Jornet, Jesus M. (2016). Methodological analysis of the PISA international evaluation. RELIEVE, *22*(1), Art. M1. DOI: http://dx.doi.org/10.7203/relieve22.1.8293

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Methodological analysis of the PISA project as an international assessment

Análisis metodológico del proyecto PISA como evaluación internacional

Jornet Meliá, Jesús M.
Universidad de Valencia

Abstract

This paper aims to perform a comprehensive analysis of PISA as an international assessment program, and of the methodological characteristics of the metrical processes that underlie the instruments used in the project: achievement tests and questionnaires. The strengths and weaknesses of the project are identified in various areas: from the design and validation of the project as an international assessment program, its uses and ways of communicating results, to the metric characteristics of its instruments. Alternatives are proposed to optimize using the project in general, and in the involved countries in particular. It is concluded that given its methodological quality and socio-political impact, it is a valuable project, although some aspects as an evaluation program and its educational measurements could improve to provide higher quality information that could guide decisions for improvement.

Reception Date
2016 April 2

Approval Date
2016 June 16

Publication Date:
2016 June 17

Keywords:

PISA; methodological analysis; learning assessment; background questionnaires; assessment of educational systems; education; measurement.

Resumen

En este artículo el objetivo es realizar un análisis global del Proyecto PISA como programa de evaluación internacional. Asimismo, se analizan características metodológicas de los procesos métricos que sustentan los instrumentos que se utilizan en el proyecto: las pruebas de logro y los cuestionarios de contexto. Se identifican las fortalezas y debilidades del proyecto, en diversas áreas, desde el Diseño y validación del proyecto como programa de evaluación internacional, sus usos y modos de comunicación de resultados, hasta las características métricas de sus instrumentos. Se proponen alternativas para optimizar el uso del proyecto en general y en los países involucrados en particular. Se concluye que se trata de un proyecto que, por su calidad metodológica e impacto socio-político es valioso, si bien tiene diferentes aspectos que podrían mejorarse -tanto como programa de evaluación, como en sus fundamentos de medición educativas- para que pudiera aportar informaciones de mayor calidad y que pudieran orientar decisiones de mejora.

Fecha de recepción
2 Abril 2016

Fecha de aprobación
16 Junio 2016

Fecha de publicación
17 Junio 2016

Palabras clave:

PISA; análisis metodológico; evaluación del aprendizaje; cuestionarios de contexto; evaluación de sistemas educativos; educación, medición.

Assessing educational systems has increasingly become a common practice in recent years. National and international assessments have been established in the socio-educational scenario to provide information that has demonstrated educational quality as a central element in government policies.

Despite such systems being conducted internationally since the

1950s, it seems obvious that until the PISA Project was launched, their social impact was minimum. After PISA was launched, not only political interest was drawn to its results, but also social interest, and it even surpassed the information provided by the professionals involved in and from educational research.

In virtually all Latin American countries, the commitment of PISA to increase education quality and its efforts to ensure education would become a key factor to personal and social development, to eliminate inequalities and to prevent social exclusion. It has favoured the impact of international assessments to become increasingly stronger in the 21st century. Within this framework of interests and commitments, the PISA project has been established as an evaluation programme that has acted as a driver of concern for educational improvement.

Despite, however, PISA offering sound methodological development and good control, like all evaluation studies, some of its elements can improve so it has a real positive impact on societies and, in particular, on Latin American countries. This paper starts with a previous work (Jornet, 2014) that analyzed the unfinished business of large-scale assessments. So to a great extent, it continues with the working mode employed in this previous work to better specify the most interesting lines to work with, from our viewpoint, to improve the PISA project.

When we analyze any assessment program, we sometimes forget that evaluating and measuring are two completely different actions. In most large-scale assessments nowadays, measuring and assessing are presented as a single action known as an evaluation or even assessment (by way of example, see any national or international report on systems assessments). What measuring intends is to collect information about the frequency of a phenomenon, while with an assessment, by comparing the measured amount with well-specified value criteria, provides evaluative judgments. It is clear that measuring is instrumental for evaluating, but measuring should not be confused with assessing. Although it may seem quite an obvious point, it is not as we find in the literature and assessment practice many actions that are assessed according to their metric quality, but which forget validity, and even criteria and explanatory factors to a greater extent, that facilitate the understanding of the evaluated fact. Therefore, they represent an authentic assessment that provides information to be improved. For example, in the diagnostic assessments established in Spanish legislation, a systems approach is emphasized, although what are usually provided in practice as an evaluation are only students' achievement measures in a given centre, area, etc. Unfortunately in practice, the same situation is observed in most national and international assessments, including PISA.

For this reason, we will analyze the qualities of the PISA project from both perspectives: as an assessment program and as a measuring project.

In the aforementioned work, we arranged the analysis in three core workstations: a) large-scale assessment types: PISA's characteristics; b) pending subjects as assessment plans; and c) the pending subjects of these plans from the Educational Measurement perspective. In this case, we understand that this structure can help organise the PISA Project analysis, which we take as a reference to conduct the present article.

Finally, it should be noted that this article, together with that of Martínez-Rizo (2016), intends to provide an overview on the PISA analysis, and as a way to present all the papers included in this special issue. It involves renowned prestigious authors whose contributions

range from PISA's impact in Latin American context to specialized methodological contributions that can guide the Project's improvement and optimization processes.

The PISA project. Its characteristics as a large-scale assessment program.

We no go on to mention the history and development of PISA as it is well-known and there quality references that describe it from OECD (OECD, 1999), and those by other researchers (Cordero, Crespo & Pedraja, 2013; Martinez, 2006).

From our point of view, PISA's characteristics can be summarized as follows:

- This is a project whose *Analysis Unit* (on which the assessment is to be guided) is international, and which focuses mainly on reporting the differential result of educational systems from different countries based on analyzing the skills acquired by 15-year-old students in three key areas (maths, reading and sciences) [1]
- The *Educational Domain* (ED, hereinafter) or Measuring Universe, refers to the educational goals that represent the achievement of the learning or competencies that constitute the ultimate goal of an education system. Thus we refer to the ED as the series of objectives, activities and tasks that refer to an educational program in general or to a particular subject (Jornet & Suarez, 1989). Since the PISA Project came into being, it has had to face the difficulty of identifying a universe of measures that can be considered common for all the evaluated countries. As it was an international study, it was obvious that a specific curriculum design could not be taken as a reference because this would compromise the validity of the tests. For this reason, it took a theoretical educational construct as its reference, for which (in its three competences) efforts had to be made to provide a definition, which is reflected in the Project's theoretical frameworks (OECD, 1999, 2003; 2006; 2009; 2012; 2015). Here a cognitive approach was established where the idea of Competence was emphasized as an example of using acquired knowledge until the target age the Project considered for troubleshooting or for using habitual information in daily life.
- In all cases, *the target population* is a statistically representative sample of the countries that participate to complete each report. Yet each country has the possibility of expanding its sample to report on any stratum of local interest. For example in Spain, in the various waves of PISA, Spain's Autonomous Communities (ACs hereinafter) have been integrated as independent analysis units for the study to have representative samples in Spain, and in some -or most- ACs (MEC, 2009; MECD, 2012). This allows inferences to be made about Spain and the ACs to better represent the differential operation of the educational achievements made in the country.
- *The methodology used to design this type of testing is matrix sampling*. In this strategy, a conceptual framework that defines competition is designed to make assessments by identifying the key components of the characteristics of the competence to be evaluated. After the conceptual framework has been validated by experts, an agreement is reached that will act as a reference to design items. Normally a very large bank of items is generated because the purpose is to adequately represent the competence to be evaluated to ensure construct validity and content [2]. By matrix sampling, it is from this bank of items that booklets are produced to "sample" the mastering of content. These booklets include equivalent items in difficulty terms and succinctly represent content, and in

such a way that if all the booklets were applied, a detailed overview of the evaluated ED would become available. However, the matrix sampling strategy intends to provide viable assessment logistics as it would be impossible to apply all the items to all the subjects. So by applying a few of them, an inference can be made, with a certain level of error, about students' behaviour in the series of items. The aim of such testing is not to assess students individually because each student responds to a different set of booklets. Instead the intention is to infer a representative level of the different sample strata. However, all the booklets include anchor items that allow an estimation of students' achievement ro be made if they were administered all the items. This, in turn, allows the estimation of possible values to assign each student, known as plausible values (OECD, 2002; 2005; 2009b; 2012b; 2014).

- *All the PISA Project waves since 2000 have included background context questionnaire systems* to collect data about variables and/or indicators of input, process and context. They are taken as sources of information: students, teachers, management teams and families. With its different waves, background questionnaires have been refined and various complex indicators or compounds by different items have been included.

In short, and following the aforementioned typology, the PISA Project falls in the International Assessments category, based on theoretical constructs of the sample type.

Research areas to consider to improve the PISA Project

In order to arrange this review, we follow the scheme presented in Table 1. Seven lines to improve the Project and Assessment Program are identified, and three refer to its measuring instruments.

1. The PISA Project as an Assessment Program.

In the PISA Assessment Program, we focus on three elements that we find interesting: a) program design and validity; b) using its results; and c) communicating its results.

Any evaluation program must result from a particular educational quality concept, from which the criteria used to structure the way the value judgment is made that the assessment represents are specified.

Checking the level of competences acquired by the 15-year-old students (when they finish compulsory education) to help them develop in society (OECD 1999) is one declared PISA Project Assessment objective.

Table 1. Summary of the methodological review lines of the PISA Project

Methodological analysis lines	**Assessment**	Assessment design and validation (4)
		Uses of the Assessment (2)
		Communicating the Results (1)
	Measurement	Validation of Achievement Tests (2)
		Background Context (1)

However, the declaration made about the purpose of the *assessment involves a quality concept for assessment based on the excellence concept;* e.g., students' achievements are taken as a reference, and it is assumed that the higher average level of scores obtained by subjects in each country, the better the education system is. The PISA Project results from the aim to obtain educational achievement indicators to complete the framework of indicators that the OECD usually works with (MECD, 2014; 2015). Therefore, it is not surprising that it implies this inference made of the underlying quality concept of education systems. Laws on educational in any country tend to be more ambitious and emphasise the comprehensive training of people and its consequences for human development and social transformation as an objective of the education system. Obviously, if we represent the product of an educational system based on the achievements shown by external standard tests in three competitions, it is not possible to respond as to whether a system meets the psycho-socio-educational objectives in order to cover personal and social development requirements through each country's legislation.

The affective psycho-social aspects that best represent education and human development are not present. Likewise, as assessments do not focus on the internal assessment processes developed in each country, nor on the conditions which education operates in (input variables), nor on their contexts, we can hardly acquire a realistic image of quality of systems. As De la Orden stated (1997; 2007), according to how the effects considered in educational assessments are described, we point out that PISA provides information about the effectiveness of achieving results in the evaluated competences. Other aspects, such as efficiency (the extent to which the resources and means to meet objectives are optimized) or functionality (the extent to which the system responds to the specific social needs of each country), are not involved in the Project design draft. Even though we are aware that we simplify the analysis, the Project responds to only one question: what levels of achievement do the participating countries obtain in the measured competence? Secondarily, analyses that involve more variables, such as socio-economic and cultural levels (SECL, hereinafter), have been included as the normal reference used to interpret whether achievement is greater than expected for its SECL, or not.

Likewise from the contributions made to allow statistical data analyses, an analysis of the system`s fairness was introduced. This analysis is based on comparing intra- and inter-centre variance and its relations with the SECL of the families from the centres included in the study. From our point of view, although excellence and fairness are reported, it is difficult to establish any coherence between the two assessment objectives with a simplified approach. Nor can the current design report about the impact and relevance of education in a given country.

As a result of the described situation, we find that the results in many cases are misinterpreted by users (political and educational administrators). *This means that erroneous inferences are made of the project approach as an evaluation study* -see the studies about Consequential Validity by Mehrens (1997), Popham (1997) or Martínez-Rizo et al (2015).

Thus the PISA Project comes closer to a system based on indicators than an evaluation study because it is an analysis of achievements guided from the excellence concept, and not an analysis

of how educational processes occur in countries, nor of the impact and relevance that education can have in each country (Jornet, Sánchez-Delgado & Perales, 2015). The approaches that the OECD values as being recommended for teaching education, such as inclusive approaches, are not valued by PISA. In short, the Project is clearly a comparative and differential study in terms of the results, but is not an analysis based on a comparative methodology. So it is very difficult to draw lessons about educational innovation.

Can we improve this situation? Two lines can be addressed to make the Project more useful. First, *the Project's ED could be extended* by including the educational outcome concept in areas of the psycho-socio-emotional and social integration types. Moreover, it is necessary to consider that the quality of an education system cannot be represented only by students' results. To this end, the input variables, process and context of the systems used in each country should be analyzed, and the functionality of the educational system should be analzed in each case. Should this approach be undertaken by the OECD? Obviously, many more resources would be required, which could prove unfeasible. However, if we wish to make full use of the Project information, then national institutions should undertake this task. In order to confer coherence to the analysis, it would be desirable for PISA staff to establish trends to conduct these studies. A systems and holistic approach could better meet information needs to guide decisions at the macro-system level.

Another element to consider is the distance between the definition of the constructs worked in the PISA Project and the educational goals found in the curricula designs of each country's education system. The fact is that we cannot presently respond to a basic question from the assessment validity point of view: *To what extent does what PISA assesses represent each country's curriculum?* It might be more representative of one curriculum design than of others. This lack of information does not allow the interpretation of national scores to be contextualized, and could in fact prove to be an evaluative bias should there be different distances between the ED measured in PISA and the ED in each national curriculum design. We could state that if we differentiated among the original ED of the tests (the conceptual frameworks of the competencies measured in the Project), the ED implemented in each country (according to its own syllabi and circumstances) and the ED measured by tests, there would be a gap in the logic continuity to help us analyze study validity: we need to analyze the ED/the origin of the ED to test each national ED. In this case the analysis would not be very expensive to perform as it could be sustained in expert committees and could be a further guarantee to support a more accurate interpretation of the Project results -see Figure 1.

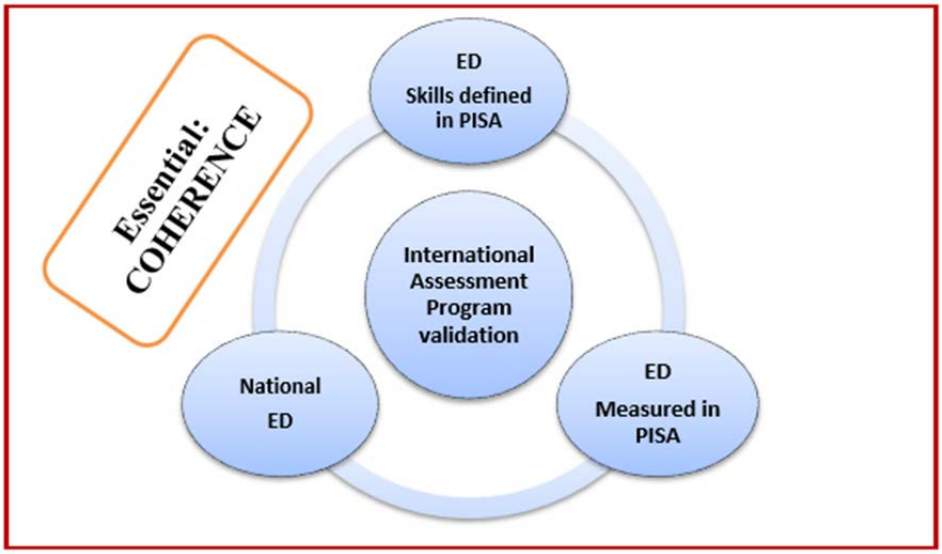

Figure 1. Graphical representation of the reference Educational Domains among which coherence must be set to achieve Project validity

Finally, as regards the PISA Project design as an international assessment, *we thought it necessary to insist on the need to emphasize having to analyze input variables, process and* context, and their relationship with the results. From a systems approach, "... equity, for example, will be reported from process variables, without disregarding the influence of context, the impact on the results, or the relationships between input and processes-, but will focus essentially on analyzing whether equal opportunities are created in the system (processes), and not only in the results analyzed with a single reference input (the family's SECL). To do this, it should be based on an orientation that would allow an approach to assess requirements (in the same sense as Tejedor proposes, 1995) (Jornet, 2014, p. 119).

Despite the considerable efforts made, including secondary research analyses done with Project data by both the OECD and some national assessment institutes, and the contributions made by private studies, such as *PISA In Focus*, we believe that still more efforts should be made.

Another problem that affects the PISA Project, and one that is common to other large-scale assessment studies, is the selection of subjects. Undoubtedly one of the strong points of PISA is that a thorough sampling and collection process is followed that deals with information representativeness and quality factors. However, it is also true that samples are set as being representative of a given ountry. Should a region in a country wish to be further analyzed as a separate unit of the country, its sample can be extended. However, when secondary analyses are done that refer to other sample strata, we ought to remember that work is already done with groups, and not with samples; this implies that the inferences made are affected by an increased sampling error.

Overcoming this fact is difficult because it is very costly to increase the number of cases to be included, and to achieve statistically representative samples in more strata. However with those studies in which work is not done with statistically representative samples, and those which expressly state that work is done in groups, but not with samples, any inferences made would be conditioned by a higher level of error. It is advisable that if the participating countries wish to use the PISA Project, they should answer some specific questions, and get involved in extending the sample for the required strata. In fact, most of the information that they provide in national differential studies by means of the variables that differentiate lower population strata, and are supported by the Project data, is performed with groups, save a few cases. This being the case, as previously

mentioned, the interpretations of any differential analyses done with groups, and not with samples, need to be relativized.

A third element to consider is that *the surveyed student is always a "normative" case;* e.g., people with diagnosed learning impediments, difficulties or disorders are not included; what this means is that work is done without considering the vulnerable groups for which education should particularly show its potential for change [3]. We understand that personal vulnerability goes beyond socio-economic vulnerability, which is usually that considered to refer to "vulnerable groups". Analyzing the education system's capacity as an inclusion element requires studying the processes that are set up to address these student types, and also studying the results obtained through reinforcement programs and structures (personal and material media) set up to help them overcome their difficulties (Booth & Ainscow, 2000).

Finally, the fourth element to consider in PISA Project design and validation are some further considerations that we present of differential analyses and the assessment of educational change.

We understand that systematisizing the PISA Project as regards the schedule it is to be applied in is initially appropriate. Any periodic study must be based on a well-structured schedule which sets the elements to be evaluated. However as previously mentioned, this fact is positive, but does not seem to be well-understood when those interested in the evaluation receive the results. We go on to offer a few examples.

As it is an international study, its periods of application are linked to structured innovation, which may have taken place in the participating countries. Assessing whether education improves, or not, in relation to a change in the legislation of a country requires a specific assessment process that considers pre-post and follow-up measures as the Programs Assessment must answer these questions. A project such as PISA can provide indirect information, but will always be partial, and not only due to what has been discussed above, but also because the schedule to apply it is independent of what happens in each country and its changes in legislation. In addition, educational change should be assessed when a law has been completely implemented, particularly as facts or changes in the internal work mode take place in schools and systems that may have a specific meaning. These meanings would, in any case, be those that should guide the focal point of the assessment, and answers are certainly not always found in school results.

Moreover, the follow-up studies done from the Project are necessarily cross-sectional (or transectional in the terms of Hernández, Fernández & Baptista, 2010). When comparing the level of the obtained results during different time periods, it is necessary to consider that it is evidently a matter of different subjects as the results of 15-year-olds are always analyzed-, but also any changes in the schools sampled in each wave may entail effects that cannot be controlled, although sampling is done with the same guarantees of representativeness and degree of trust. Indeed different cohorts from schools from each country are compared.

We have sometimes observed how national institutions have presented these data as if longitudinal studies were being done, without even questioning the real equivalence of the tests, nor the error in the difference which may appear among tests from one wave to another. When working with large samples, the most widely observed result is stability as years go by. Lack of clarification of analysis type and the way information is contributed is evident among different PISA waves. In any case, social transformations in general and, changes caused by education, are observed in the mid-long term. During 3-year periods it is difficult to verify significant changes in achievement levels. Implementing changes in ways of working in each country in general, and in schools in particular, requires time. Bear in mind that 15-year-olds are always assessed, so

any changes to be noted from one PISA wave to another should have been made in what has been worked, at least from 12 to 15 years of age. If we bear in mind the differential emphasis obtained in the analysis of each competence, we should be able to note changes from the application made in the first wave and in the third one, in which a re-evaluation is made by placing the same emphasis on each competence. This means that since PISA was applied on the first occasion, changes should be immediately introduced so that the course taken by compulsory education is taken completely –or almost completely, in each participating country- until the age of 15. In any case, from the time information is collected until reports are published, an almost 1-year latency time is considered. As previously mentioned, socio-educational changes and, more specifically changes in teaching practices, need more time because efforts must be made to understand and identify the system's weak points, to arbitrate solutions to overcome them (organizational changes at school, changes in didactic methodologies, in teacher training, etc.), and more time is needed to implement them. If success is achieved through the changes made, and we can see these changes when a complete generation has been educated within this innovation frame, we may require some 15-20 years to see such changes. So, although we have indicated that the schedule seems adequate, perhaps it will be necessary to prolong times among waves at least every 5 years, and then leave these intermediate times to better exploit information nationally, to arbitrate improvement measures, etc.

Next we understand the socio-cultural diversity in our world as a wealth rather than a problem. Globalization effects always have two sides; a positive one and a negative one. Tension between appreciating a local culture and internationalization is an on-going debate as a more profound analysis is required to strike a balance between respect and appreciating a country's culture, and what is assumed to be a reference internationally. As some authors have indicated, what actually occurs is that the definition of the competences measured in the Project is oriented to a type of society in which engineers and technologists eventually form the elite that allows a country's economy to develop in the same development model (Martínez-Rizo, 2016). Training requirements to adequately develop personally in a given geopolitical and social space do not have to be the same in all countries. Hence the need to reconsider the competences to be assessed and analyzed in their definition of local characteristics, which are not just those defined in a national curricular design. We can learn from others, but it is necessary to recognize the potential in each country, and according to its circumstances, to contribute social improvement, inclusion, co-existence and participation elements. A single desirable model does not exist. So we have to identify a relevant challenge for the PISA Project in particular and for any international assessment project: promote actions to recognize diversity. Despite it seeming anecdotal, an example can help us to better understand this perspective. Scientific aspects are evaluated about the way climate changes can be understood; yet a farmer or a shepherd in a "third world" country, whose training is based on living in the countryside all their lives and, at any rate, listening to their elders, can read and predict changes by simply contemplating clouds up in the sky (how high they are, their colors, shapes, etc.), watching the direction the wind moves in, noticing the smell of the land and plants. Quite often we check that their predictions are more successful than those offered by weather stations which resort to large quantities of information that is analyzed by stochastic models. Children and adolescents from developed countries will probably never locate the most important stars or constellations. Our world is diverse, and so is knowledge. An assessment program that is not open to include diversity, and is considered based on the opinions of a group of specialists from a few developed countries [4], which markedly lack assessment validity, constitute an element that guides toward knowledge globalization, which

leads us to forget socio-cultural diversity. As an assessment program, *it needs to bear this in mind because a program that ignores diversity runs the serious risk of its assessment lacking justice and equity.*

Finally, analyzing effect size is often not generalized (Borges & Sánchez-Bruno, 2004; Frías, Pascual & García, 2000; Ledesma, Macbeth & Cortada De Kohan, 2008). Quite often small differences are mentioned as being verified, but it is questionable that they are really meaningful from the substantive perspective.

So, *will we be able to improve this situation?* In this case, we understand that technical solutions to improve analyses can be found with time. We do not mean that the result would be more realistic if the same schools always acted as references because it is likely that as they would be subjects to be analyzed, they would end up guiding education according to what the Project assesses and how it is assessed –typology of instruments-, with what the results would falsify in practice. Random sampling distribution in each wave by controlling the equivalence of schools' characteristics during each period is more valid, and is currently the case, but by increasing the control over selecting the participating schools, and by orientating toward the homogeneity of the typology of schools analyzed between waves. In any case, we feel that the solution can be found by enhancing the culture of assessing and better defining the project's usefulness, which we will look at later. Among such proposals, it might be possible to: a) specifically inform about the real equivalence levels studied among the tests in the different waves (especially the error in the difference); b) report on effect size; c) inform about the degree of homogeneity in the typology of the schools studied in each wave; d) explain as explicitly as possible the limits of interpreting the obtained results; and e) include a methodological complementarity consideration (Bericat, 1998) by supporting countries so that studies can be given with local good practices from their national institutes through a qualitative approach based on case studies. With this work, both the OECD –in charge of the Project- and National Institutes have categorical work to do, which should be done jointly by marking lines as to how to understand information on the changes in education that the Project offers and how such studies are undertaken.

Regarding the PISA Project uses as an assessment program, we stress three aspects that affect the best way to put its results to good use.

First, we indicate that initially, and possibly given its origin as a study of indicators, it is *mainly descriptive; that is, it describes the levels of achievement obtained by students in each country, but does not emphasize the study of the factors that can explain them.* Lack of explanatory studies (although this aspect has been progressively alleviated ever since the Project came into being) makes understanding why certain levels are obtained, but not others, very difficult. Having very few explanatory studies no doubt limits its usefulness as this situation does not favor understanding results. It is true that the OECD, national institutes and independent education researchers have been conducting explanatory studies. The contributions made by multilevel studies (Hox, 2002; Murillo, 2008; Andréu, 2011; Murillo & Hernández-Castilla, 2011a; 2011b) have increased the usefulness of such assessments in general, and of the PISA project in particular. Nonetheless, the low level of quality information they provide should indicate that the efforts made are not enough. Highly sophisticated techniques are employed to identify information that was verified many years ago. Indeed very little progress has been made since what Coleman et al (1966) verified about the socio-economic level and family culture being the best predictors of academic achievement. Almost all the studies done afterward with assessment projects, including PISA, have provided similar data (Gaviria, Martínez & Castro, 2004; Lizasoain & Joaristi, 2010; Murillo & Hernández-Castilla, 2011).

The fact that today's analyses have become more sophisticated and background questionnaires have improved (as evidenced in the PISA Project) has allowed other related factors to be identified. However, usefulness of information does not improve what has been contributed from other research-type approaches (Murillo, 2003, 2007).

No doubt *the PISA Project has been designed to meet a macro-analytical vision, rather than a meso- and micro-analytical vision.* The groups involved in education (from lawmakers and politicians to teaching staff) expect to find more responses in the project to improve educational organization in general, and the teaching practice in particular (Jornet, García-García & González-Such, 2014). As in former cases, we once again wonder, *could we adopt another strategy to improve this situation?*

Logically the effort made to carry out the PISA Project and its socio-political impact requires considering ways to enhance its usefulness. We go on to mention some options.

Different studies have indicated that from the macro-analytical viewpoint, and beyond the influence of countries' socio-economic and cultural level, other variables related to their socio-economic structure must exist, which may be related with the objective and subjective social value conferred to education (Jornet, Perales & Sánchez-Delgado, 2011; Sancho-Álvarez, Jornet & González-Such, 2016). It is likely that completing the Project at the most comprehensive and systemic macro-analytical level could maximize its utility to orientate political decisions, but it would require a model that works achievements comprehensively with its mediate and immediate contexts.

Drawing conclusions for education practice is no easy task as the Project is not designed for this, although it provides some indications (Carabaña, 2015). However in the terms of De la Orden (2012), any assessment is optimizing in nature. So we understand that an effort must be made from national institutes, and not by the OECD. We can find one good example in the work done by the INEE in Mexico with its works into PISA for teachers (e.g., see INEE, 2005). "Read the Project Results" which intends to learn to improve (in collaboration with specialists in measuring-assessing and teachers), and implies clear benefits. Nor can we expect a project with such characteristics to provide us everything. Guiding the teaching practice requires other evaluative considerations. Likely with the line that emerges from PISA for schools it is possible to assess whether it is a good reference for improvement. However, we believe that this line should not be prioritized when the project is underway because it would enter "the market of accreditation or certification evaluations" like any other model would (along with the European Foundation for Quality Management Excellence Model –EFQM-, ISO Standards, etc.), when it has not been designed for this. For the time being, it is merely a start but, as it is praiseworthy, we estimate that national institutes should establish the necessary connection between the report results and any potential users of each assessment project, and also in this case, as a form of development of their assessment culture.

Finally, we simply point out that the PISA Project has published its databases ever since it began. Likely there is an excellent research opportunity lying in them. It is true that they are complex, and it is necessary to have had a high level of training to use them suitably. Yet we estimate that evaluation and education researchers have not generally made the best of these chances. Could we improve this situation? No doubt we could. It would be worthwhile increasing the number of researchers who examine data from the Project in depth. As researchers we cannot, and must not, expect the Project to offer us all its potential without doing anything. Researchers from each country are in the best position as they know the culture, socio-educational interests and problems of their country more directly to put forward study hypotheses and objectives. We should get more involved to maximize the benefits that evaluation studies

in general, and the studies that result from the PISA Project in particular, may offer us. Indeed, the PISA Project contributes documents and training alternatives so that we researchers learn the structure of the bases and how to deal with some peculiarities of the data, e.g., about plausible values.

In conclusion as regards PISA's Design and Validity as an assessment project, *communicating the results is possibly one of the areas that should be most carefully reviewed.* In assessments, it is always necessary to differentially consider the potential users or audiences involved (Green, 1988; Weiss, 1984). The report is structured as any type of research report would be, and addresses technicians, and not the possible users of an assessment of this kind. So it is not surprising that deficient interpretations are often made of the Project results. Doubtlessly this report is carefully prepared technically, and has included contributions and innovations over the years. Some uses of graphs that were not frequently resorted to have become popular among researchers. Indeed researchers seem to understand these graphs, but this is not always the case with professional users in education. We find in a recent work that professionals (with a degree and a master degree) did not understand some of the graphs most frequently used in this Project (García-Bellido, 2015). Is this a PISA Project problem? It certainly does not produce such a problem, but does not consider the level of comprehension of those it addresses.

In any case, the way its results are transmitted is incorrect. Use of ranking is stressed in particular. It has been rejected by many specialists in evaluations as an assessment practice to communicate results (Martínez-Rizo, 2015; Martínez-Rizo, 2016; Ravela, 2002, 2003, etc.). Why? There are several reasons for this and we go onto provide a few.

Merely ordering countries according to a single criterion leads to errors when appreciating the levels that each country shows. In many cases, minor differences (1 or 2 points) can be ignored by whoever receives the report, who only examines the position that the target country of its analysis occupies. A ranking is never completely contextualized.

In the PISA Project, attempts have been made to contextualize levels of achievement with the SCEL of countries, but this is not sufficient. The internal characteristics of countries determine whether their results indicate a greater or lesser extent of functionality for their own particular circumstances, whether the system adequately responds to social expectations insofar as it is able to. So it does not answer basic questions in education: have the goals that our system set been achieved?; does the system provide more or less what can be achieved in our country?; what does the education system contribute – identified by the level of accomplishment at a given international position- for the personal social development of the citizens of our country? These would be the questions that we would be interested in evaluation studies of education systems responding. In a diverse and complex world, one defined by inequalities in development opportunities, information contributed as a ranking only provides mistaken visions that arouse ill-informed political debates that do not help to guide improvement processes. Competitiveness, and not competition, is what stimulates one way of transmitting results of this type. A ranking is justifiable in sports competitions where winning is what is important, and winning by 1 or 5 points, or by 1 second or tenths of a second, is of no importance. The social assessment cannot be treated with simplification that leads to clear mistakes. We have too much evidence for this, which has resulted in not only pointing out Consequential Validity as a quality factor of assessment studies, but also in a shortage of most large-scale assessment projects, and also with the PISA Project (on which we work with this monographic work, and on which several studies have been conducted; e.g.: Taut & Palacios, 2016).

Simple orderings are done without informing about the level of measuring error which could explain that certain differences would be meaningless, which would be a fallacy. Can any expert explain with ranking data whether a difference between 498 and 490 points is due to some people acquiring such and such a competence, but others do not? When we are evaluating, is it not necessary to improve, provide the keys to understand fundamental differences between facts and evaluation-based phenomena?

Could this situation improve? Populations seem to be used for rankings, and no doubt it is a tempatation for assessors to offer their results via this method. Once again, it is a matter of working to enhance the evaluation culture. Different alternative or complementary solutions could be contributed. Some have been passed by some national institutes (e.g. the INEE of Mexico or Spain).

It would first be necessary to identify the analysis –macro-, meso- and micro- analytical-plans in which it is possible to understand whether certain political practices in organizing education or in school/classroom administration may be related with achievement levels, or not. Would this task be one to be managed by the OECD? Not necessarily. The OECD could play a role like an orchestra conductor; that is, without actually playing a musical instrument directly, it harmonizes information so that each country better understands what its results contribute and where it could make improvements; that is, draw lines of use for each analysis level. Distributing responsibilities to assume improvement is essential when we work according to an overall international study approach.

A second question is if data are provided as rankings, whether any statistically significant differences exist among countries will be specified. Processing this information with the number of participating countries is a complex task, but if real criteria standards are established (Jornet & González-Such, 2009) to interpret achievement levels, this option would minimize the negative effects of observing rankings or grouping countries by normative standards to a great extent.

Another action that could be arbitrated from the OECD is that of identifying groups of countries from structural variables; e.g., population size and dispersion, the socio-economic structure of geopolitical regions, kinds of offers made by teaching centres, etc. These clusterings could be dealt with from analyzing conglomerates (e.g., using k-means). Next achievement results would be analyzed for clusters (each cluster's average and if there were any differences among clusters). In this way the obtained results could be better contextualized.

Along the same line, the conglomerates analysis could be performed according to socio-cultural characteristics (existing language –or languages- in countries, levels of immigration and internal population migration, etc.), or conglomerates of countries could be analyzed according to indicators of economic development or per geopolitical region.

Basically, it would be a matter of overlooking rankings and integrating information about countries into categories and contextualizing these categories by context variables, and not only by their SCEL.

Evidently with the information provided, it is extremely difficult, if not impossible, to make improvement recommendations when we should understand that an assessment that does not contribute this type of information can hardly be useful.

All in all, it is a matter of marking interpretation lines that guide users to what can be stated according to a country's results and what is not licit to state.

2. The PISA Project: some notes about its metric characteristics

In reviewing the PISA Project characteristics according to their metric aspects, we use three main nuclei; two refer to

achievement tests and one to background questionnaires.

First with academic achievement tests we refer to problems related to the Design of Test Contents which, to a great extent, have been previously mentioned regarding the Project Validity as an Assessment Programme. In this case, we center specifically on the elements that affect the validity of the measuring instruments themselves.

First we notice the marked lack of basic informaction: no studies exist on aligning these tests with the construct from which they originate, nor with interpretative standards. Empirical evidence is lacking about the validity and content of the construct. Studies on aligning tests with the original ED and interpretation standards are fundamental to ensure the validity of the interpretation of test scores. What does obtaining a score of 523 mean? What kind of competences have been acquired and which ones have not? It is not easy to answer these questions, and when attempts are made to answer them, they are done quite primarily by referring only to some examples of which items have respondents not known how to answer in one country or another, which normally serve only to offer the press headlines. If alignment studies are done by taking a cognitive taxonomy as a reference, they can provide the necessary bases to respond to this type of questions, which are basically those that express whether the test results are valid and, consequently, of any use. *Can this situation be improved?* Enough methodologies are found in the test validation studies domain to offer more exact responses based on committees of specialists who analyze alignment levels (Martínez-Rizo, 2015; Rothman, 2004; Webb, 1997; Webb, Herman & Webb, 2007). They are not complex studies, but are thorough, and not only result in a better interpretation of scores, but also offer the evidence required to improve test validity.

Along the same review lines, we find something else is lacking: Cultural Validity studies. An international project necessarily implies risks from which invalidity factors appear through cultural and linguistic differences. Some studies exist with various international projects that demonstrate the functioning of differential items (DIF, its acronym in English), or even a bias, which would involve favoring the observation made of differentiated achievement levels linked to items being improperly devised. In national studies, cultural and/or linguistic diversity appears in the country that is the study object. As in the previous case, solutions for this type of problems can be arbitrated. First, items being analyzed by judgment committees that follow methodologies like those described by Solano-Flores, Contreras-Niño and Backhoff, (2006), Solano-Flores (2009, 2013) or Basterra (2011). A study that currently addresses over 80 countries cannot be expected to test work with the same validity guarantees, and not even with the same metric properties.

Any standardized test developed in a given socio-educational context requires an adaptation/validation study to be applied in a different context (Hambleton, 2005; Solano-Flores, 2008). It would be worthwhile informing about the pilot studies conducted in the participating institutions and integrating into a specific study about the validity and metric properties of tests. This would guarantee the interpretability of the study, would help to qualify the uses and interpretations of scores, and would provide better quality guarantees to project users.

Institutional collaboration with the OECD will be fundamental to be able to make such information available.

One unquestionable contribution made by the OECD through PISA is the number of research reports it has generated (see: http://www.oecd-ilibrary.org/education/oecd-education-working-papers_19939019?page=1). However for DIF and bias, it is striking that although an adequate methodology exists in an international study, studies with such characteristics do not appear in the Project reports for this purpose (e.g., see Camilli, & Shepard, 1994; González-Montesinos & Jornet, 2010). It would be an

interesting area for independent researchers given this shortage. In this case, we understand that the responsibility should fall on those who devise the tests and, if applicable, on collaboration with national institutes.

Another relevant factor is that the PISA Project is based on the specific design of items, which is typical, characteristic. A normalized assessment has "its own personality" (Ruiz-Primo & Li, 2015; Ruiz-Primo, Li & Minstrell, 2014; Shavelson, et al., 2002). The advances that could have been made to develop them and in the manner of dealing with designing items may, conversely, be an uncontrolled factor of a possible DIF and/or bias for not clearly reflecting the way students are normally assessed in class. Doubtlessly, the analysis should be greater than that we do along these lines. We realize that for a competence to be considered to have been acquired, students should be able to solve any problem related to it, presented in any evaluative format, even though it differs considerably from what they normally face. However, it is well-known that people tend to study depending on how they are evaluated. The form an evaluation takes conditions the expression of the achievement level as it reveals the way in which it has been taught. Related solutions also pose questions to be answered.

As the Project releases items in every wave, should teachers be trained to design similar items so that throughout compulsory education (until the age of 15) students would also be evaluated with such item formats? Among learning resources, should students be submitted to different evaluation formats, including the PISA format, or that of other evaluation projects? From our point of view, one solution could be to evaluate students in different ways as it would benefit generalization of learning. Yet this is merely an opinion as it would condition the work of teachers, who are the people who best know how to orientate teaching more if it is done with an individualized, personalized character. This could be a matter for the Education Committee to debate.

One fact that could improve the understanding of the scores observed with the Project would be to analyze the forms of assessment to which the students who form part of each wave are submitted, and if the distance between the ways in which they have been evaluated were related with the obtained scores were investigated. This could prove to be further evidence of validity. Indeed studies on instructional sensitivity demonstrate that distances between evaluation formats may be an explanatory factor of the obtained scores (Ruiz-Primo & Li, 2015). Nonetheless, this would imply establishing assessment protocols on the way in which teachers assess students in each country. It certainly is no easy task, is no doubt complex and will not be easily apprehended from only an external study based on questionnaires that address the teacher or the student. All in all, satisfactory experiences exist on which a quite suitable representation of the assessment practices that teachers undertake can be obtained (Martínez-Rizo, 2012). Clearly this type of analysis should be led from the PISA Project (by establishing lines). However for one to be done, national institutes would have to take it on because it should be dealt with from methodological complementariness (quantitative-qualitative – Bericat, 1998-).

Finally, the need to inform about the degree of implementing the curriculum and its distance with the tests done in each country is stressed. This aspect we mention generically when the factors that intervene in the PISA Project Design and Validity as an assessment programme are analyzed. In this case, one verification that should at least be made from the data that teachers can offer through background questionnaires, but one that specifically refers to the items that tests include, could provide very useful information to better understand the scores obtained.

The second analysis nucleus to which we refer is that related to studying the metric properties of the PISA Project tests. We will not spend too much time on this aspect as it is probably one of the best attended ones in the

Project. However, as to the way of using the Item Response Theory (IRT, its acronym in English) was a controversial point at one time. Difficulties have been overcome and now it is considered one of the strongest points in the PISA Project. In any case, it is worth remembering the need to follow protocols to adapt tests when it comes to international tests by following the recommendations and protocols provided in works like those by Hambleton, Merenda and Spielberg (2005) or, more recently, those in the work of Muñiz, Elosúa and Hambleton (2013).

The matrix sampling technique to design and develop tests is complex. In fact very few people are really specialized in this methodology. One problem that we believe is not being clarified is the real equivalence of the booklets designed for each wave and between waves. We understand this information will be key to assume whether the final process employed actually matches the desirable metric properties, and whether it maintains indications of content validation between booklets. Another point is that tests are translated and/or adapted to different languages from distinct countries, and have different socio-educational contexts, curricula, etc. It will even be necessary to verify if the distributions of booklets among countries are equivalent or not, simply with a frequency analysis of the booklets that have finally been responded in each country. It will be necessary to include such studies as a guarantee that matrix sampling processes suitably fulfil their purpose.

Regarding the final aspect, background questionnaires, it is noteworthy that since the first PISA wave to the present-day, their quality and use of its indicators have improved greatly, regardless of them being simple or complex. The characteristics of these questionnaires do not differ much from those normally employed in other national or international evaluation studies. Among their strong points, it is worth stressing that considerable efforts have been made to include what we call complex indicators, and also to identify simple indicators (items) which have a differential capacity for achievement levels. As in so many other background questionnaire systems, difficulties arise from their initial design. No clear theory can be identified that orientates as to which explanatory factors need to be measured. Clearly, support is partially provided by education research findings about some factors that can explain achievement, but no systemic theory exists that confers sense to the series of elements considered in questionnaires.

Nor are there data about their metric properties –or about the subscales they include-, nor evidence for their validation. As measurement instruments, and also as achievement tests, they probably receive less attention in the Project as a whole. Invariance among countries does not occur, conversely to what happens in achievement tests, which is logical.

Doubtless, education is a political and cultural manifestation. Therefore, it is a direct expression of the psycho-socio-cultural way the role that dducation plays in each country is understood. From our viewpoint, it would be desirable within the Project to mark lines for the final background questionnaires' configuration, but for there to be freedom to determine a wide sampling area of the information obtained from them in each country. Other evaluation systems differentially identify indicators by countries to better represent the characteristics and data that are interesting for each country. For instance, Laeken Portafolios; although it is a model of evaluations based on indicators that are not comparable with PISA, it introduces a strategy that could be useful in the Project we are considering. The following are distinguished: Primary, Secondary and Tertiary Indicators. It is compulsory to apply the Primary and Secondary ones in each country to collect fundamental information, which is common among countries. The Primary ones will provide key information, while the Secondary ones will help qualify the former by including minor socio-cultural

adaptations from each country. The Teriary ones will be freely available to each country so they will be able to sample any data that could be of interest during the socio-educational development of each country. We are aware that, from the PISA Project, up to three items in each country can be included, which differ from the common series of items. Nonetheless, we believe that this is a low proportion as it does not allow progress to be made in specific knowledge of problems that might be of local interest.

The final usefulness of background questionnaires is based on their capacity to explain achievement and their status to describe the procedural, input and background elements that must serve to better understand the obtained results, regardless of whether or not analyses are established that relate background data with the results. This we feel is a weak point, but one that can be improved if measures are taken to improve its validity and the control of its metric properties (for a methodological review to develop background questionnaires, see Jornet, López-González & Tourón, 2012).

Conclusions

When an assessment project that has reached a high level of social and political impact is analyzed, its critics are watchful to find reasons in technical works to help them make it disappear. Its supporters expect quite the opposite; that is, they seek basic reasons to affirm its value and permanence.

In this case, we believe that the PISA Project can be improved. Yet precisely because of the socio-political impact it has had, it presents a value that other projects have not obtain, regardless of them being national or international: *place education in the center of social and political concern.* If only for this reason, it is worth working to optimize their ways of doing things and their contributions. This is doubtlessly a chance for societies to pay attention to improve their education systems and, if possible, their organization ways and teaching practices.

Apart from this appreciation and general position, we believe that it is necessary to manifest any shortages or gaps that can be improved. This is not doubt a highly ambitious, complex project with considerable technical sophistication. As a project that is orientated from a very powerful institution like the OECD, we understand that it has every chance of surviving and being established internationally, which we consider positive. Indeed it is worthwhile participating in it and maintaining it. However, we estimate that it should be re-evaluated according to its capacity to overcome relevant themes:

- Problems with its validity, as an assessment program, by improving the validity of curricular characteristics, the socio-economic and cultural characteristics of each country.
- Increase the technical quality of the achievement tests by also working to improve their validity, curricular alignment and criterial standards, and by devising items.
- Tacking different approach for developing background questionnaire by conferring more roles to national institutions.

  Help provide data that allow Compared Studies to be conducted rather than just a differential comparative study about achievement levels.
- Establish analysis plans that allow their results to be better understood by project users, which would improve their consequential validity.
- Work hard on the ways by which assessment information is provided so that it is genuinely clear and understandable for users.
- Finally, remember that any assessment project implies value judgments. Therefore, it is action whose whole methodological overview must be adapted to ensure ethical poisitons that affirm justice and equity. Hence the need for DIF and bias studies.

We understand that the development lines that have been adopted, such as “PISA for centers”, or its “computer applications”, are not precisely a priority, nor will they provide the

improvements required for the Project to be more valid as an assessment program.

Only by improving the quality of the Project (which is already high) will we be able to make comprehensible data available and, in short, to increase its credibility and usefulness, two key factors for validating assessment programs.

Briefly, and as previously set out, from a conceptual perspective, socio-cultural diversity is one of humanity's assets, and not a problem. International studies should be committed to combine local perspectives with international references, and the latter do not entail a risk for diversity to progressively diminish. Education is also immersed in globalization and cannot constitute an element that dilutes the typical characteristics of each country, but contributes to improve it. As we understand it, *the PISA Project is an opportunity for international dialog.* Hence we believe that in order to improve the Project, it is necessary to combine, to a greater extent, the characteristics oriented from the OECD with those which can prove functional for each country. We encourage national institutes and agencies to collaborate to dialog among all the authorities to help find solutions that help solve some of the problems mentioned herein. The OECD's role of marking lines and active collaboration by national institutions, will no doubt allow improvements to be made. We already have an example of such: the PISA Latin American Group (GIP). In this line of collaboration and openness, the work by Andreas Schleicher in this monograph about the future of PISA is most interesting. We are grateful to him for it.

It presents us with proposals to open up to psycho-socio-affective constructs and the willingness of PISA personnel to collaborate with participating countries within a model that conceives education in an ever increasingly interconnected world; aspects –which mainly fall in line with those mentioned herein- that can no doubt contribute to increase the Project's usefulness.

References

Andréu, J. (2011). El análisis multinivel: una revisión actualizada en el ámbito sociológico. *Metodología de Encuestas*, Vol. 13, 161-176.2 ISSN: 1575-7803

Backhoff, E., Bouzas, R., Contreras, C., Hernández, P. & García, P. (2007). *Factores escolares y aprendizaje en México. El caso de la educación básica*. México: INEE.

Basterra, M. R. (2011). Cognition, culture, language, and assessment. In M. R. Basterra, E. Trumbull, and G. Solano-Flores (Eds.), *Cultural validity in assessment* (pp. 72-95). New York: Routledge.

Bericat, E. (1998). *La Integración de los métodos cuantitativos y cualitativos en la investigación social. Significado y medida*. Barcelona, Ariel Sociología.

Booth, T. & Ainscow, M. (2000). *Index for Inclusion*. Bristol: CSIE.

Borges, A. & Sánchez-Bruno, A. (2004). Algunas consideraciones metodológicas relevantes para la investigación aplicada. *Revista Electrónica de Metodología Aplicada*, *9*(1), pp. 1-11.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousands Oaks, CA: Sage.

Carabaña, J. (2015). *La inutilidad de PISA para las escuelas*. Madrid: La Catarata.

Coleman, J. S.; Campbell, E.; Hobson, C.; McPartland, J.; Mood, A.; Weinfeld, F. & York., R. (1966). *Equality of educational opportunity.* Washington: Government Printing Office.

Cordero, J. M., Crespo, E. & Pedraja, F. (2013). Rendimiento educativo y determinantes según PISA: una revisión de la literatura en España. *Revista de Educación*, 362, 273-297. DOI: http://dx.doi.org/10.4438/1988-592X-RE-2011-362-161

De la Orden, A. (2007). Evaluación de la calidad de la educación. Un modelo sistémico como base para la construcción de un

sistema de indicadores. En INEE, *Conceptos, metodologías y experiencias para la construcción de indicadores educativos* (pp. 6-21). México: Instituto Nacional para la Evaluación de la Educación (INEE).

De la Orden, A. (2012). La función general de la evaluación y la optimización educativa. Ponencia invitada en el *I Foro Iberoamericano de Evaluación Educativa*. México: Ensenada, UABC-IIDE (5-7 Noviembre). Recuperado de http://uee.uabc.mx/uee/eventos/primerForoRIEE/ponencias/6.pdf

De la Orden, A. (Dir.) (1997). Desarrollo y validación de un modelo de calidad universitaria como base para su evaluación. *RELIEVE*, *3*(1), art.2. DOI: http://dx.doi.org/10.7203/relieve.3.1.6334

De la Orden, A., & Jornet, J.M. (2012). La utilidad de las evaluaciones de sistemas educativos: la consideración del contexto. *Bordón*, *64* (2), 69-88.

Ercikan, K., & Solano-Flores, G. (2016). Assessment and sociocultural context: A bidirectional relationship. En G. T. L. Brown & L. Harris (Eds.), *Human Factors and Social Conditions of Assessment*. New York: Routledge.

Ercikan, K., Roth, W.-M., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education*, *27*, 275-285.

Frías, M.D., Pascual, J., & García, J.F. (2000). Tamaño del efecto del tratamiento y significación estadística. *Psicothema*, 12(2), 236-240.

García-Bellido, Rosario (2015). *Diseño y validación de un instrumento para evaluar la competencia "Aprender a aprender" en profesionales de la educación.* Tesis Doctoral. Universitat de València. Recuperado de http://roderic.uv.es/handle/10550/43599

Gaviria, J. L., Martínez, R., & Castro, M. (2004). Un Estudio Multinivel Sobre los Factores de Eficacia Escolar en Países en Desarrollo: El Caso de los Recursos en Brasil. *Policy Analysis Archives*, *12*(20). DOI: http://dx.doi.org/10.14507/epaa.v12n20.2004

González-Montesinos, M.J. & Jornet, J.M. (2010). *Modelo para detección de funcionamiento diferencial de reactivos (DIF) en pruebas INEE. Informe Técnico 2010*. México: INEE, Dirección General de Pruebas, Documento interno.

Green, J. (1988). Stakeholder Participation and Utilization in Program Evaluation. *Evaluation Review*, April (12), 91-116. DOI: http://dx.doi.org/10.1177/0193841X8801200201

Hambleton, R.K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R.K. Hambleton, P.F. Merenda, & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.

Hernández, R., Fernández, C., & Baptista, P. (2010). *Metodología de la investigación*. México: Editorial Mc Graw Hill.

Hox, J. (2002). *Multilevel Analysis: Techniques and applications*. London: Lawrence Erlbaum Associates.

INEE -Instituto Nacional para la Evaluación de la Educación de México- (2005). *PISA para docentes: La evaluación como oportunidad de aprendizaje*. México D.F.: INEE. Recuperado de http://www.inee.edu.mx/index.php/84-publicaciones/materiales-para-docentes-capitulos/455-pisa-para-docentes-la-evaluacion-como-oportunidad-de-aprendizaje

Jornet, J. M. & Suárez, J. M. (1989). Conceptualización del dominio educativo desde una perspectiva integradora en Evaluación Referida al Criterio (ERC). *Bordón*, 41, 237-275.

Jornet, J. M. (2014). Asignaturas pendientes en las evaluaciones a gran escala. En M. C.

Cardona, y E. Chiner. (Eds.). *Investigación educativa en escenarios diversos, plurales y globales*. (pp. 115 – 128). Madrid: EOS.

Jornet, J. M., Perales, M. J. & Sánchez-Delgado, P. (2011). El Valor Social de la Educación: Entre la Subjetividad y la Objetividad. Consideraciones Teórico-Metodológicas para su Evaluación. *Revista Iberoamericana de Evaluación Educativa*, *4*(1), 51-77. Recuperado de http://www.rinace.net/riee/numeros/vol4-num1/art3.pdf

Jornet, J.M. & González-Such, J. (2009). Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios sobre Educación*, *16*, 103-123. Recuperado de http://dadun.unav.edu/bitstream/10171/9172/1/16%20Estudios%20Ee.pdf

Jornet, J.M., García-García, M. & González-Such, J (Eds.)-(2014). *La evaluación de sistemas educativos. Informaciones de interés para los colectivos implicados*. Universitat de València –PUV-.

Jornet, J.M., López-González, E., & Tourón, J. (Coords.) (2012). Los cuestionarios de contexto en la evaluación de sistemas educativos. *Bordón*, *64*(2). (Monográfico).

Jornet, J.M., Sánchez-Delgado, P. & Perales, M. J. (2015) *La evaluación del impacto y la relevancia de la educación en la sociedad.* Universitat de València -PUV-.

Ledesma, R., Macbeth, G. & Cortada de Kohan, N. (2008). El tamaño del efecto: una revisión conceptual y aplicaciones de la vista con sistema de estadística. *Revista Latinoamericana de psicología*, *40*(3), 425-439.

Lizasoain, L. & Joaristi, L. (2010). Estudio Diferencial del Rendimiento Académico en Lengua Española de Estudiantes de Educación Secundaria de Baja California (México). *Revista Iberoamericana de Evaluación Educativa,3*(3), pp. 115-134. Recuperado de http://www.rinace.net/riee/numeros/vol3-num3/art6.pdf

Martínez Rizo, F. (2009) Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *REDIE*, *11*(2), 1.18. Recuperado de http://redie.uabc.mx/redie/article/download/231/388

Martínez Rizo, F. (2012). *La evaluación en el aula: promesas y desafíos de la evaluación formativa*. México: Universidad Autónoma de Aguascalientes.

Martínez-Rizo, F. (2016). Impacto de las pruebas en gran escala en contextos de débil tradición técnica: Experiencia de México y el Grupo Iberoamericano de PISA. *RELIEVE*, *22*(1). DOI: http://dx.doi.org/10.7203/relieve.22.1.8244

Martinez-Rizo, F. (Coord.) (2015). *Las pruebas ENLACE y Excale. Un estudio de validación. Cuaderno de Investigación No. 40*. México. DF: Instituto Nacional para la Evaluación de la Educación. Recuperado de http://publicaciones.inee.edu.mx/buscadorPub//P1/C/148/P1C148_01E01.pdf

Martínez, R. (2006). La metodología de los estudios PISA. *Revista de Educación*, núm. Extraordinario, 111-129.

MECD. (2014). *Panorama de la educación. Indicadores de la OECD 2014*. Madrid: MECD. Recuperado de http://www.mecd.gob.es/dctm/inee/indicadores-educativos/panorama2014/panorama2014web.pdf?documentId=0901e72b81b20622

MECD. (2015). *Panorama de la educación. Indicadores de la OECD 2015*. Madrid: MECD. Recuperado de http://www.mecd.gob.es/dctm/inee/internacional/panorama-de-la-educacion-2015.-informe-espanol.pdf?documentId=0901e72b81ee9fa3

Mehrens, W. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, *16*(2), 16-18.

Ministerio de Educación, Cultura y Deporte. (2012). *PISA 2012. Programa para la Evaluación Internacional de los alumnos. Informe español.* Madrid: MECD. http://www.mecd.gob.es/dctm/inee/internacional/pisa2012/pisa2012lineavolumeni.pdf?documentId=0901e72b81786310

Ministerio de Educación (2009). PISA 2009. *Programa para la evaluación internacional de los alumnos. Informe español.* Madrid: Ministerio de Educación. http://www.mecd.gob.es/dctm/ievaluacion/internacional/pisa-2009-con-escudo.pdf?documentId=0901e72b808ee4fd

Muñiz, J., Elosua, P. & Hambleton, R.K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema*, *25*(2), pp. 151-157. DOI: http://dx.doi.org/10.7334/psicothema2013.24

Murillo, F.J. (2003). Una panorámica de la investigación iberoamericana sobre eficacia escolar. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, *1*(1), 1-14.

Murillo, F.J. (2007). *School Effectiveness Research in Latin America*. En T. Townsend (Ed.), International Handbook of School Effectiveness and Improvement, (pp. 75-92). New York: Springer.

Murillo, F.J. (2008). Los modelos multinivel como herramienta para la investigación educativa. *Magis, Revista Internacional de Investigación en Educación*, 1, 45-62.

Murillo, F.J. & Hernández-Castilla, R. (2011a). Factores escolares asociados al desarrollo socio-afectivo en Iberoamérica. *RELIEVE*, *17*(2). DOI: http://dx.doi.org/10.7203/relieve.17.2.4007

Murillo, F. J., & Hernández-Castilla, R. (2011b). Efectos escolares de factores socio-afectivos. Un estudio Multinivel para Iberoamérica. *Revista de Investigación Educativa*, 29(2), 407-42. Recuperado de http://revistas.um.es/rie/article/view/111811

OECD. (1999). *Measuring student knowledge and skills. A new Framework for assessment.* Paris: OECD. Recuperado de http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33693997.pdf

OECD. (2002). PISA 2000 Technical Report. Paris: OECD. Recuperado de https://www.oecd.org/pisa/pisaproducts/33688233.pdf

OECD. (2003). *The PISA 2003. Assessment Framework. Mathematics, Reading, Science and Problem Solving knowledge and skills.* Paris: OECD. Recuperado de http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33694881.pdf

OECD. (2005). *PISA 2003. Technical Report.* Paris: OECD Publishing. Recuperado de https://www.oecd.org/pisa/pisaproducts/pisa2003technicalreport.htm

OECD. (2006). *Assessing Scientific, Reading and Mathematical Literacy. A Framework for PISA 2006.* Paris: OECD. Recuperado de http://www.oecdilibrary.org/docserver/download/9806031e.pdf?expires=1465551043&id=id&accname=guest&checksum=02E5A7F7B73F1CCFA0DA6E8336A2F3D8

OECD. (2009). *PISA 2009. Assessment Framework. Key competencies in reading, mathematics and science.* Paris: OECD. Recuperado de https://www.oecd.org/pisa/pisaproducts/44455820.pdf

OECD. (2009b). *PISA 2006. Technical Report.* Paris: OECD Publishing. Recuperado de https://www.oecd.org/pisa/pisaproducts/42025182.pdf

OECD. (2012). Pisa 2012. *Assessment and analytical Framework. Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD. https://www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf

OECD. (2012b). *PISA 2009. Technical Report.* París: OECD Publishing. Recuperado de

https://www.oecd.org/pisa/pisaproducts/pisa2009technicalreport.htm

OECD. (2014). *PISA 2012. Technical Report.* París: OECD Publishing. Recuperado de https://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm

OECD. (2015). PISA 2015. *Assessment and Analytical Framework. Science, Reading, Mathematic and Financial Literacy*. Paris. OECD. http://www.keepeek.com/Digital-Asset-Management/oecd/education/pisa-2015-assessment-and-analytical-framework_9789264255425-en#page201

Popham, W. (1997). Consequential validity: Right concern - wrong concept. *Educational Measurement: Issues and Practice*, *16*(2), 9-13.

Ravela, P. (2002). *¿Cómo presentan sus resultados los sistemas nacionales de evaluación educativa en América Latina?* Documento de Trabajo No. 2. Santiago de Chile: PREAL. Recuperado de http://www.preal.org/docs-trabajo/ravela_n22.pdf

Ravela, P. (2003). *¿Cómo aparecen los resultados de las evaluaciones educativas en la prensa?* Grupo de Trabajo sobre Estándares y Evaluación del PREAL. Recuperado de http://www.preal.cl/GTEE/pdf/prensa.pdf

Rothman, R. (2004). Benchmarking and alignment of state standards and assessment. En S. Fuhrman and R. Elmore (Eds.), *Redesigning accountability systems for education (pp. 96-114).* New York: Teachers College Press.

Ruiz-Primo, M.A. (2006). A Multi-Method and Multi-Source Approach for Studying Fidelity of Implementation. CSE Report 677. SEAL, Stanford University/CRESST.

Ruiz-Primo, M. A., & Li, M. (2015). The relationship between item context characteristics and student performance: The case of the 2006 and 2009 PISA Science items. *Teachers College Record*, *117*(1), 1-36.

Ruiz-Primo, M. A., Li, M., & Minstrell, J. (2014). Building a framework for developing and evaluating contextualized ítems in science assessment (DECISA). Proposal submitted to the DRL–CORE R7D Program to National Science Foundation. Washington, DC: National Science Foundation.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. & Klein, S. (2002). On the evaluation of systemic education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39(5), 269-393. DOI: 10.1002/tea.10027

Ruiz-Primo, A., Jornet, J.M. y Backhoff, E. (2006). *Acerca de la Validez de los exámenes de la calidad y el logro educativos (Excale).* México: Instituto Nacional de Evaluación Educativa. Recuperado de http://www.uv.es/gem/gemhistorico/publicaciones/Acerca_de_la_Validez_de_los_examenes_de_la_calidad_y_el_logro_educativos_Excale.pdf

Ruiz-Primo, M.A.; Li, M.; Wills, K.; Giamellaro, M.; Ming-Chih, L.; Mason, H., & Sand, D. (2012). Developing and Evaluating Instructionally Sensitive Assessments in Science. *Journal of research in science teaching. 49*(6), pp. 691–712.

Sancho-Álvarez, C., Jornet, J. & González-Such, J. (2016). El constructo Valor Social Subjetivo de la Educación: validación cruzada entre profesorado de escuela y universidad. *Revista de Investigación Educativa*, 34(2). DOI: http://dx.doi.org/10.6018/rie.34.2.226131

Shavelson, R. J., Ruiz-Primo, M. A., Li, M., & Ayala, C. C. (2002). *Evaluating new approaches to assessing learning*. CSE Technical Report 604. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) University of California, Los Angeles.

Solano-Flores, G. & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation: An International Journal on Theory and Practice*,

19(2-3), pp. 245-263. DOI: http://dx.doi.org/10.1080/13803611.2013.767632

Solano-Flores, G. (2008, July). A conceptual framework for examining the assessment capacity of countries in an era of globalization, accountability, and international test comparisons. Paper given at the *6th Conference of the International Test Commission*. Liverpool, UK.

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. En M. R. Basterra, E. Trumbull, & G. Solano-Flores, *Cultural validity in assessment* (pp. 3-21). New York: Routledge.

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice, 28*(2), pp.9-18.

Solano-Flores, G., Contreras-Niño, L.A., & Backhoff, E. (2006). Test translation and adaptation: Lessons learned and recommendations for countries participating in TIMSS, PISA, and other international comparisons. *REDIE: Electronic Journal of Educational Research*, *8*(2). Recuperado de http://redie.uabc.mx/vol8no2/contents-solano2.html

Tejedor, F. (1995). Perspectiva metodológica del diagnóstico y evaluación de necesidades en el ámbito educativo. Metodología en el diagnóstico y evaluación en los procesos de intervención educativa. *Actas del V Seminario de Modelos de Investigación Educativa.* Murcia: AIDIPE.

Webb, N. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education. Council of Chief State School Officers and National Institute for Science Education.* Madison, Wisconsin: Wisconsin Center for Education Research, University of Wisconsin.

Webb, N.M., Herman, J. & Webb, N.L. (2007). Alignment of mathematics state-level standards and assessment: The role of reviewer agreement. *Educational Measurement: Issues and Practice, 26*(2), 17-29.

Weiss C.H. (1984). Toward the future of stakeholders approaches in evaluation. En R.F. Conner, D.G. Altman & C. Jackson (Eds), *Evaluation studies. Review Annual, 9*, pp. 255-268. Beverly Hills, California: Sage Publishers.

Notes

[1] Although compulsory education does not finish in all countries at the age of 15, this agreement was taken as it was the most representative.

[2] Used in most large-scale sampling-type tests: TIMSS, PIRLS (both of the IEA), State diagnosis tests from Spain, EXCALE of the INEE from Mexico, for example.

[3] Some evidence exists that such cases are eliminated from the sample upon sample collection. The comparative scheme of achievement levels entails that, in many countries, these students are avoided from being considered as they tend to show a better level.

[4] See in Annex I the summary of the specialists who have participated in developing the theoretical frames of each competence. The diversity of specialists –who are no doubt very well trained- fom which –and how many- countries they come. Hence the need for contributions made from GIP as a beginning to consider typical Latin American characteristics.

Acknowledgments

This work has been conducted as part of the R&D&I project of an Educational and Social Conhesion System: designing an evaluation model of requirements (SCEL/EVALNEC). Ref. EDU2012-37437, financed by the Spanish Ministry of Economy and Competitiveness.

Annex 1

Countries to which the specialists responsible for the conceptual frames of the tests belong

2000	2003	2006	2009	2012	2015
Mathematics					
Netherlands, Italy, Ireland, Spain, Denmark, Korea, USA, Austria	Netherlands, Japan Germany, 2 USA, Slovak Republic, Spain Ireland, Poland, Denmark, Korea, ,	Netherlands, Germany, USA, Poland, Denmark, Japan	Netherlands, Germany, USA, Poland, Denmark, Japan	2 Australia, Germany, 3 USA, Japan, Poland, Denmark, UK	
Reading					
2 USA, UK, Canada, Japan, Nederland, Belgium, Finland, France, Germany,	2 USA, UK, Canada, Low Countries, Belgium, Finland, France	Netherlands, 2 USA, UK, Canada, Belgium, Finland, France	2 USA, Japan, Low Countries, UK, Belgium, Korea, France, Germany, Spain		
Sciences					
UK, Australia, Switzerland, Korea, Norway, Germany, 2 USA	UK, Australia, Switzerland, Norway, Germany, 2 USA, Korea,	USA, Poland, Australia, Slovak Republic, Italy, UK, Norway, 2 France, Japan, Germany	Australia, Norway, Japan, 3 Germany, 2 France, China, Netherlands, 2 USA, Finland,		2 UK, Germany, South Africa, USA, France, Australia, Singapore
Problem solving					
	2 USA, Hungary, UK, 2 Netherlands, Germany, Greece			2 Germany, Hungary, 3 USA, Luxemburg, Singapore	
Technical experts					
			5 USA, Australia, 2 France, Belgium, 2 Netherlands,	Germany, Mexico, Singapore, 3 USA, 2 Netherlands,	4 USA, 2 Germany, Chile, Norway, Japan, Cyprus, Netherlands
				Financial Literature	**Scientific Literature**
				2 USA, France, New Zealand, Australia, Czech Republic, Canada, UK	3 USA, Denmark, Netherlands, Italy, Japan, China

Source: Obtained from the OECD 2000; 2003; 2006; 2009; 2012; 2015.

Note: a number that appears alongside a country indicates the number of specialists from this country who have intervened in designing the conceptual frame in each case

Authors / Autores

Jornet-Meliá, Jesús M. (jesus.m.jornet@uv.es).
Professor of Educational Evaluation and Measurement from the Department of Research and Diagnosis Methods in Education from the University of Valencia (Spain). He is one of the co-editors of this monographic section about "international Assessments: PISA". His address is: Facultad de Filosofía y Ciencias de la Educación. Universidad de Valencia. Avda. Blasco Ibáñez, 30. 46010, Valencia (Spain).

To know more / Saber más

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa
E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]

© Copyright, RELIEVE. Reproduction and distribution of this article is authorized, provided its content has not been modified and its origin is indicated (RELIEVE Journal, volume, number and electronic address of the document).

© Copyright, RELIEVE. Se autoriza la reproducción y distribución de este artículo siempre que no se modifique el contenido y se indique su origen (RELIEVE, volumen, número y dirección electrónica del documento).