociodemográfico compuesto por 58 preguntas, las cuales buscan conocer algunos aspectos relacionados con el contexto familiar, social, económico y cultural del estudiante. En la primera parte del estudio se hace una revisión de la literatura especializada con el fin de definir las variables y dimensiones que se utilizarán en la construcción del nivel socioeconómico. Una vez se seleccionan las posibles variables a incluir en el índice, se pasa a la segunda parte en la que se realiza un análisis de componentes principales (ACP), resumiendo así la máxima información posible en un solo componente que se denomina índice socioeconómico, utilizando la metodología planteada en Flores et al. (2008). Dado que la mayoría de variables son categóricas, se propone usar el método de asignación óptima con el fin de atribuirles pesos que permitan analizar dichas categorías de manera cuantitativa (De Leeuw & Mair 2009). Por último, se realiza un análisis del desempeño académico de los estudiantes en dichas pruebas con el índice socioeconómico, donde se presenta una correlación alta entre el puntaje obtenido y el índice construido.

Palabras clave: índice socioeconómico, asignación óptima, análisis de componentes principales.

1 Introduction

The importance of education in a society goes beyond of what we imagine. Education in the tool of social mobility that reproduces privileges and tries to compensate and make a new distribution between the different strata or social treads. The evaluations measurement allow us to find and set out strategies to get an equable distribution of social' resources (Bowles & Gintis 1997, UNESCO 2008, Pérez 2000, Sánchez & Otero 2012).

In this sense, the present educational systems have been worried about the coverage, quality and equity measurement for which they have developed methodologies to evaluate and qualify education, as well as the impact of associated factors.

The ICFES, an organism attached to the National Education Ministry, whose mission is to evaluate education in all its levels, and research about the factors that influence its quality, in order to offer information to improve it.

Among the factors that influence educational quality it is the socioeconomic level, defined as the interaction between the family incomes, the grade of familiar education, occupation and social capital. The students of low socioeconomic level obtain lower learning results and at the same time a lower probability of finishing their studies and continue with superior studies (Gil 2013). Previous researches

proved that family socioeconomic conditions are a good predictor of academic accomplishment of students. ICFES built a socioeconomic index that summarizes information of students' socioeconomic characteristics.

Socioeconomic index represents the hierarchical ranking of an individual or his family in terms of access or control of some combination of products as richness, power and social status (Mueller & Parcel 1981). It is also an index that reveals a relative position against a number of dimensions that describe socioeconomic level.

There is a diversity of socioeconomic index in literature, each one aims to find a better way to reflect socioeconomic *status* of people through a set of relevant items. Such composition has depended on the advance of studies, the availability of information and the different models proposed that have nourished the index capacity to adequately predict the social and economic characteristics of individuals (Center for Education Statistic. 2012).

This article explains the construction of a socioeconomic index proposed by the ICFES and the different results that are obtained through an analysis of the students scores in the Saber tests 11° and the interpretations of the last with respect to the socioeconomic index.

2 Conceptual construction

2.1 Background

Since the year 2009 the ICFES has tried to develop different socioeconomic indexes. From a missionary perspective it is of vital importance for ICFES to study the effect of socioeconomic characteristics about academic performance; and like index is a significant agent and differentiator of academic achievement, the constructions of socioeconomic index is fundamental; in 2012 a socioeconomic index was developed using methodologies of the item response theory (TRI), logistic one-dimensional models for TRI models of dichotomic and polytomous items de (for nominal and ordinal case). This methodology was implemented in 2012 and applied in the Saber tests 3°, 5° and 9° made annually since 2012.

2.2 Variable selection

The population on which this research was made is built by the students that took the Saber 11° test in the period 2012-I and 2012-II corresponding to A and B calendar, respectively.

When a student presents the state test Saber 11°, he is asked to fill a registration form of 50 questions asking for information about housing, family and parents occupation among other topics. These variables will be selected for the construction of the socioeconomic index over the 424.916 students that presented the Saber

11° tests during 2012 and answered the whole application form of which 55 % are women and 45 % are men.

When addressing the measurement of socioeconomic status there is a consensus on the inclusion of three fundamental aspects: income, educational level and parents occupation (Hauser 1994). However, in some other studies variables as overcrowding and housing endowment, among others, are specified (Piñeros & Rodríguez 1998, Ravela 2005).

In a first analysis it has been decided to include three dimensions: *family income*, *parents occupation and education* and, finally, *household crowding and endowment*; each of these was defined based on the questions and variables available in the registration form and will be described below.

2.2.1 Family income

In this dimension we'll count on three variables: the first corresponds to Sisben¹, punctuation given to an individual by creating an index for measuring poverty; the second corresponds to the monthly family income, and finally, the price of the pension in the educational institution. This dimension is built on the assumption of a positive correlation between these variables, which, in this case, correspond to three ordinal variables.

In the case of Sisben, we have 5 possible answers: *Level 1*, *Level 2*, *Level 3*, *classified in other level and is not classified*, being level 1 and 2 subsidized and the 3 not subsidized²

2.2.2 Occupation and education of parents

The level of education of parents is a nominal variable. Among possible response categories for the parents occupation are: *entrepreneur*, *small entrepreneur*, *employee with position as director or manager*, *worker or operator*, among others.

Despite finding in some applications the possibility to analyze only the highest level of occupation (either the father or the mother), it was decided to use the two levels, both the mother and the father, because of the differences presented for Colombian case³.

Educational level of parents is measured by the variable of *maximum level of education reached*. In this variable are possible response categories that correspond to an ordinal scale like: *none*, *incomplete primary*, *complete primary*, *incomplete secondary*, *complete secondary*, *complete technician*, *incomplete technician*, *complete technologist*, *incomplete technologist*, *incomplete university*, *complete university* and *postgraduate*.

¹Sisben was designed as an instrument for targeting social programs, ensuring that the social spending is assigned to the poorest and most vulnerable groups.

²Level 1= from 0 to 47,99 points, level 2= from 48,00 to 54,86 points

³53.3% of mothers have the house for occupation, in front of a 12% of fathers.

2.2.3 Household crowding and endowment

In housing we have the crowding as first variable, calculated as follows:

$$Overcrowding = \frac{\text{Number of people}}{\text{Number of rooms}} \quad (1)$$

The above is a continuous variable that has a minimum value of 0.1 and a maximum of 12. This variable is coded as 1 when it is less than or equal to 3 and zero otherwise.

According to the definition of critical crowding we say that a 4.8% of students are in critical crowding⁴.

It also has the endowment household, defined from dichotomous variables corresponding to having or not having certain resources such as automobile, telephone and internet. These variables are chosen, since they represent an advantage or disadvantage over the study and the academic support that a student could need, besides they represent the mayor discrimination in the population.

As presented throughout this section, the construction of a socioeconomic index not only involves choosing appropriately the variables, given the context of the evaluation and available resources, it also involves a detailed analysis of how it is going to be built and the assumptions that will be made.

Despite the difficulty of building a descriptive and accurate socio-economic index, the importance of its inclusion in the studies and the information that it offers convert it in an useful tool for different contexts (such as education and health), each culture is unique and therefore it is necessary to have an appropriate measurement for each social context. In the next section the tools that make the statistical methodology that supports the construction of the proposed index are selected.

2.3 Statistical methodology

After doing a descriptive analysis of the variables and debugging the database, a process of reduction (principal component analysis) is done, this for each of the identified and described above dimensions: *income, level of occupation or parental education, household crowding and endowment*.

The decision to conduct an analysis of main components, start from the facility that gives us this methodology to be applied, besides the facility in the interpretation of obtained results. It is noteworthy that when making a reduction of variables for each dimension, we ensure the correlation between these and the reduction allows us a higher percentage of variability explained by the first dimension.

⁴This is an indicator that seeks to capture the critical occupation levels of housing resources by the group that inhabits it. Homes with more than three people per room are considered in this situation. (DANE unmet basic needs).

Because there are categorical variables within the dimensions (ordinal or nominal), it is necessary to use a transformation that allows us to notice the categorical variables as continuous. For this, we make an analysis of homogeneity in order to maximize the correlations between variables and thus an optimal numerical scale is assigned to the categorical variables values.

The general idea of homogeneity analysis is to make a plan set in the space of all the n individuals and the p categories of all variables. The technique, in its simplest form, makes that the sum of squares of the $n \times p$ lines is as small as possible subject to the normalization of the scores of objects (the matrix $n \times p$ must be orthonormal) (De Leeuw & Mair 2009).

The construction of the socioeconomic index at the end results of the construction of a single index from the scales created for each dimension.

3 Results

For each dimension was obtained a single factor that explains a 72.38% of the total variability for the case of family incomes, a 47.3% for occupation and educations of the parents and a 50% for household crowding and endowment. We used, additionally, the parallel analysis of Horn (1965) to confirm the selection of the first component, both at the level of each dimension as the total. Tables 1, 2 and 3 show the coordinates and contributions of each one of the variables in the constructed dimensions for the socioeconomic index proposed.

Prior to analysis of principal component, the optimal quantifications for the proper treatment of categorical and quantitative variables were generated using the Homals methodology. However, we omitted it in this paper, since they are not considered as relevant in the analysis as the weights of the variables and the index validation shown below.

Table 1: *Coordinates and contribution of variables to the family income dimension.*
Source: own elaboration.

	Variable	Coordinate	Contribution
	Sisben	0.80	29.69
	Monthly family income	0.87	35.17
	School pension	0.87	35.14

It's relevant to mention that home's socioeconomic conditions are mostly characterized by the value of the pension at school than by the Sisben level, this generates a wake-up call for the government in improving the Sisben to better capture the socioeconomic conditions of the homes and invites universities to use the pension's value as a criteria to define the enrollment price for students of higher education.

For the parental occupation and education dimension it's important to highlight that educational levels of both parents have a bigger relevance than occupation,

Table 2: *Coordinates and contributions of variables in the household crowding and endowment dimension. Source: own elaboration.*

Variable	Coordinate	Contribution
Mother's occupation	0.62	19.44
Father's occupation	0.51	13.27
Mother's education	0.82	34.75
Father's education	0.80	32.54

Table 3: *Coordinates and contributions of variables in the household crowding and endowment dimension. Source: own elaboration.*

Variable	Coordinate	Contribution
Auto	0.63	20.06
Internet	0.82	34.26
Phone	0.78	30.54
Overcrowding	0.55	15.14

because education levels allow to differentiate better homes of high and low socioeconomic conditions.

In the third dimension proposed we found that the household endowment allows a bigger discrimination of socioeconomic conditions than its crowding. Possessions as the car, internet or telephone present a bigger contribution to the socioeconomic index formation than endowment.

It's important to understand that till this point we have the construction of 3 scales corresponding to each dimension. The decision of doing a scale for dimension instead of using the chosen variables comes from the variability's reduction in the dimension (since we looked for each dimension variables a positive correlation), based on obtaining a better index following the methodology used by Gil Gil (2013).

After obtaining the weights of variables by dimension, it's also necessary to know the weight of each dimension in the constructed index, or what could be called factorial weight. Table 4 shows the corresponding weights of each dimension in the obtained index.

Table 4: *Contribution of dimensions in the socioeconomic index (INSE). Source: own elaboration.*

Variable	Dim.1
Family income	35.98
Parents occupancy and education	30.80
Overcrowding	33.22

Distribution of socioeconomic index is of big importance, since it helps us to interpret the individuals' behavior and the possible concentrations. It has been

created so that it has an average of 50 with a standard deviation of 10, it reaches a maximum of 84.36 and a minimum of 24.26 as it's shown in table 5. This allows to consider that a student that presents Saber 11° tests in a different year may have lower or higher socioeconomic conditions than the reached here.

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

Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.
24.26	41.95	47.71	50.00	55.63	84.36

Figure 1 shows the graphic of density corresponding to the socioeconomic index built, it has an asymmetric to the right, which implies a higher concentration of individuals with a low socioeconomic level. However it's necessary to find variables to validate the obtained results for socioeconomic index.

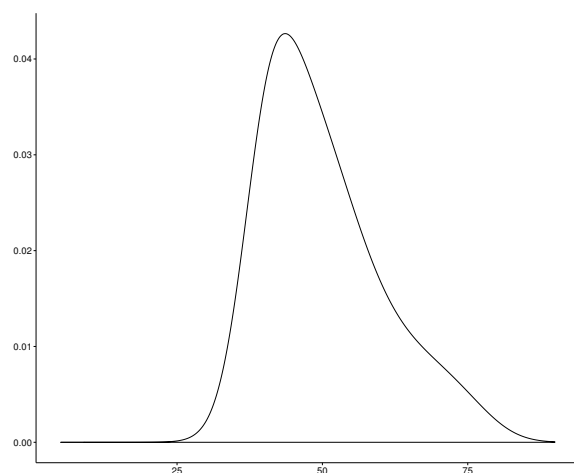


Figure 1: *Distribution of socioeconomic index. Source: own elaboration.*

The socioeconomic stratus defined by the DANE (National Administrative Department of Statistics) is built from the characteristics of the housing and its surroundings; this variable was not taken into account during the index construction that's why it may be used as a validation variable.

Figure 2 shows the distribution of socioeconomic index proposed by socioeconomic stratus. Socioeconomic index shows low values for individuals stratified in stratus 1, 2 and 3. It shows besides high values for individuals in stratus 4, 5 and 6. On the other hand you can see a change in the index's asymmetry of each stratus, characterized by a higher concentration of low resources students in stratus 1 and of high resources in stratus 6.

The importance of this relation between the constructed index and the socioeconomic stratus is precisely the validation of a direct relation between those two

variables, having into account that socioeconomic stratus was not used in the index's construction.

There is a set of variables that were chosen in order to validate and understand in a large scale the index meaning. These variables were the calendar and the nature of school, in calendar there are two possible options: calendar A and calendar B; although this difference should be a difference with regard to the beginning period and the ending of classes. There's also a difference with regard to family income, since schools in calendar B are also more expensive schools (taking as reference the price of the pension in school), which would represent a higher difficulty of access for people with no money⁵

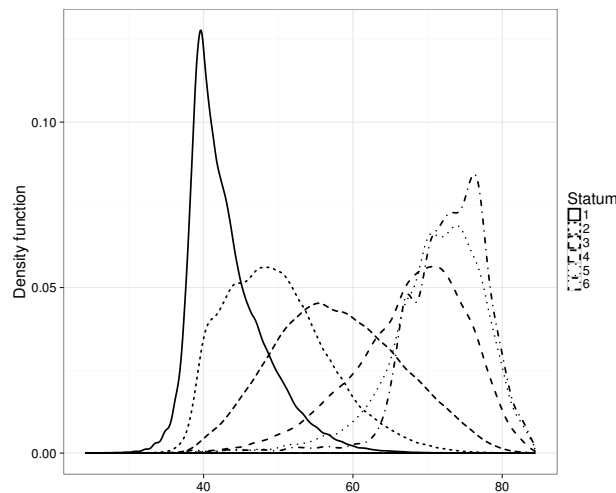


Figure 2: *Index by economic stratum. Source: own elaboration.*

As for the nature of the school, it would be expected the socioeconomic index to present a difference between official and non-official schools, since family incomes and level of education of parents represent a better chance for the child of accessing to a non-official school.

Figure 3 shows the obtained graphics of socioeconomic index with respect to the nature and calendar of school which student attends.

Last but not least, achievement or academic accomplishment of students in relation to their socioeconomic level represents a fundamental aspect of the study, since the construction of it lies in the objective of measuring or approaching the measurement of academic success, given the student's socioeconomic level, or at least, finding if there is a relationship between the index created and academic

⁵In calendar B more than 51% of students pay more than 250.000 COP, compared to 4% of students who say to belong to A calendar and pay more than 250.000 COP. Besides 82% of students of calendar A don't pay pension compared to a 12% of calendar B that don't pay it.

achievement.

It was found in the analysis that one of the biggest differences regarding the performance in the tests Saber 11 of 2012 is in the English score of the student body, the socioeconomic index quartiles were used for this, classifying the first, second, third and fourth quartile as a low level, low medium, high medium and high of index respectively.

Table 6: *Students score by quartiles of INSE. Source: own elaboration.*

INSE level	Mathematics	Physics	Chemistry	Language	Social Sciences	English.
Low	42.74	42.91	44.15	43.90	42.13	40.12
Low Medium	44.37	43.59	45.00	45.40	43.48	41.70
High	53.26	48.93	50.15	51.81	49.99	54.50

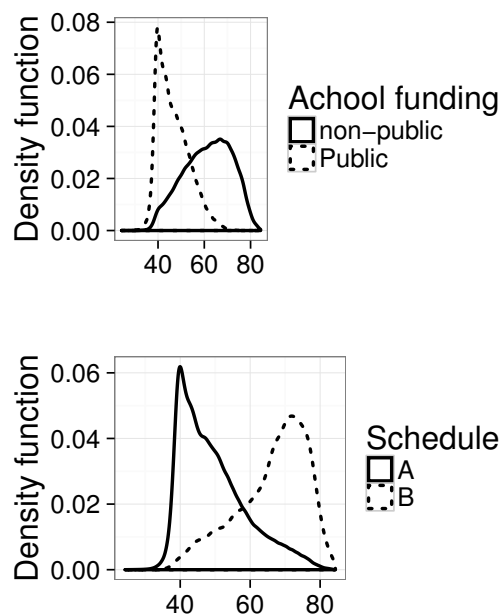


Figure 3: *Index by school's nature and calendar. Source: own elaboration.*

As you can see in Figure 4, socioeconomic status of a student has a strong relationship with obtained scores in the English, mathematics, critical Reading and sciences. However, the bigger difference is in the English test, that is scored and classified according to the Common European Framework, where the best score is for students classified in a B+ level and the worst for students classified in an A-level. Figure 4 shows the distribution of index by level of English.

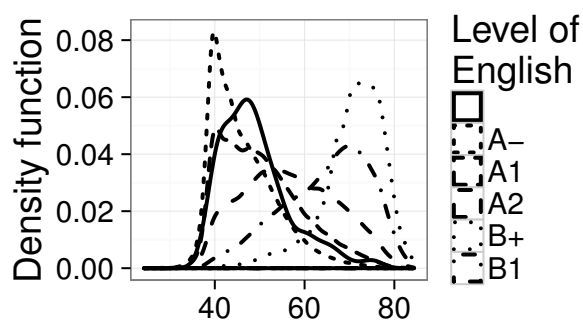


Figure 4: *Index by English level classification. Source: own elaboration.*

4 Conclusions

The construction of a family socioeconomic index is a necessary and useful exercise; that's why it must be used and understood as a key player in the development of studies and analysis of education in Colombia. Educational achievement of students in standardized tests Saber 11 has a high correlation with socioeconomic index developed, which in turn lets us understand the importance of each of the factors used in the construction of this index.

Despite showing differences in all components of the test Saber 11 among low, low medium, high medium and high index levels, the score obtained by students classified in the high index level is the highest with 14.38 points of difference between the average score of the high level and the low index level. This score is 2.3 times the difference in the chemistry test in which scores are those with a smaller difference.

Received: February 14, 2016

Accepted: April 20, 2016

References

- Bowles, S. & Gintis, G. (1997), 'Recasting Egalitarianism', *University of Massachusetts*.
- Center for Education Statistic., N. (2012), 'Improving the Measurment of Socioeconomic Status for the National Assesment of Educational Progress. A theoretical foundation.', *Child development*.

- De Leeuw, J. & Mair, P. (2009), ‘Gifi Methods for Optimal Scaling in R: The Package *homals*’, *Journal of Statistical Software* **31**(4), 1–20.
- Flores, C. E., Espinosa, F., Sánchez, L. M. & Angulo, R. (2008), ‘Diseño del Índice Sisben en su tercera Version: SISBEN III.’, *Departamento Nacional de Planeación-DNP Resúmen Ejecutivo*.
- Gil, F. J. (2013), ‘Medición del nivel socioeconómico familiar en el alumnado de educación primaria.’, *Revista de Educación* **362**.
- Hauser, R. M. (1994), ‘Measuring Socioeconomic Status in Studies of Child Development’, *Child Development* **65**(6), 1541–1545.
- Horn, J. L. (1965), ‘A rationale and test for the number of factors in factor analysis’, *Psychometrika*. **30**, 179–185.
- Mueller, W. C. & Parcel, L. T. (1981), ‘Measures of socioeconomic status: Alternatives and recommendations.’, *Child Development* **52**(1), 13–30.
- Pérez, C. (2000), ‘¿Equidad en la educación?’, *Revista Iberoamericana de Educación* **23**.
- 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.
- Ravela, P. (2005), Estudio de los factores institucionales y pedagógicos que inciden en los aprendizajes en escuelas primarias en Uruguay, Technical report, ANEP/MECAEP/UMRE.
- Sánchez, A. & Otero, A. (2012), ‘Educación y reproducción de la desigualdad en Colombia.’, *Reportes del Emisor: Banco de la República de Colombia*.
- UNESCO (2008), ‘Conferencia internacional de educación: La educación inclusiva el camino hacia el futuro.’.