Propósitos y Representaciones
ISSN 2307-7999
e-ISSN 2310-4635

Dec. 2019, Vol. 7, N° SPE, e347
Special issue: Psychological Evaluation: Perspectives and Challenges of Educational Psychology
http://dx.doi.org/10.20511/pyr2019.v7nSPE.347

RESEARCH ARTICLES

Quality of Education Systems: Evaluation Models

Calidad de los sistemas educativos: Modelos de evaluación

Rubén Fernández-Alonso*
Consejería de Educación del Gobierno del Principado de Asturias, Oviedo, España
Universidad de Oviedo, Oviedo, España
ORCID: http://orcid.org/0000-0002-7011-0630

José Muñiz
Universidad de Oviedo, Oviedo, España
ORCID: http://orcid.org/0000-0002-2652-5361

Received 01-06-19 **Revised** 13-08-19 **Accepted** 29-10-19 **On line** 07-11-19

***Correspondence**

Email: fernandezaruben@uniovi.es

Cite as:

Fernández-Alonso, R., & Muñiz, J. (2019). Quality of Education Systems: Evaluation Models. *Propósitos y Representaciones*, *7*(SPE), e347. doi: http://dx.doi.org/10.20511/pyr2019.v7nSPE.347

© Universidad San Ignacio de Loyola, Vicerrectorado de Investigación, 2019.

This article is distributed under license CC BY-NC-ND 4.0 International (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Summary

The Psychometric Research Group of the University of Oviedo participated in different presentations and workshops organized within the framework of the II International Congress of Psychological Evaluation held in November 2018 at the San Ignacio De Loyola University (Lima, Peru). This work gathers part of those contributions. Specifically, the aim of this paper is to present the mathematical models and methodological procedures available for the analysis of data in the evaluation of educational systems. The design, execution and dissemination of the results of an evaluation program of the education system is a complex task that poses a challenge in different areas, among which data analysis stands out. These programs have two main purposes: to know and describe the level of knowledge and skills of the student population and to identify and analyze the context and process factors associated with educational outcomes. In order to fulfil both purposes, the evaluation of education systems has been provided with unique and specific methodological solutions. Three of them are presented in this paper. Two are aimed at expressing learning outcomes: plausible values and cut-off methods, while the last focuses on analyzing the relationship between school factors and outcomes.

Keywords: Evaluation of Education Systems; Item Response Theory; Plausible Values; Hierarchical-linear Models; School Effectiveness

Resumen

El grupo de Investigación Psicometría de la Universidad de Oviedo participó en diferentes ponencias y talleres organizados en el marco del II Congreso Internacional de Evaluación Psicológica celebrado en noviembre de 2018 en la Universidad San Ignacio De Loyola (Lima, Perú). El presente trabajo recoge parte de aquellas aportaciones. En concreto el objetivo de este escrito es presentar los modelos matemáticos y procedimientos metodológicos disponibles para el análisis de los datos en las evaluaciones de sistemas educativos. El diseño, ejecución y diseminación de resultados de un programa de evaluación de sistema educativo es una tarea compleja que supone un desafío en diferentes ámbitos, entre los que destaca el análisis de datos. Estos programas tienen dos grandes finalidades: conocer y describir el nivel de conocimientos y competencias de la población de estudiantes e identificar y analizar los factores de contexto y proceso asociados a los resultados educativos. Para cumplir ambas finalidades la evaluación de sistemas educativos se ha dotado de soluciones metodológicas singulares y específicas. En este escrito se presentan tres de ellas. Dos están orientadas a expresar los resultados del aprendizaje: valores plausibles y métodos de punto de corte, mientras que la última está centrada en analizar la relación entre los factores escolares y los resultados.

Palabras clave: Evaluación de sistemas educativos; Teoría de Respuesta al ítem; Valores Plausibles; Modelos jerárquico-lineales; Eficacia escolar.

Introduction

The evaluation of educational systems will be 60 years old in 2019. In June 1959, under the sponsorship of the United Nations Educational, Scientific and Cultural Organization (UNESCO), the *Twelve-Country Study* was launched, a transnational cooperation program that explored the possibility of making rigorous international comparisons of academic performance and which is considered the world's first school performance assessment study. As it corresponds to a development stage, the *12-country Pilot Study* pointed out the limitations and challenges faced by the evaluation of educational systems (translation and cultural adaptation of tests, logistics of application, comparability of results...), but it also pointed out that, under certain conditions, comparison was possible (Foshay, Thorndike, Hotyat, Pidgeon& Walker, 1962).

Since then, the evaluation of educational systems has grown and become widespread. The first national assessment, *The National Assessment of Educational Progress* (NAEP), was organized in 1969 by the U.S. Department of Education. In Latin America, most countries began to evaluate their education systems in the 1990s, although in some cases (e.g., Chile, Mexico, or Costa Rica) the beginning is earlier (Woitschach, 2018). Several worldwide studies are currently underway: *Trends in International Mathematics and Science Study* (TIMSS), *Programme for International Student Assessment* (PISA), *Progress in International Reading Literacy Study* (PIRLS), *International Civic and Citizenship Education Study* (ICCS), and *International Computer and Information Literacy Study* (ICILS). International regional assessments are also available. In Latin America and the Caribbean, the Latin American Laboratory for Assessment of the Quality of Education (LLECE) is noteworthy, although it is possible to cite examples from each continent. In Africa, *Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SAMEQ) and *Programme for the Analysis of Education Systems* (PASEC); in Asia, *Southeast Asia Primary Learning Metrics* (SEA-PLM); and in Oceania, *Pacific Islands Literacy and Numeracy Assessment* (PILNA).

Despite their diversity, all of them pursue two purposes: to know and describe the level of knowledge and skills of the student population, whether at a specific moment or throughout schooling; and to identify and analyze the context and process factors associated with educational outcomes (Fernández-Alonso, 2004). The objective of this paper is to present the main methodological and analytical solutions for this dual purpose. Therefore, the work is organized into two sections: the first will show mathematical models and methods for estimating and describing learning outcomes, and the second will recreate models for analyzing factors associated with school performance.

Quality of education systems: Evaluation models

The evaluation of the education system has two ways of reporting student results. On the one hand, the average scores (*scale scores*) aggregated at the population level, of strata or other variables of interest (demographic, type of center, etc.). These are synthetic scores that allow comparisons between groups and, therefore, tend to have a media impact. The results are also presented as *achievement levels*, which are performance standards that describe the knowledge and skills of the population.

Results expressed as average scores

In the early days, education system assessment programs expressed student cognitive outcomes using the fundamentals of Classical Test Theory (Foshay et al., 1962). However, in order to maintain an adequate validity of content, these evaluations handle a large number of items, which forces them to be distributed in different booklet models following the principles of experimental design (Adams & Wu, 2002; Allen, Carlson &Zelenak, 1999; Allen, Donoghue &Schoeps, 2001; Beaton, 1987; Fernández-Alonso &Muñiz, 2011; Frey, Hartig& Rupp, 2009; Mullis, Martin, Kennedy, Trong& Sainsbury 2009; Olson, Martin & Mullis, 2008). When the test is applied, each student only responds to one booklet model, that is, he or she is confronted with a subsample of the entire bench, with the aggravating circumstance that the books are far from being perfectly parallel (Lord, 1962). In this context, classical models are inappropriate for reporting student results (Muñiz, 1997, 2018). These limitations in the equation andcomparability of results were not solved until the last quarter of the 20th century, when NAEP first employed mathematical models derived from Item Response Theory (TRI, Beaton, 1987; Bock, Mislevy& Woodson, 1982; Messick, Beaton & Lord, 1983) which, since then, has been the dominant approach for expressing cognitive outcomes in education system assessments.

TRI models are logistic functions that estimate the competence or ability of students in the variable evaluated based on their responses to a set of items and the parameters or metric properties of those items. Its mathematical formulation is the following (Mazzeo, 2018):

(1) $$p(u_p; \beta|\Theta)$$

Where, $\Theta \equiv (\Theta_1, \Theta_2, \dots, \Theta_m)$ is the vector of the student's competence or ability p conditioned by his/her vector or pattern of responses to the test items$u_p \equiv (u_{p1}, u_{p2}, \dots, u_{pn})'$and by the vector of the parameters of the items $\beta = (\beta_1, \beta_2, \dots, \beta_i)'$. The number of item parameters determines the TRI model used in each study. For example, LLECE, SACMEQ, PILNA, ICCS and PISA (in this case until 2012) combine the Rasch model for dichotomous items and the partial credit model for politomous items (Adams & Wu, 2002; Hungi, 2011; Martin & Kelly, 1997; Pacific Community, 2016; Schulz, Carstens, Losito&Fraillon, 2018; UNESCO-Regional Bureau of Education for Latin America and the Caribbean [UNESCO-OREALC], 2016a). From 2015 onwards, PISA combines the Birnbaum model (2-parameters) for binary items and Muraki's generalized partial credit model for items with three or more categories (Organisation for Economic Cooperation and Development [OECD], 2017). NAEP, TIMSS and PIRLS as well combine three models according to the format of the items: 3-parameters for multiple-choice items, 2-parameters for binary open items and Muraki's model for polytomous items (Martin, Mullis & Hooper, 2016, 2017; National Center for Education Statistics [NCES], 2018).

The TRI models have undoubted advantages over the classic approach (Muñiz, 2018). In return, they are less intuitive since the scale of scores(Θ) is indeterminate, it moves between infinite extremes. In order to solve the indeterminacy the results are offered in transformed scores. The best known expresses the results on a normal scale with mean 500 points and standard deviation 100 [N(500,100)] (Hungi, 2011; Martin et al., 2016, 2017; OECD, 2017), although other values are possible (NCES, 2018; Pacific Community, 2016; UNESCO-OREALC, 2016a; UNESCO-OREALC, & LLECE, 2016a).

Traditionally, psych educational assessment calculates function (1) using weighted maximum likelihood point estimators or Bayesian procedures (Muñiz, 1997, 2018). However, in the evaluation of education systems, individual scores are of no interest. Mazzeo (2018) calls these studies *group-score assessement* to emphasize that their objective is to estimate and compare population parameters (e.g., country averages) and not to evaluate individual performances as it occurs in most educational and psychological research. In addition, it has been demonstrated that point estimators present biases when recovering population parameters (Beaton, 1987; Mislevy, Beaton, Kaplan & Sheehan, 1992; von Davier, Gonzalez &Mislevy, 2009). Therefore, the evaluation of educational systems has developed a unique and specific procedure to report cognitive results: plausible values (PV).

A *PV* can be defined as *a random sample taken from a posteriori multivariate density function that contains the distribution of a student's probabilities of obtaining a subject score evaluated based on his/her responses to a parameterized bank of items and their socio-demographic and personal characteristics.* Mathematically the model is expressed in the following way (Mazzeo, 2018):

(2) $$f(\Theta|u_p, x_p) \propto p(u_p; \hat{\beta}|\Theta)\emptyset(\Theta; \Gamma' x_p, \Sigma)$$

The term$f(\Theta|u_p, x_p)$ represents the a posteriori density function of the student's level of competence (θ)conditioned by their responses to the items $(u_p)$and their socio-demographic and personal characteristics (x_p). This density function collects the distribution of probable student scores and is the product of two probability distributions. On the one hand, an IRT model seen in (1) that estimates the level of competence of students conditioned by their responses to some items of known parameters $[p(u_p; \hat{\beta}|\Theta)]$and, on the other hand, a model of population structure $[\emptyset(\Theta; \Gamma' x_p, \Sigma)]$where$\Gamma'$is the matrix of regression coefficients of the socio-demographic population variables on the results and $\Sigma$is the matrix of variance-covariance of the characteristics of the population. Therefore, the second term of the product is a function of continuous density

that estimates the probability that a student has a certain level of competence conditioned by his socio-demographic and personal characteristics, the effect that these characteristics have on performance at the population level and the relationship that exists between the variables used to define these characteristics. Socio-demographic characteristics are understood as variables such as gender, age, socioeconomic and cultural level of the student, average school results, as well as other factors extracted from the analysis of main components of the responses to the context questionnaires (Martin et al., 2016, 2017; Mazzeo, 2018; OECD, 2017; NCES, 2018; UNESCO-OREALC, 2016a).

The PV estimation procedure is carried out in two phases. The first is similar to the adjustment of an ordinary item bank: an equal unweighted number of cases is selected for all groups (e.g., same N of students for all countries or strata) that functions as a calibration sample. From the response vector, the parameters of the items are calculated using some point estimation procedure. In the second phase we work with all the cases and their corresponding sample weights to estimate the density function a posteriori from which the PV will be extracted randomly. In this phase, the matrix includes the vector of the students' responses to the items and all the information about their socio-demographic characteristics and the factors extracted from the analysis of main components. The parameters of the items of the first phase are set as prior information for all groups and the socio-demographic variables function as covariates in a multiple regression model. The estimation of the multivariate density function is done separately for group (country or sample stratum) so that item parameters remain constant in all countries, and are complemented by the specific effect of covariates on scores in each country or stratum. The description of the logic and fundamentals of PV can be found in Mazzeo (2018), NCES (2018) and von Davier et al. (2009) and details for execution in Wu, Adams, Wilson and Haldane (2007).

From the estimated density function for each student, as just described, a certain number of VPs are randomly taken, between 5 and 20, which are likely student scores (OECD, 2017; Martin et al., 2016, 2017; NCES, 2018). Figure 1 shows the density functions of two students who responded to the same items and whose socio-demographic and personal characteristics are similar. Student 2 matched more items than student 1 and therefore his density function is located more to the right on the N scale (500,100). However, the probable values of each student are very broad. In this example, 5 VP are randomly extracted for each student. Note that, in general, student 2's likely values are higher than student 1's. However, student 2's VP-2 (about 480 points) is lower than student 1's VP-4. (around 520). It is for this reason that PVs, unlike point IRT estimators, cannot be used to report individual results and are only used in the evaluation of education systems to describe population parameters (Mazzeo, 2018; NCES, 2018).

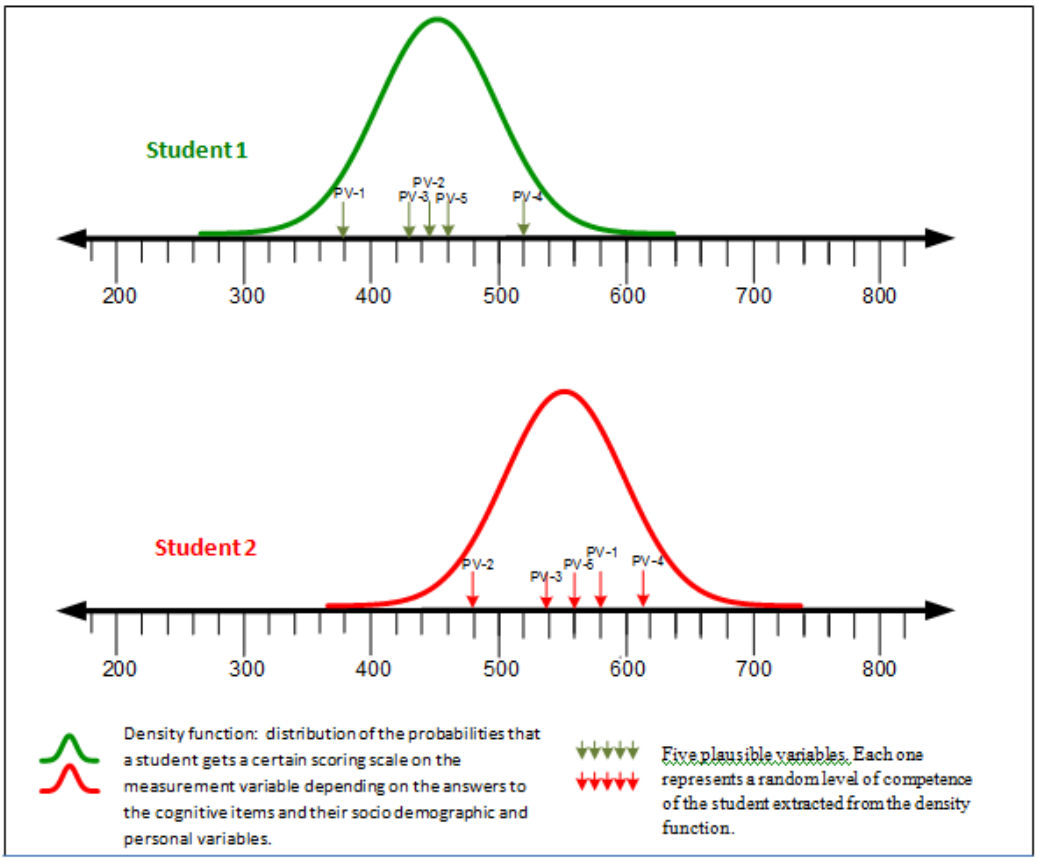

Figure 1. The logic of assigning scores: a posteriori density function$\left[f\left(\Theta\middle|u_p,x_p\right)\right]$ and selection of plausible values for two students.

Finally, the score of each group (country, jurisdiction, stratum...) is expressed as the average of the VP. However, the standard error of the mean is not calculated as the ratio between the standard deviation and the square root of the number of cases since students are not selected by a simple random procedure, but by cluster sampling in two or more stages. This requires re-sampling methods to calculate typical estimator errors. The details of this calculation can be found in OECD (2009).

Results expressed as qualitative scales: performance levels

The synthetic scoring scales described above summarize population parameters and make it possible to locate and compare group performances. However, the numerical score does not provide information on the knowledge and skills actually attained by the students (Fernández-Alonso, 2004). To answer these types of questions, a methodology is used, called performance levels, whose purpose is to establish cut-off points on the continuous scale and to analyze the knowledge, skills and abilities that shows the group of students above a certain level or cut-off point (Kelly, Mullis & Martin, 2000; Martin et al., 2016, 2017; UNESCO-OREALC, 2016a).

It is an arbitrary procedure, but at the same time very practical and efficient. Its logic is similar to the textile industry's use of anthropometric measures (OECD, 2017; Educational Evaluation Service of the Principality of Asturias, 2018a). Physical traits, like the results of a cognitive test, are very variable and are expressed in numerical and continuous scales. For example, the width of the hips of men normally oscillates between 65 and 150 centimeters, that is, in a range of 85 centimeters. However, textile producers collapse or group this range into a few categories: size S (between 78 and 85 cm.); size M (between 86 and 94 cm.) and so on. Apart from distances, determining cut points follows the same idea: to arbitrate limits or intervals on a continuous scale

to group scores into a few levels of performance. The procedure is executed in two major phases (Educational Evaluation Service of the Principality of Asturias, 2018a, 2018b):

- Determine cut-off points on the outcome scale to establish performance groups or performance levels and assign items to those levels.
- Prepare descriptions that summarize student competencies at each of the performance levels.

Determine cut-off points and assign items to performance levels. There are different procedures for establishing cut-off points (Muñiz, 2018). In its first editions, TIMSS indicated the cut-off points a priori on the percentile scale (Kelly et al., 2000). At present, both TIMSS and PIRLS set four cuts on the N(500,100) scale: 400, 475, 550 and 625 points, creating as many groups with students who obtain a score of ±5 points on these marks (Martin and Mullis, 2012; Martin et al., 2016, 2017). Thus, for example, the Low Level group is made up of students who scored between 395 and 405 points (Figure 2).

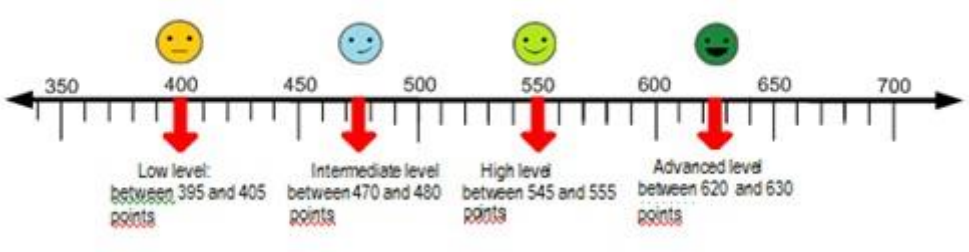

Figure 2. Cut-off points and performance level groups in TIMSS and PIRLS
Source: Educational Evaluation Service of the Principality of Asturias

The following compares the success rate of the four groups in each of the items. There are several criteria for assigning items to achievement levels, although they are all based on the same principle: an item will be assigned to an achievement level when the majority of students at that level (at least 60%) respond correctly to the item while the majority of students at the next lower level fail it (less than 50% of right answers). For example, an item will be assigned to the High Level if it is successful by at least 60% of the students in that group and more than half of the students in the Intermediate Level fail it.

Taking *a priori* scores on the N(500,100) scale is not the only way to establish cut-off points. Other studies such as PISA, LLECE or PILNA determine their cut-off points by consensus of a group of experts on the characteristics of the items (Hungi et al., 2010; OECD, 2017; UNESCO-OREALC, 2016a). In this case a panel of experts works with the items ordered by their level of difficulty and collegially agree to indicate the cut-off points. At this point the most critical marks are upper and lower limits. Determining the lower limit means identifying the items that ask for basic and elementary aspects, so that students who do not respond correctly to them will become part of the group of lesser competence. In order to establish the upper limit, it is necessary to isolate the items with greater complexity, those that can only be resolved by advanced or excellent students. Once the extreme marks are delimited, the range of points between the two is divided equally into as many points as groups are necessary.

Figure 3 exemplifies the procedure for establishing 6 performance levels with a 20 item test (Educational Evaluation Service of the Principality of Asturias, 2018a). The central part of the graph shows together the distribution of results [N(500,100)] and the items ordered by their difficulty. Since the student score and the difficulty of the items are on the same scale, it is possible to predict, for example, that the student body that obtains 600 points has a high probability of guessing the 15 items whose difficulty is below this mark (all those between item 05 and item

18). Likewise, it is also more likely that this student body will fail the 5 items whose degree of difficulty is above 600 points (ít09, 07, 20, 17 and 19).

In this example, the group of experts agreed that the basic items were 12, 13 and 18. Of the three, the graph indicates that item 13 is the most difficult and, therefore, the lower limit is located immediately above the difficulty of that item, in this case 360 points. It will then be said that the probability of hitting a basic item by students who obtain less than 360 points is less than pure chance ($p < 0.50$). In the case of the upper limit, the panel of experts agreed that only the most competent student body will guess items 17 and 19. Between them, the ít17 is easier and, therefore, the upper limit is established below the difficulty level of this item (680 points). In this way it is predicted that a student who obtains more than 680 points will have a higher probability at random ($p > 0.50$) of guessing a very complex item. Therefore, 680 points marks the difference between the advanced or excellent student body and the rest of the students evaluated. As this example sought to establish six levels of performance, the space between the lower and upper marks is divided into four parts of 80 points each. The items that fall within each quadrant are assigned to their respective performance level.

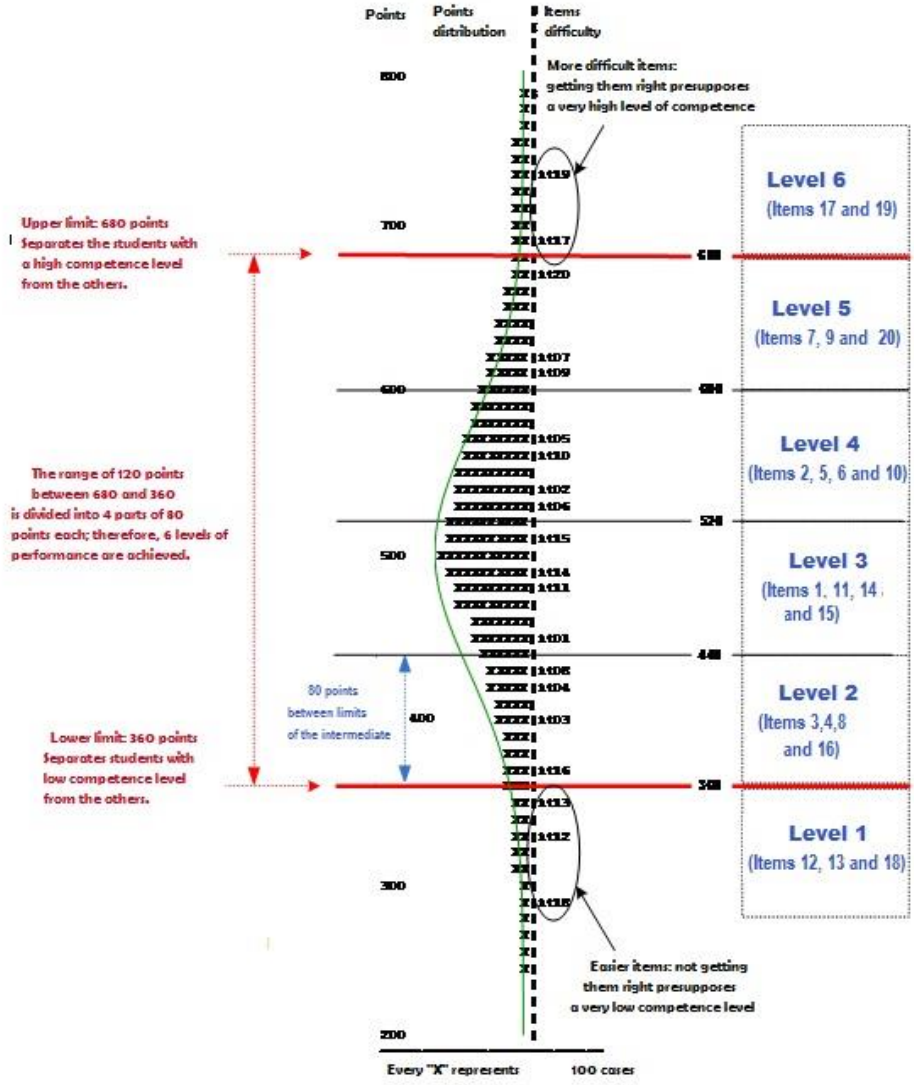

Figure 3. The logic of assigning items to performance levels in expert agreement methods
Source: Educational Evaluation Service of the Principality of Asturias

Develop descriptions that summarize competencies. With the items sorted by level of difficulty or assigned to performance levels, a panel of experts in the subject and evaluated course analyzes

their content, taking on three tasks (Nungi et al., 2010; OECD, 2017; UNESCO-OREALC, 2016a):

- Develop short descriptions that define the specific skills and abilities needed to respond correctly to each of the items. Generally these descriptions are very punctual and concrete since they refer to the knowledge and cognitive processes that are put into play at the moment of responding to certain items.
- The set of items at each level makes up the range of competences, knowledge and skills of the students at that level. Therefore, the second task consists of writing a general description that summarizes and characterizes each of the levels of performance.
- Select a group of items that exemplify the competencies of each group or performance level. These items will be released and published in the corresponding results reports.

The descriptions contained in the scales of competence described have four characteristics (Educational Evaluation Service of the Principality of Asturias, 2018a):

- They are hierarchical and inclusive: the model predicts that students at a given level have a high probability of responding correctly to items at lower levels.
- They are probabilistic, not deterministic: it cannot be concluded that all students at one level will respond correctly to the same items or that, because they belong to the same level, their real competence in the subject is identical.
- They are empirical: the descriptions contained in the levels of achievement come primarily from responses to specific items and, therefore, represent effective student achievement.
- They have the potential to guide educational practice: because of the way in which levels are constructed, they have the potential to predict future learning that the student is able to approach succesfully.

How are factors associated with educational outcomes identified and analyzed?

The evaluation of the education system has to offer educational policy guidelines for school improvement (Fernández-Alonso, 2004; UNESCO-OREALC, 2016b). This implies identifying and studying the elements linked to educational outcomes. In order to gather the necessary information to synthesize these factors, context questionnaires are applied to students, families, teachers, school principals and, occasionally, educational authorities of the participating countries. This information allows for the construction of simple variables and complex indices. The former reflect observable or documentarily verifiable facts (e.g., gender, age, etc.) and are generated through recodifications and arithmetic calculations. Complex indices summarize unobservable facts or latent variables (e.g., personal attitudes and beliefs, classroom climate, pedagogical leadership, etc.) and are constructed through confirmatory factor analyses or TRI models (Adams & Wu, 2002; OECD, 2017; UNESCO-OREALC, 2016a).

Prior to the construction of the indices and in order to organize the analysis, theoretical frameworks based on the systemic approach are developed (Adams & Wu, 2002; Mullis et al., 2002, 2009; Servicio de Ordenación Académica, Formación del Profesorado y Tecnologías Educativas del Principado de Asturias (Academic Planning, Teacher Training and Educational Technologies Service of the Principality of Asturias, 2011). Figure 4 exemplifies a theoretical framework that functions as a matrix of double-entry specifications for selecting and locating the variables and indices considered in the analysis.

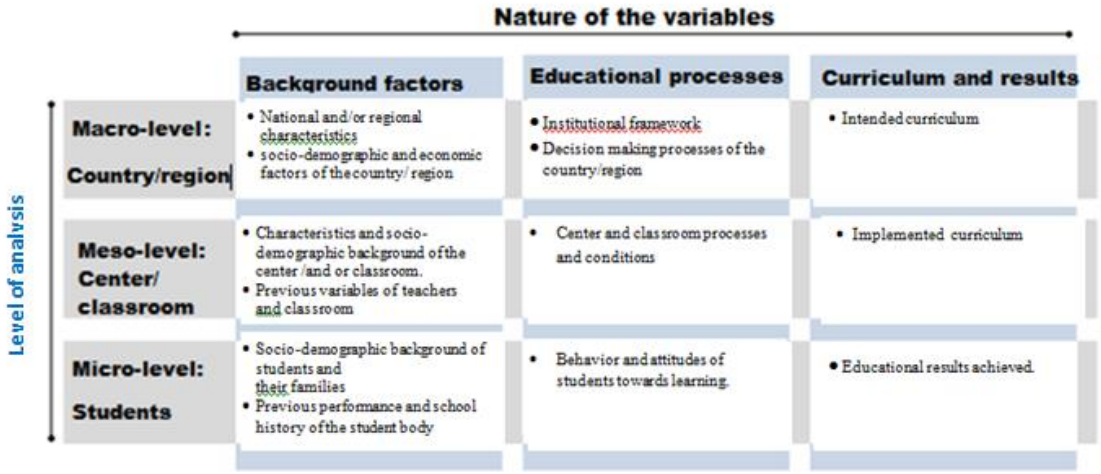

Figure 4. Theoretical framework for a study of associated factors

The coordinate axis distinguishes three types of variables according to their nature: background and socio-demographic context factors that, by definition, are stable and not very permeable to educational action; process factors that, due to their moldable nature, have greater potential and capacity for school improvement; and educational results, understood in a broad sense since they include cognitive, affective results and other desirable products such as user satisfaction with the educational service (Muñoz-Repiso et al., 1995; Murillo, 2003). The second axis of the table indicates that the data present a hierarchical or multilevel structure (Scheerens&Bosker, 1997; Scheerens, 2016): students (micro-level or Level 1) are educated in classrooms, these form centers (meso-level or Level 2) and these are located in geographic areas within the same educational system (macro-level or Level 3).

The analysis of associated factors must be coherent with the theoretical framework, an issue that is reflected both in the type of mathematical models used and in the strategy of adjustment and comparison of these models. Both aspects are developed below.

Mathematical models in the analysis of associated factors.

In a multilevel structure, students who share higher order hierarchical groupings (e.g., classroom groups) tend to be more similar to each other and their performances more homogeneous than those who do not share such groupings. In this context, the assumption of independence of observations that underpins the analytical solutions of the general linear model cannot be maintained. In fact, classic multiple regression models present important limitations for analyzing nested data, since they underestimate measurement errors when they do not contemplate hierarchical structures of a higher order, or they destroy the internal differences of groups when they eliminate hierarchical structures of a lower order (Openshaw, 1982; Robinson, 1950).

Three decades ago, the first works were published with multilevel models (Paterson & Goldstein, 1991), also known as linear hierarchical models (Raudenbush&Bryk, 2002) or random coefficients (Longford, 1993). Together they make up a family of mathematical models developed specifically to analyze data of a complex nature. At present their use is widespread because they are very versatile; they can be implemented on criteria variables measured at any scale: continuous, ordinal, discrete or binary; and they allow modulating data in longitudinal or growth designs, repeated measurements, experimental studies with control group and cross-classification structures, to name but a few the most recurrent applications in educational research (Hox, 1998; Raudenbush&Bryk, 2002).

In the evaluation of education systems, the data respond to the structure of a nested design, in which the cases (students, Level 1) are grouped into broader units of information (classrooms,

centers..., Level 2) and these in turn into higher order structures (population strata, regions, countries..., Level 3). In this type of design, the most commonly used multilevel models are variance analysis of a random effects factor, regression analysis with averages as results, covariance analysis of a random effects factor, regression analysis with random coefficients, and regression analysis with averages and slopes as results (Gaviria Soto & Castro Morera, 2005; Pardo, Ruiz & San Martín, 2007; Raudenbush&Bryk, 2002).

A hierarchical-linear model can be understood as a classical regression model with regressors at different levels and, therefore, the classical regression model is a good starting point for understanding the logic of hierarchical-linear analysis (Gaviria Soto & Castro Morera, 2005). Let us suppose that we want to predict the result of a student in a test based on his score in a socioeconomic and cultural index (ISEC). The simple regression model indicates that the student's true score (y_i) will be an additive model of three terms. Two of them fixed and common to all cases: the intercept (β_0) which is the expected score for students whose ISEC(X_i) is equal to the mean($X_i - \bar{X}_{\cdot} = 0$); and the slope ($\beta_1$) which is the expected gain (or loss) in the result for each unit that increases (or decreases) the student's ISEC ($X_i - \bar{X}_{\cdot} \neq 0$). The third term is the estimation error (ε_i) which is assumed to be random (Raudenbush&Bryk, 2002).

(3) $$y_i = \beta_0 + \beta_1(X_i - \bar{X}_{\cdot}) + \varepsilon_i$$

All the same, a hierarchical design assumes that intercepts and slopes are not fixed, but that each school has its own. The intercepts vary because the centers obtain different average scores for students of the same ISEC, and the slopes also vary because the differences between students of low and high ISEC are greater in some centers than in others. These variations in slopes and intercepts make it necessary to specify a multilevel model that introduces new terms into the random part of the equation. For this example, where only the ISEC of each student is included (level 1 measure) and there are no predictors in level 2, a regression model is specified with random coefficients (intercepts and slopes), which is mathematically defined as (Raudenbush&Bryk, 2002):

(4) Level 1: $$y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{\cdot j}) + \varepsilon_{ij}$$

Level 2: $$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

And that in its compact form it looks like this:

(5) $$y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\cdot j}) + [\mu_{0j} + \mu_{1j}(X_{ij} - \bar{X}_{\cdot j}) + \varepsilon_{ij}]$$

In this case the prediction of the result of student *i* in school $j(y_{ij})$ has a fixed part and a random part, the latter contained in the bracket. As in the classical model the fixed part contains two terms, γ_{00} y $\gamma_{10}(X_{ij} - \bar{X}_{\cdot j})$, whose meaning is similar to that of β_0 and $\beta_1(X_{ij} - \bar{X}_{\cdot j})$ in equation (3). However, now the random part is more complex. In addition to the estimation error associated with the student (ε_{ij} , two new variation terms are included: one associated with the fact that centers have different intercepts(μ_{0j}) and a second variation because the effect of ISEC on performance $[\mu_{1j}(X_{ij} - \bar{X}_{\cdot j})]$ is different in each school.

Figure 5 graphically represents the terms of these equations with a fictitious but very plausible example. The ordinate axis collects the test scores on the N scale *(500,100)* and the coordinate axis the ISEC scores on an *N* scale *(0,1)*. The central green line is the regression line that summarizes the effect of ISEC on the results in the whole population (all centers and students): the general intercept (γ_{00}) equals 500 points and the line has a slope (γ_{10}) of 31 degrees, which according to the scale in this graph assumes that for each unit that increases ISEC 15 points of gain are predicted in the test result.

In the figure we have also selected two schools that have different intercepts and slopes and we have pointed out a case, which we will call student 7 of school 2 (y_{72}), who achieved 565 points in the test and whose ISEC is equal to 1 point. According to equation (3) the student y_{72} would have obtained 50 points above the expected value depending on his individual ISEC, since:

$$y_{72} = \beta_0 + \beta_1(X_i - \bar{X}_.) + \varepsilon_i =$$
$$565 = 500 + 15(1 - 0) + \varepsilon_i =$$
$$\varepsilon_i = 565 - 515 = 50$$

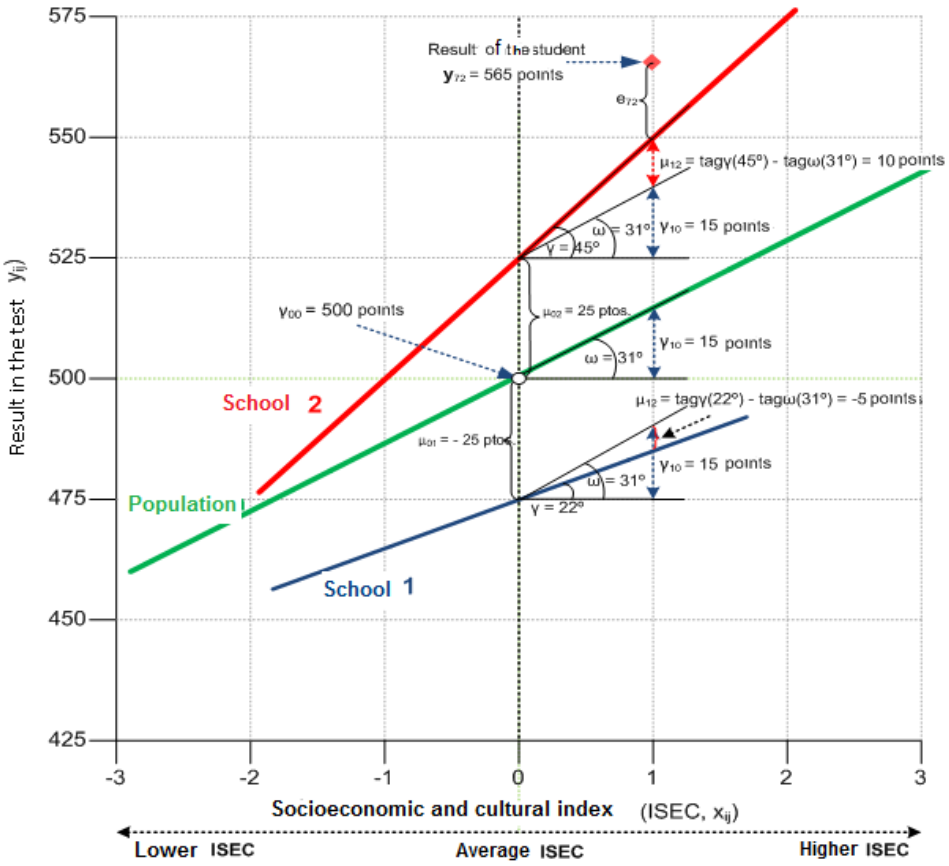

Figure 5. Representation in the plane of the terms of a multilevel equation

However, the student y_{72} attends School 2 where the intercept and slope summarizing the relationship between the results and the ISEC are different from the general parameter and also from that of other schools. Therefore, the student outcome y_{72} can be explained not only by his ISEC, but also by the school he attends. In the first place, the intercept of School 2 stands at 525 points, that is, student y_{72} *is* enrolled in a center whose average is 25 points higher than the average population ($\mu_{02} = 25$). Note that in the case of School 1 the situation is the opposite: $\mu_{01} = -25$. In addition, in School 2 the regression slope (μ_{12}) has an inclination of 45º, that is to say, the gain in the result for each unit that increases the ISEC is greater than the 15 points predicted by the population model (γ_{10}). Precisely, for the scalar values in this graph: $\mu_{12} = tag(45º) - tag(31º) = (1 \times 25) - (0{,}6 \times 25) = 1$. Therefore, substituting the values of (5):

$$y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{.j}) + [\mu_{0j} + \mu_{1j}(X_{ij} - \bar{X}_{.j}) + \varepsilon_{ij}] =$$
$$565 = 500 + 15(1 - 0) + [25 + 10(1 - 0) + \varepsilon_{ij}] =$$
$$\varepsilon_{ij} = 565 - (515 + 25 + 10) = 15$$

The example shows that of the 50 points of error that the simple regression model imputes to the student, 25 points are explained because the student attends a school with good results and 10 additional points because in his school the effect of ISEC on the results is greater than the estimated for the whole population, so that finally the error related to the student is reduced to 15 points. This illustrates one of the advantages of hierarchical-linear models over classical regression: they allow the variance of results to be identified and broken down into different levels: individual, center/classroom and education system. In general, analyses indicate that higher order structures accumulate less variance than lower levels, which is totally compatible with the educational reality: school performance has an important component of motivation and individual effort, so it is expected that individual factors explain a large part of the differences. Similarly, it is very plausible that classroom variables (orderly climate, teaching methodology, etc.) have a much greater impact on student outcomes than factors in the education system whose effect on results is always more indirect (Woitschach, Fernández-Alonso, Martínez-Arias & Muñiz, 2017).

On the other hand, the use of hierarchical-linear models is not only recommended because it relates data from complex patterns of variability. It also happens that many educational phenomena are multilevel in nature, that is, the same measure can have different meanings and present different effects depending on the level of analysis at which it is considered. Homework is an example of this type of variable (Trautwein, 2007). Assume questions such as: how often do you do your homework or how long does it take you to do it? Analyzed at the individual level, the measures reflect the student's work habit or dedication. However, if the responses are averaged per classroom, the measure has a different meaning because it describes the teacher's homework policy, i.e., the frequency or amount of homework assigned. In addition, the effects on performance differ according to the level of analysis: in general it has been found that the effect of homework time at the individual level is negative or, at best, not significant, while the frequency or size of homework tends to be positive and significantly associated with the results (Fernández-Alonso, Álvarez-Díaz, Suárez-Álvarez&Muñiz, 2017; Fernández-Alonso, Suárez-Álvarez&Muñiz, 2015, 2016; Fernández-Alonso et al, 2019; Trautwein, 2007).

Adjustment and comparison of models to analyze associated factors

The adjustment strategy of a hierarchical-linear analysis begins by specifying very simple models to which variables are added, while maintaining those significant factors in the previous models. This makes it possible to compare the increase in the percentage of variance explained by the successive models and the improvement experienced by the adjustment parameters with the introduction of new variables. The specification of the models must be coherent with the theoretical framework of the study (see figure 4) and the strategy is very flexible and allows for the establishment of different models depending on the objectives of the study and the variables of interest. However, in the analysis of associated factors there are three basic models that, in one way or another, are usually included in all studies.

Null model. The multi-level strategy begins with a model without predictors. It is, therefore, an analysis of variance of a random effects factor that is known as a null or empty model and covers three purposes: it estimates the magnitude of the total variance and how it is distributed among the different levels of aggregation; it serves as the basis for comparing the adjustment and improvement of the explanatory capacity of the rest of the models; and it makes it possible to estimate the center effect, that is, the proportion of the differences in the result that are attributable to the educational action of the schools. The estimation of the center effect is the earliest and oldest line of school effectiveness (Scheerens, 2016; Scheerens&Bosker, 1997; Scheerens, Witziers& Steen, 2013; Teddlie& Reynolds, 2000; Towsend, 2007) and in Latin America, research on school effects has been carried out for more than two decades, for which reason a wide range of works is available (among others, Casas, Gamboa&Piñeros, 2002; Cervini, 2012; Cervini, Dari & Quiroz, 2016; Murillo, 2003, Murillo &Román, 2011; UNESCO-OREALC & LLECE, 2000, 2010, 2016b). The effect of the center indicates, fundamentally, the percentage

of variations related to differences in the quality and formative offer of the centers. It is generally assumed that in more equitable education systems the size of this effect is smaller since differences in school performance tend to be smaller (Woitschach et al., 2017).

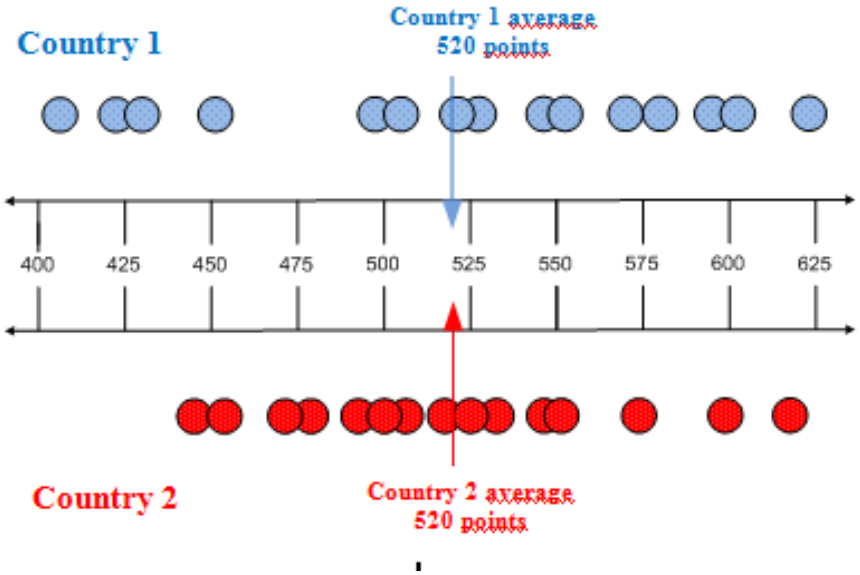

Figure 6. The logic of the center effect
Source: Educational Evaluation Service of the Principality of Asturias

Figure 6 shows the simulated results of two countries composed of 15 centers (each represented by a circle). The position of the circle indicates the average of each center on the N scale (500,100). The average of both countries is identical (520 points), but in Country 2 the differences between the centers, understood as the variance in the center averages is much smaller (approximately 50% smaller). Therefore, the size of the effect of the center in Country 2, that is, the differences or inequalities between its schools are smaller and it is concluded that, in relation to Country 1, its results are more equitable.

Adjustment model or background models. The second model includes as predictors the available information on school background and variables of the students' socio-demographic and school context. The most commonly used variables are the index that summarizes the socioeconomic and cultural level of the student body or, failing that, variables such as studies and professions of the parents, number of books in the home, material possessions or characteristics of the dwelling (Palardy, Rumberger& Butler, 2015; Peña Suárez, Fernández-Alonso & Muñiz, 2009; Sirin, 2005). Other variables widely used in adjustment models are gender, mother tongue or migrant status, and in Latin American studies it also seems important to be indigenous and to reconcile work and studies, which are generally not considered in research with developed countries (UNESCO-OREALC & LLECE, 2016b). On the other hand, the variables related to school background with the greatest effect on results are, in this order, previous performance, school repetition, and early schooling.

Background models can be specified with predictors at a single level (e.g., regressions with means as results or regressions with random coefficients). However, explanatory power is increased by including adjustment factors at all levels using covariance analysis of a factor or regression analysis with means and slopes as results. Therefore, it is highly recommended that background models include variables and factors at all hierarchical levels of analysis (Scheerens, 2016).

The results of the background variables model can be interpreted in terms of educational inequality: the higher the percentage of variance explained by the predictors included in this model, the greater the determination of the results by background factors and, therefore, the

greater the level of inequality. Figure 7 compares the "effect of student socioeconomic and cultural status (ISEC) on student scores" in two countries with identical test scores (500 points) and socioeconomic and cultural status (ISEC = 0 points). In Country 1, for each point that ISEC increases, 25 points of gain are predicted in the test, while in Country 2, 10 points of gain are predicted in the test for each point that ISEC increases. Therefore, it is concluded that Country 2 seems more equitable since the results of its students are less determined by socioeconomic and cultural background than in the case of Country 1.

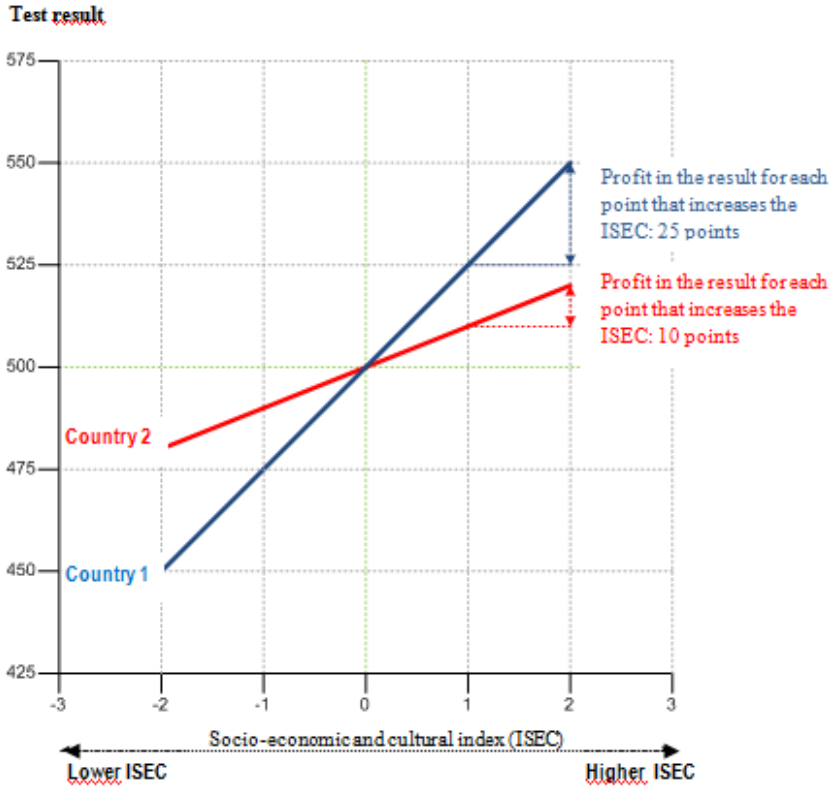

Figure 7. The Logic of the Effect of Sociological Background on Outcomes
Source: Educational Evaluation Service of the Principality of Asturias

Models of educational processes and variables of interest. Once a model with the background variables is available, the next step in the strategy is to add to it the variables and factors that describe the educational processes. As already pointed out, the background model discounts the variance imputable to background and demographic factors. Therefore, the variance explained by the process models can be interpreted as a net and uncontaminated effect. In other words, the processes that appear statistically significant will be so after discounting or neutralizing the effect of the antecedents and, therefore, it can be ruled out that the results of the model are affected by alternative hypotheses relative to the characteristics of social and demographic context.

The specifications of the process models can be very varied. The most analytical strategy is to introduce one by one all the variables of interest on the fit model (UNESCO-OREALC & LLECE, 2016b). It is the most detailed and in general the most likely to show statistical significance in the associated factors. It is also possible to specify models that include all process variables at the same level of analysis. For example, introduce all the process factors measured at the student level and in this way explore the overall effect on the results of variables such as reading habits, attitudes, motivation, academic expectations of the student body, etc., and estimate what percentage of the variance is explained by the personal variables of the student body.

Another solution, probably more substantive from the theoretical point of view, is to specify a model for studying specific processes including variables measured at different levels of analysis. A possible example would be the study of classroom processes (Academic Organization Service, Teacher Training and Educational Technologies of the Principality of Asturias, 2011). In this case, the model includes variables measured at the individual level (for example, the assessment of teaching work by students), and other measures at the classroom level (work climate, time effectively devoted to learning, profiles of teaching methodology, etc.). With this type of model it is possible to analyse the interaction between variables at different levels, answering very interesting research questions such as, for example, studying the influence of a certain teaching methodology (classroom variable) on student learning with different levels of comprehension (individual variable), being able to identify teaching-learning methodologies that greatly benefit students with greater comprehension problems.

In general, the latest model tends to include all process variables at all levels (UNESCO-OREALC & LLECE, 2000, 2010). This makes it possible to estimate the predictive capacity of the whole model and to compare the effects of the set of variables and factors analysed. Those variables that continue to maintain their statistical significance in the final model can be considered the most relevant factors on which to support the conclusions of the study and guide educational policies for the improvement of the education system.

Conclusions

The evaluation of education systems poses a challenge in different areas: availability and logistics of resources; specification of theoretical frameworks; analysis of data and communication of results. In relation to data analysis, the main challenge is to respond to the two purposes of these studies: to express the results and competencies of the school population; and to identify and study the factors associated with educational outcomes that allow policy decisions to be guided for the improvement of education systems.

In order to meet the first objective, two unique procedures have been developed. The scores are expressed as plausible values and not as point estimators. In addition, numerical scores are translated into descriptions of competencies using cut-off methods that establish performance levels. For its part, the analysis of associated factors is inseparable from a robust theoretical framework, which assumes the existence of factors of diverse nature, intimately related and maintaining hierarchical relationships. In coherence with this theoretical framework, the analysis of associated factors uses hierarchical-linear models that must be adjusted according to a strategy that breaks down the variance in the different levels of aggregation and controls and discounts the part of the variations due to the antecedent factors. This is the only way to estimate a net effect of the educational and non-polluted processes, and to orient educational policies to the improvement of the system.

References

Adams, R. J., & Wu, M. L. (2002). *Technical report for the OECD Programme for International Student Assessment*. Paris: OECD Publications.

Allen, N. L., Carlson, J. E., &Zelenak, C. A. (1999). *The NAEP 1996 Technical Report*. Washington, DC: U.S. Department of Education / NCES.

Allen, N. L., Donoghue, J. R., &Schoeps, T. L. (2001).*The NAEP 1998 Technical Report*. Washington, DC: U.S. Department of Education / NCES

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 Technical Report*. Princeton, NJ: NAEP / Educational Testing Service.

Bock, R. D., Mislevy, R., & Woodson, C. (1982).The Next Stage in Educational Assessment.*Educational Researcher, 11*(3), 4-16. doi: https://doi.org/10.3102/0013189X011003004

Casas, A., Gamboa, L. F., & Piñeros, L. J. (2002). *El efecto escuela en Colombia, 1999-2000*. Colombia: Universidad del Rosario.

Cervini, R. (2012). El efecto escuela en países de América Latina: Reanalizando los datos del SERCE. *Archivos Analíticos de Políticas Educativas, 20*(39), 1-25.

Cervini, R., Dari, N., & Quiroz, S. (2016). Las determinaciones socioeconómicas sobre la distribución de los aprendizajes escolares. Los datos del TERCE. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 14*(4), 61-79. doi:http://dx.doi.org/10.15366/reice2016.14.4.003

Fernández-Alonso, R. (2004). *Evaluación del rendimiento matemático* (tesis doctoral). Universidad de Oviedo. Recuperado dehttp://hdl.handle.net/10651/16615

Fernández-Alonso, R., Álvarez-Díaz, M., Suárez-Álvarez, J., & Muñiz, J. (2017). Students' achievement and homework assignment strategies.*Frontiers in Psychology, 8*, 286.doi: http://dx.doi.org/10.3389/fpsyg.2017.00286

Fernández-Alonso, R., & Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las competencias básicas. *Aula Abierta, 39*(2), 3-34.

Fernández-Alonso, R., Suárez-Álvarez, J., & Muñiz, J. (2015). Adolescents' homework performance in mathematics and science: Personal factors and teaching practices. *Journal of Educational Psychology, 107*(4), 1075-1085. doi:http://dx.doi.org/10.1037/edu0000032

Fernández-Alonso, R., Suárez-Álvarez, J., &Muñiz, J. (2016). Homework and performance in mathematics: the role of the teacher, the family and the student's background. *Revista de Psicodidáctica, 21*(1), 5-23. doi: http://dx.doi.org/10.1387/RevPsicodidact.13939

Fernández-Alonso, R., Woitschach, P., Álvarez-Díaz, M., González-López, A. M., Cuesta, M., & Muñiz, J. (2019). Homework and academic achievement in Latin America: A multilevel approach. *Frontiers in psychology, 10*, 95. https://doi.org/10.3389/fpsyg.2019.00095

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53.doi: https://doi.org/10.1111/j.1745-3992.2009.00154.x

Gaviria Soto, J. L., & Castro Morera, M. (2005). *Modelos jerárquicos lineales*. Madrid: La Muralla.

Hox, J. J. (1998). Multinivel modeling: When and Why. En I. Balderjahn, R. Mathar y M. Schader (Eds.). *Classification, data analysis and data highways* (pp 147-154). New York: Springer.

Hungi, N. (2011). Accounting for Variations in the Quality of Primary School Education, *SACMEQ Working Paper, 7*.Recuperado de http://www.sacmeq.org/sites/default/files/sacmeq/publications/07_multivariate_final.pdf

Hungi, N., Makuwa, D., Ross, K., Saito, M., Dolata, S., van Capelle, F., Paviot, L., &Vellien, J. (2010). SACMEQ III Project Results: Pupil Achievement levels in Reading and Mathematics. *SAMEQ WorkingDocument, 1*. Recuperado de http://www.sacmeq.org/sites/default/files/sacmeq/reports/sacmeq-iii/working-documents/wd01_sacmeq_iii_results_pupil_achievement.pdf

Kelly, D. L., Mullis, I.V. S., & Martin, M. O. (2000). *Profiles of Student Achievement in Mathematics at the TIMSS International Benchmarks: U.S. Performance and Standards in an International Context*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Longford, N. T. (1993). *Random coefficient models*. New York: Oxford University Press.

Lord, F. M. (1962). Estimating norms by item-sampling.*Educational and Psychological Measurement, 22*, 259-267.doi: https://doi.org/10.1177/001316446202200202

Martin, M. O., & Kelly, D. L. (1997).*TIMSS technical report: Vol. II. Implementation and analysis: Primary and middle school years*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., & Mullis, I. V. S (Eds.) (2012). *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Recuperado de http://timssandpirls.bc.edu/methods/pdf/TP11_Context_Q_ Scales.pdf

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.) (2016). *Methods and Procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Recuperado de http://timssandpirls.bc.edu/publications/timss/2015-methods.html

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.) (2017). *Methods and Procedures in PIRLS 2016*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Recuperado de https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

Mazzeo, J. (2018). Large-scale group-score assessments.En W. J. van der Linden (Ed.) *Handbook of Item Response Theory. Vol. 3: Applications* (pp. 297-311). Boca Raton, FL: CRC Press.

Messick, S., Beaton, A. E., & Lord, F. (1983).*A new design for a new era*.Princeton. NJ: Educational Testing Service.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating Population Characteristics From Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement, 29*(2): 133-161. Recuperado de http://www.jstor.org/stable/1434599

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009).*PIRLS 2011 Assessment Framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gonzalez, E. J., Chorostowski, S. J., & O'Connor, K. M. (2002). *TIMSS assessment framewoks and specificacions 2003* (2 ed.). Chestnut Hill, MA: Boston College.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide

Muñiz, J. (2018). *Introducción a la Psicometría*. Madrid: Pirámide

Muñoz-Repiso, M., Cerdán, J., Murillo, F. J., Calzón, J., Castro, M., Egido, I., García, R., & Lucio-Villegas, M. (1995). *Calidad de la educación y eficacia de las escuelas. Estudio sobre la gestión de los recursos educativos*. Madrid: Ministerio de Educación y Ciencia.

Murillo, F. J. (2003). *La investigación sobre Eficacia Escolar en Iberoamerica. Revisión Internacional sobre el Estado del Arte*. Bogotá: CAB/CIDE.

Murillo, F. J., & Román, M. (2011). ¿La escuela o la cuna? Evidencias sobre su aportación al rendimiento de los estudiantes de América Latina. Estudio multinivel sobre la estimación de los efectos escolares. *Revista de currículum y formación del profesorado, 15*(3), 27-50.

National Center for Education Statistics (2018). *NAEP Technical Handbook: Methods and Procedures*. U.S. Department of Education: Washington, DC. Recuperado de https://nces.ed.gov/nationsreportcard/tdw/

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2008).*TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Openshaw, S. (1984).Ecological Fallacies and the Analysis of Areal Census Data. *Environment and Planning A: Economy and Space, 16*(1), 17-31.doi:https://doi.org/10.1068/a160017

OrganisationforEconomic Co-operation and Development [OECD]. (2009). *PISA Data Analysis Manual: SPSS® Users*, 2nd Ed. Paris: OECD Publishing.doi: http://dx.doi.org/10.1787/9789264056275-en

Organisation for Economic Co-operation and Development [OECD].(2017). *PISA 2015 Technical Report*. Paris: OECD Publishing. Recuperado de http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Pacific Community (2016).*Pacific Islands Literacy and Numeracy Assessment (PILNA 2015).Regional Report*. Fiji Islands: Educational Quality Assessment Program.

Palardy, G., Rumberger, R., & Butler, T. (2015).The effect of high school socioeconomic, racial, and linguistic segregation on academic performance and school behaviors.*Teachers College Record, 117*(12), 1-52.

Pardo, A., Ruiz, M. Á., & San Martín, R. (2007).Cómo ajustar e interpretar modelos multinivel con SPSS. *Psicothema, 19*(2), 308-321.

Paterson, L., & Goldstein, H. (1991) New statistical methods for analysing social structures: An introduction to multilevel models. *British EducationalResearchJournal, 17*(4), 387-393. Recuperado de http://links.jstor.org/sici?sici=0141-1926%281991%2917%3A4%3C387%3ANSMFAS%3E2.0.CO%3B2-9

Peña Suárez, E., Fernández Alonso, R., & Muñiz Fernández, J. (2009). Estimación del valor añadido de los centros educativos. *Aula abierta, 37*(1), 3-18. Recuperado de http://digibuo.uniovi.es/dspace/bitstream/10651/7869/1/AulaAbierta.2009.37.1.3-18.pdf

Raudenbush, S. W., &Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2ª ed.)*. Thousand Oaks, CA: Sage.

Robinson, A. H. (1950). Ecological correlation and the behavior of individuals.*American Sociological Review, 15*, 351-357

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness.A critical review of the knowledge base*. Dordrecht, The Netherlands: Springer.

Scheerens, J., &Bosker, R. J. (1997).*The foundations of educational effectiveness*. Oxford: Elsevier.

Scheerens, J., Witziers, B., & Steen, R. (2013).A meta-analysis of school effectiveness studies.*Revista de Educación, 361*, 619-645. doi: https://doi.org/10.4438/1988-592X-RE-2013-361-235

Schulz, W., Carstens, R., Losito, B., &Fraillon, J. (2018). *ICCS 2016 technical report.* Amsterdam: The International Association for the Evaluation of Educational Achievement.

Servicio de Evaluación Educativa del Principado de Asturias. (2018a) ¿Cómo se describen los resultados del aprendizaje en las evaluaciones del sistema educativo? *Informes de Evaluación, 14*. Recuperado de https://www.educastur.es/documents/10531/879356/2018_11_informe_evaluacion_N14_V03_r/997be880-9627-4456-9ebb-e5465c20a4df

Servicio de Evaluación Educativa del Principado de Asturias. (2018b). *Evaluación de Diagnóstico Asturias 2018. Niveles de rendimiento 6º de EP*. Oviedo, España: Consejería de Educación y Cultura.

Servicio de Ordenación Académica, Formación del Profesorado y Tecnologías Educativas del Principado de Asturias (2011). *Evaluación de Diagnóstico Asturias 2010*. Oviedo, España: Consejería de Educación y Ciencia.

Sirin, S. (2005). Socioeconomic status and academic achievement: A Meta-Analytic review of research. *Review of Educational Research, 75*(3), 417-453. doi: https://doi.org/10.3102/00346543075003417

Teddlie, C., & Reynolds, D. (2000).*The International Handbook of School Effectiveness Research*. London and New York: Falmer Press.

Towsend, T. (2007).*International handbook of school effectiveness and improvement*. Dordrecht, Netherlands: Springer.

Trautwein, U. (2007). The homework–achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort. *Learning and Instruction, 17*, 372-388.doi: https://doi.org/10.1016/j.learninstruc.2007.02.009

UNESCO-OREALC.(2016a). *ReporteTécnico.Tercer Estudio Regional Comparativo y Explicativo, TERCE*. Santiago de Chile: UNESCO. Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000247123

UNESCO-OREALC. (2016b). *Recomendaciones de Políticas Educativas en América Latina en base al TERCE*. Santiago de Chile: UNESCO.

UNESCO-OREALC, & LLECE. (2000). *Primer estudio internacional comparativo sobre lenguaje, matemática y factores asociados, para alumnos del tercer y cuarto grado de la educación básica. Segundo Informe*. Santiago de Chile: UNESCO.

UNESCO-OREALC, & LLECE. (2010). *SERCE. Factores asociados al logro cognitivo de los estudiantes de América Latina y el Caribe*. Santiago de Chile: UNESCO.

UNESCO-OREALC, & LLECE. (2016a). *Informe de resultados del Tercer Estudio Regional Comparativo y Explicativo. Logros de aprendizaje*. Santiago de Chile: UNESCO.

UNESCO-OREALC, & LLECE. (2016b). *Informe de resultados del Tercer Estudio Regional Comparativo y Explicativo. Factores Asociados*. Santiago de Chile: UNESCO.

vonDavier, M., Gonzalez, E., &Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series, Volume 2*, 9−36.

Woitschach, P. (2018). *Evaluaciones educativas a gran escala en Latinoamérica: TERCE* (tesis doctoral). Universidad Complutense, Madrid.

Woitschach, P., Fernández-Alonso, R., Martínez-Arias, R., & Muñiz, J. (2017). Influencia de los Centros Escolares sobre el Rendimiento Académico en Latinoamérica. *Revista de Psicología y Educación*, *12*(2), 138-154. doi: https://doi.org/10.23923/rpye2017.12.152

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007).*ACER ConQuest 2.0: generalised item response modelling software*. Camberwell, Victoria: Australian Council for Educational Research.