Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

A Methodological Critique of the PISA Evaluations

Una crítica metodológica de las evaluaciones PISA

Fernandez-Cano, Antonio
University of Granada (Spain)

Abstract

This paper conducts a methodological evaluation of the PISA international evaluations, giving a critical analysis of their shortcomings and limitations. A methodological review or meta-evaluation has been carried out on the multiple PISA reports in an attempt to demonstrate the plausible validity of the inferences that PISA maintains given a series of methodological limitations such as: an inconsistent rationale, opaque sampling, unstable evaluative design, measuring instruments of questionable validity, opportunistic use of scores transformed by standardization, reverential confidence in statistical significance, an absence of substantively significant statistics centered on the magnitudes of effects, a problematic presentation of findings and questionable implications drawn from the findings for educational norms and practice. There is an onus on PISA to provide and demonstrate more methodological rigor in future technical reports and a consequent need to be show great caution lest unfounded inferences are drawn from their findings.

Reception Date
2016 January 21

Approval Date
2016 May 10

Publication Date:
2016 May 11

Keywords:

Evaluation; Meta-evaluation; Methodology of evaluation; PISA

Resumen

En este trabajo realizamos una evaluación metodológica de las evaluaciones internacionales PISA, presentando un análisis crítico de sus deficiencias y limitaciones. Presentamos una revisión metodológica o meta-evaluación de los múltiples informes PISA, en un intento de demostrar la validez plausible de las inferencias que PISA mantiene, teniendo en cuenta una serie de limitaciones metodológicas tales como: una lógica incoherente, toma de muestras opacas, diseño evaluativo inestable, instrumentos de medición de validez cuestionables, el uso oportunista de las puntuaciones transformadas por la normalización, la confianza reverencial en la significación estadística, la ausencia de estadísticas sustantivamente importantes centradas en las magnitudes de los efectos, una presentación problemática de los hallazgos e implicaciones cuestionables extraídas de los resultados para las prácticas y las legislaciones educativas. Recae sobre PISA la responsabilidad de proporcionar y demostrar mayor rigor metodológico en los futuros informes técnicos y la consiguiente necesidad de ser cuidadosos para no mostrar inferencias sin fundamento a partir de sus hallazgos.

Fecha de recepción
21 Enero 2016

Fecha de aprobació
10 Mayo 2016

Fecha de publicaci
11 Mayo 2016

Palabras clave:

Evaluación; Meta-evaluación; Metodología of evaluación; PISA.

Programme for International Student Assessment (PISA) is a standardized comparative program related to large-scale international assessments run by a series of participating countries, either members or associates of the Organization for Economic Co-operation and Development (OECD). As a large study PISA can have implications for a range of stakeholders; thus, it is important to review the information the study sets forth with a critical eye.

It was first administered in 2000 and has been run every three years since, to assess three basic academic competences: reading comprehension (in paper and digital formats),

Corresponding author / Autor de contacto

Fernandez-Cano, Antonio. Universidad de Granada. Facultad de Ciencias de la Educación. Campus Universitario La Cartuja. 18071-Granada (España). (afcano@ugr.es).

mathematical competence and scientific competence, although a different competence is given special emphasis during each three-year cycle. The most recent assessment (2012) was administered to 510,000 students aged 15, their school grade varying according to the participating country, who represented a population of 28 million young people from 65 participating countries (OECD, 2013a).

The data provided for the analytical treatment involve 65 countries distinguishing between several groups of variables related to students achievement and parents considerations (470) and variables related to the educational center or school (249) but it does not consider soundly pedagogical variables about best practices of an efficient teaching for school learning; only in the last report, PISA 2012, it does some considerations in a section of questions titled "School Governance, Curriculum, and Assessment" (OECD, 2015a).

PISA belongs to a long line of international educational large-scales assessments whose predecessors are the IEA (International Association for the Evaluation of Educational Achievement) studies, carried out since 1958 by IEA (2012), and whose most recent and important studies are: TIMSS (*Trends in International Maths and Science Study*) and PIRLS (*Progress in International Reading Literacy Study*). Both TIMSS and PIRLS differ from PISA inasmuch as they emphasize the curricular dimension of classroom practice, while PISA emphasizes learning that takes place alongside the school curriculum and permits students to apply processes and content to the real-world context.

A meta-evaluation of the program according to the Scriven's guidelines is of paramount importance for its improvement, interpretation and dissemination, as well as for an appropriate use of its findings (Berliner, 2011; Scriven, 2011). Nevertheless, the methodological consistency of the PISA assessments displays certain questionable features and the program theory has yet to test the validity of their underlying assumptions, and especially their internal validity, which some (i. e. Hanberger, 2014) consider poor.

In the light of American Joint Committee on Standards for Educational Evaluation (Yarbrough, Shulha, Hopson & Caruthers, 2011), questions about standards of accuracy and are exposes seeking to ensure that an evaluation as PISA will reveal and convey technically (methodologically) adequate information. We must remember that the accuracy standards are intended to increase the dependability and truthfulness of evaluation representations, propositions, and findings, especially those that support interpretations and judgments about quality.

Objective of this study

This study is not intended as a devastating critique of PISA; rather, it seeks only to provide a methodological "meta-evaluation" of an assessment program, in this case PISA, which constitutes an immense undertaking, laudable but open to improvement, with thousands of participants, millions of testable students and an enormous economic outlay. So then, this paper is not a mere critique of the survey but how PISA could be improved.

The central objective is twofold: to carry out a specific methodological analysis of the PISA assessments and a general analysis of its large-scale survey method; reviewing many PISA-related topics to make for an effective argument: its improvable research methodology. It is not realistic for PISA to set itself up as the sole arbiter when it comes to evaluating educational systems, and therefore its results and indicators ought to be interpreted within a specific cultural context and a particular conceptual framework, namely the idea of real-life problem-solving or competence (Meyer & Benavot, 2013).

The methodology used here is methodological critical analysis or meta-evaluation (Scriven, 2011; Stufflebeam, 2011). Meta-evaluative studies such as this ought to be regarded as a habitual and desirable practice, whilst recognizing the difficulty and great risk inherent in every

meta-evaluation as a threat to its validity: that it may end up being no more than a string of unconnected statements with little bearing on the issue being evaluated. In addition, another threat related to the lack of methodological accuracy and inadequate coverage of important information about the meta-evaluation of international programs was detected by Bollen, Paxton and Morishima (2005).

In short, this paper addresses a number of methodological limitations of PISA that impact the inferences that can be drawn from the results.

Reviewing PISA's methodological approach

Official versions about the methodology used in PISA are available in Adams (2003); Adams, Berezner and Jakubowski (2010); Adams, Wu and Castensen (2007), and above all in the successive OECD technical reports, especially the 2009 technical reports (OECD, 2012).

But methodological critics came soon. The first predominantly methodological criticisms of PISA were voiced by Prais (2003), targeting PISA 2000 data for Great Britain. Prais queried the construct underlying the instruments, on the grounds that they bore little relevance to the school syllabus; the use of inappropriate samples, since they included students older than 15, above all those who had repeated a year; and very low response rates (60 %). Prais suggested that methodological flaws in PISA had resulted in an apparent improvement in the attainment of British students, particularly when compared to their Swiss and German counterparts. Adams (2003), a PISA 2000 coordinator, promptly accused him of being unaware of the methodology used by international assessment studies and by PISA in particular.

Brown, Micklewright, Schnepf and Waldmann (2007) questioned the robustness of the PISA findings and those of other international assessment studies (TIMSS & PIRLS) and the rationale for generating aggregate scores on the basis of respondents' answers. Drechsel, Carstensen and Prenzel (2011) addressed methodological problems and set out requirements for constructing tests for future assessments in the light of their significance for scientific education.

PISA and comparable international studies are generally vulnerable to criticism, especially in terms of the survey methodology they use, the questionable validity of their measurement instruments due to their tenuous relationship with national curricula (Dolin & Krogh, 2010), and the serious mis-interpretations of results by people ill-qualified to do so in many studies (Carnoy & Rothstein, 2013).

Methodological issues

The various methodological issues examined below serve to highlight the potential limitations of the PISA inferences about the general standard of accuracy.

Inconsistent rationality

As an external, accumulative evaluation, PISA is full of inconsistencies revealing the irrationality of the process; specifically it should be noted that the OECD does not impose PISA rather national ministries of education choose to participate but the evaluative model is imposed –not sought, or desired, or agreed to; that is a political imposition without any consideration to students, teachers or parents' opinions.

The underlying correlational hypotheses in the evaluative-economics model are ill-stated and their successive concatenations are ill-founded (Banks, 2012; De Witte and Kortelainen, 2013). PISA hypothesize that assumes economic wealth correlates positively with its students' mastery of knowledge and competences, and that scholastic performance predicted on the basis of certain scholastic variables depends on certain characteristics of the schools, as a consequence, if the schools in an academic system fulfill certain performance-boosting characteristics, the wealth of the country concerned will rise. Both the initial hypothesis and the conclusions drawn from it

are somewhat simplistic. This assumption would be maintained by an initial finding from PISA 2003 estimated a average correlation with statistical significance of 0.32 (accounting for about 10% of variation in achievement scores) but this varied considerably across participating countries.

The explanatory factors that it puts forward as important are correlational rather than causal in nature; therefore no causal decision could generalize as surreptitiously PISA seeks to although in PISA multiple regression models or multilevel models are used. These models are statistical models and obviously do not rely on causal relationships but PISA findings obtained and exposed even in the proper reports are badly interpreted as causal inferences *ad nauseam*; a relation of cautions about inferences from the case of PISA 2009 could read in Ercikan, Roth and Asil (2015). Concretely its inferences about what makes schools successful (OECD, 2015b; Volume IV); the relation students and money (OECD, 2015b; Volume VI); the spurious causation by unique and lineal that ready to learn is conditioned by students' engagement, drive and self-beliefs (OECD, 2015b; Volume III).

PISA is an evaluation that has no direct consequences in promoting the person being evaluated (15-years student), hence the examinee's plausibly low level of commitment to the tasks. An index of the use of assessment was derived from eight items asked to school principals but only one item was used with the purpose to make decisions about students' retention or promotion (OECD, 2013b) oscillating from 1% in Iceland to 98 % in Portugal (OECD, 2013b, p. 149). Its purpose is not the improvement of the students or the educational system, in the Spanish case (Couso, 2009), in the Russian case (Zuckerman, Kovaleva & Kuznetsova, 2013), in Taiwan (Zhang & Sheu, 2013) or in Germany with students who repeat a class learn in mathematics (Ehmke, Drechsel & Carstensen, 2008); rather, it is a subtle exercise in accountability, and competitiveness aimed at justifying current spending and rationalizing future spending on education.

It is difficultly assumable the PISA's unified model based on competencies for real-life situations (DeSeCo, 2008; OECD, 2013a, p. 3; OECD, 2014a, p. 4; Rychen & Salganik, 2003) because the vital performances of an Egyptian female peasant, an Andean male shepherd, an female executive of Tokyo or a German male industrial operator are not the same. In addition, competencies in the PISA frameworks have been subject to considerable evolution across the six PISA surveys completed so far.

Opaque sampling

PISA provides an extensive technical documentation in the manifold reports about exclusion rates and sampling procedures; particularly sample sizes are carefully developed and described. However, some omissions about the sampling are manifest; and above all an assumption is adopted questionably that there are comparable groups of students in the respective national samples. But neither age-based nor grade-based sampling strategies can achieve balanced samples in terms of both age and schooling (Strietholt, Rosén & Bos, 2013).

In this line, Krohne, Meier and Tillmann (2004) have previously complained in the German context about the exclusion of students with special educational needs, who were evaluated differently or left out; this, coupled with the inclusion of students who are repeating a year and the diversity of trajectories (or "tracks"), raises questions about the appropriateness of the population surveyed. These restrictions make it questionable whether PISA can be regarded as an inclusive program.

Sampling assumptions are not usually set out explicitly, either in terms of the sampling technique used, the sample sizes, associated sampling errors or, most importantly, the mortality rates given the low rate of response, of around 15% in PISA 2006, 25% in PISA 2009 and 20% in PISA 2012, as well as the proportion of excluded students, which ought

not to exceed 5% of the students selected from the target population; or worse still, the existence of differential exclusion rates depending on the countries concerned (differential mortality).

Moreover there are countries where schooling is not compulsory at 15; the age that PISA rigidly insists on adopting[1]. The sampling becomes yet more opaque in view of the fact that there is no certainty regarding what international agreements are made regarding a high mortality rate that may prove critical.

Another worrying aspect of the sampling is the sample size of the countries. The variety of sizes, out of proportion to the population of the country, remains questionable despite the statistical flourishes that are produced in their defense; for example, it is difficult to accept the comparability of a country like Austria, with scarcely eight million inhabitants, appearing in PISA 2009 with a total sample of 6,590 students from an accessible population of 82,135, and Japan, with 126 million inhabitants supposedly represented by 6,088 students drawn from an accessible population of 1,060,381 (OECD, 2012, p. 188). Then Knipprath (2010) claimed that the quality of Japanese education has not been well investigated because a non representative with a shorted-scale datasets.

By other side and as a positive feature, some national samples are over-sampled, for example in 2012 some countries such as Belgium, Colombia, Italy and especially Spain invested in oversampling and could obtain data at regional level.

PISA relies on the use of resampling[2] and in particular on balanced repeated replication (BRR) using Fay's method (Judkins, 1990), a statistical technique used to generate country specific standard error estimates and obviously not a re-sampling of individual students; but if the initial sample is not representative it does not make sense to rely on any statistical resampling, or any technique of replicability (whether it be bootstrapping[3], jackknifing, BRR or cross validation).

Flawed evaluative design

The PISA design is very simple, basically involving a trends design for the observation of tendencies, but not, unfortunately, a longitudinal design involving a panels or cohorts, like the one that is widely used in the analogous NAEP assessments (National Center for Education Statistics, 2012). PISA still falls some way short of NAEP[4], especially in terms of methodology because above all NAEP use longitudinal designs.

The measurement structure for evaluating educational effectiveness in PISA continues to rely increasingly on one sole indicator (performance of large samples in a large-scaled test) and the use of these test scores in isolation of other indicators that also capture what it means to be effective.

In this line of metrical pluralism, a consistent longitudinal study operating with all of PISA three-annuals applications would undoubtedly be welcome and easy to achieve given that PISA raw data have been made available in various reports (OECD, 2009; OECD, 2013a), but with the reservation that the instruments, comparing one edition to the others, are neither equal nor parallel, with some items having been discarded, new ones introduced and previous ones amended; notwithstanding PISA dispose calibrated item banks. It is present the question of linking errors between items, already recognized and addressed in PISA 2009 (OECD, 2012, p. 143-146), stemming from discrepancies in difficulty indices ("international percent correct", to use the PISA expression) between one application and another (OECD, 2012, p. 215-228).

The string of causal inferences derived from PISA need to be interpreted with multiple reservations; such reservations are appropriate to all observational designs with correlational techniques of data analysis, included regression weights, even if sophisticated they may be. That venerable methodological dictum "correlation is never synonymous with cause" is pertinent here, although PISA belatedly tried to cover all bases by announcing, almost as an

afterthought, that "PISA does not measure cause and effect" (OECD, 2013a, p. 29); but plenty of the inferences that have been drawn from it have been spuriously causal.

PISA omits moderating "objective" variables that can only be included as part of a complex factorial design; it would then be appropriate to refer to factors into a multilevel modeling approach, and not just in passing as PISA does. It also fails to consider the covariates that might be included as propensity scores, as manifestly significant as spending on education, student-teacher ratios, remuneration-supervision of teaching staff and other more qualitative but equally objective factors such as the ability of students, a history of full school attendance, or teacher qualifications. In general, PISA includes sociological variables but leaves out psychological, cultural and above all pedagogical variables, such as those mentioned above, which according to earlier research (Wang, Haertel & Walberg, 1993) may well be pertinent.

Questionable validity of the instruments of measurement

The performance tests, which are characteristic of large-scale evaluations, exhibit multiple shortcomings relating above all to validity.

Unconvinced applications of Rasch model

In PISA low validity of measurements may be exacerbated by the matrix-like nature of the questions administered, for which there is no guarantee of equivalence, and not every examinee is administered every item, since the comparisons are not based on a common test, rather, different students answer different questions because it does not occur that every item of the applicable item pool is administered to every examinee. As consequence, there is no guarantee that the neither booklets nor items administered are equivalent to one another which poses currently intractable challenges to estimating individual achievements. It is a questionable reductionism to consider that the items included in each booklet are items with known difficulty indexes using only estimations previous to administration. It is also a reductionist consideration accepting as suitable the use of items matrix sampling because only it is a well-known and widely used common approach in large-scale assessments estimating a population achievement; but as Rutkowsky (2014) advises this imputation model (more commonly called a conditioning model) used in PISA data, is assumed to be fully measured, without error, although departures from this assumption can have a meaningful impact on conditioning model parameter estimates, subpopulation achievement estimates, and under- or over-estimates of subpopulation differences.

The validity of the evaluative instruments' content is also highly questionable, firstly because of the small number of items each student answered from the booklets, generated by matrix sampling, meaning there is no guarantee or evidence that the parallel structures of each test are indeed parallel, all the more so since the items' discrimination indices, which do not appear in the reports, are unavailable. More questionable is appealing to psychometric scoring procedures for adaptive assessments and relying on the use of IRT (item response theory) to generate parallel structures, specifically the Rasch model, to compare scores derived from various forms of the test, based solely on difficulty indices and without capturing all the pertinent dimensions of the test, may prove counterproductive. Kreiner and Christiansen (2014) supply ample evidence that the scaling used in PISA, based on the Rasch model, is extremely unsuitable given the high differential item functioning, which undermines the robustness of country rankings. Some countries administered easier items to low-ability students, and some even excluded items on the basis of their "poor psychometric characteristics", though these items functioned well in the vast majority of other countries (OECD, 2012, p. 132). Secondly, there is the questionable suitability of the items, given their unjustified lack of

connection with national curricula[7] and still less with the officially-sanctioned textbooks representing those curricula.

The items to be included should also consider the item discrimination index. It is not technically and statistically impossible to take into account the discrimination index. Obviously the Rasch model used in PISA must not then applicable because it involves a single parameter model based in difficulty index calculated by the number of correct responses. A two-parameter model should be considered despite that the plausible interpretations of the results were much difficult. For example, Lu and Bolt (2015) offer a two-level multidimensional item response model to provide an informative way of studying the effects (or lack thereof) of cross-country variability in response style.

The real challenge as Kubinger, Hohensinn, Hofer et al. (2011) expose is to meet constraints determined by numerous moderator variables such as different response formats and several topics of content. PISA administer the same item at different positions within a booklet are used; therefore the occurrence of position effects influencing the difficulty of the item is a crucial issue. Not taking learning or fatigue effects into account would result in a bias of estimated item difficult (Hohensinn, Kubinger, Reif, Schleicher & Khorramdel, 2011).

The multi-language nature of the tests

The multi-language nature of the tests, which need to be translated into a host of languages, raises questions about whether they can even be considered parallel, let alone equivalent, instruments amid such a diversity of cultural contexts. Indeed, although 101 national versions of the Reading material were prepared for PISA 2009 in 45 languages, an in-depth examination for technical/metrical validation of instruments was only conducted for the French and English versions, acknowledging the special complexity of translations into non-Indo-European languages such as Chinese and Arabic (OECD, 2012, p. 813). As Grisay, de Jong, Gebhardt, Berezner and Halleux-Monseu (2007) point out, referring to PISA in particular and international comparison studies in general, nobody should be surprised to find that cognitive instruments are more suited in cultural and linguistic terms to western countries, which make up the majority of the participating nations. PISA 2012 (OECD, 2014b; Chapter V) maintains the double translation and metrical validation from only two source languages (English predominantly, and French). However, national versions were made available for Spanish-speaking countries looking for a strange idiosyncratic adaptation; perhaps Spanish language is flourishing like the old Latin in manifold new idioms.

Confusion between validity and reliability

The validity of the instruments used for PISA 2009, specifically in the case of Spain (see Ministerio de Educación, Cultura y Deporte [MECD], 2010) or England (Jerrim, 2011), is highly questionable, since it is based on a validation process linked to the criterion of concurrence and calculated using "the correlation between the averages of countries and autonomous regions in PISA 2009, and between these results and those of 2006" (p. 18); proudly reaching indices in excess of 0.86. This overlooks the fact that concurrent validity relative to a criterion requires a pertinent external variable, uncontaminated from one administration to another and hitherto accepted as valid, and not the anchor values of previous items[5], when these anchor items are inadequate and scarce. What PISA 2009 is actually conducting here is a measurement of reliability based on stability, also desirable for ascertaining validity, through a test-retest process (i.e. the 2006 versus the 2009 edition) on the basis of common items linking two temporally distinct applications. The "official "determination of the instruments' reliability is actually carried out using the internal consistency of the units (items), varying by domain (Reading, Mathematics and Science, both the paper and digital versions) and depending on the method, whether weighted likelihood estimates (between 0.75 and 0.85) or

plausible values (between 0.30 and 0.86) (OECD, 2012, p. 194). The plethora of reliability coefficients PISA usually offers is no guarantee of the validity of the instruments (OECD, 2012, p. 234-238).

No correspondence between contents of instruments and national curricula

The content of the PISA tests is defined through the opinion of national experts (one supervisor per country) rather than being derived from the content of national curricula; although PISA rests on a functional curricular framework (a theory of the curriculum), it does not always fully coincide with conventional curricula, which are focused on the acquisition of more academic knowledge. The implication of this emerging, subtle hidden curriculum for national educational policies is clear: let us shape our curricula to PISA's requirements; evidence of curricular changes auspices by PISA are manifold: Chisholm (2015) in Germany and South Africa; Dolin, and Krogh (2010) in a Danish context, and especially in Turkey (Gur, Celik & Ozoglu, 2012) where the educational authorities had already decided to proceed with reform of the old curriculum much before the PISA 2003 results were out, and they made use of the PISA 2003 results to justify the curriculum reform. Then it must understand that although PISA is a large international survey and it is difficult but possible and perhaps necessary to connect it with national curricula.

Added to this is the presence of biased items, identified by Olsen and Lie (2011) and Kjaernsli and Lie (2011), owing to gender and culture differences, differences between mother-tongue and language of tuition, and different teaching traditions, which are ignored, or that PISA operates with latent constructs generated by items validated in relatively small samples.

No guaranty of standardized administration

A general issue associated with testing and as such related to PISA is the use of open questions which also raises problems of low reliability, the need for agreement among correctors/coders due to difficulties in correcting highly varied answers and the discretionary time assigned to the task during the exam itself. Contrarily, PISA also tends to use closed multiple-choice questions (mostly with four choices, sometimes five), but for these there is no certainty as to the deduction made for chance in the correct answers; in other words, it is not revealed whether Lafourcade's (1971) well-known correction is applied to the final cumulative score for such items.

There are no sufficient guarantees of standardization in the administration of the tests, since a degree of opacity persists in the reports, which do not specify whether the administrators and correctors/coders of the tests are external or internal to the centers, although the materials are sent to the person in charge of the center (principal), who does not give classes to the school's 15 year-old students. This detail may be crucial, and although PISA (OECD, 2012, p. 24) supposedly guarantees standardization in the administration, capturing and processing of data, doubts arise about whether the data generated can be comparable between countries given the diversity and multiplicity of settings; PISA allows participating countries the opportunity "to adapt certain questions or procedures to suit local circumstances, and to add optional components that are unique to a particular national context" (OECD, 2012, p. 148) and the majority of countries omit certain items and administer other items and booklets (OECD, 2012, Table 12.8; p. 195-197).

The recent use of computerized administration, likely to be widely adopted in PISA 2015, only succeeds in sowing confusion, since it is no longer possible to determine whether performance reflects the ability of a student in an area of competence or a lack of access to, or command of, computers; something along these lines may have occurred with the Spanish results in PISA 2012 for problem solving (OECD, 2014a, p. 1).

The opportunist use of transformed scores for standardization

PISA basically provides scores that have been transformed via a standardized transformation scale of the kind $x_t = 500 + 100\ x_i$; where x_t is the typical CEEB (College Entrance Examination Board) transformed score; but it does not supply direct scores, or totals classified by groups or any associated standard deviation. The utility of the CEEB scores lies in the fact that, like all standardized scores, they enable comparisons to be made irrespective of the various sizes of the measurement instruments.

The generation of transformed scores causes differences to be magnified and by extension prompts a false variance between scores. One consequence of that transformation is the identification of allegedly major differences that do not in fact exist, since they are purely metrical artifacts.

The attempt in PISA 2009 Reading to ascribe eight quasi-classificatory levels ranging from lower to higher performance (*<1b*, *1b*, *1a*, *2*, *3*, *4*, *5* and *6*) or the usual seven levels in Mathematics and Science (*<1*, *1*, *2*, *3*, *4*, *5*, *6*), and to an even greater extent the rankings of countries and regions, calculated on the basis of these levels, leaves unanswered, in the light of subsequent evidence, the question of what aggregate score or level is desirable for identifying the level attained by a particular student or population sub-group; notwithstanding, it is freely assumed that: "students who attain performance level 2 in Reading demonstrate the minimum degree of competence needed for subsequent learning and social and working life (MECD, 2010, p. 61). Whether the level agreed upon is level 2 or the more desirable and convincing level 3, it is difficult to assign competence descriptors to the appropriate level.

PISA surreptitiously tries to combine two types of evaluation: one relative to a norm provided by group performance and another supposedly relative to the judgment of experts, which ranks the quality of cognitive processes previously modeled as being of high or low cognitive requirement and subsequently ranked by performance levels (1 to 6), whereas in fact these levels are generated by norm-relative values. Thus, in the case of Spain, the intervals of transformed scores linked to each PISA 2009 Reading level are: 1b [262-334]; 1a [335-406]; 2[407-479]; 3[480-552]; 4[553-625]; 5[626-707]; 6[≥708].

Reverential confidence in statistical significance as a data analysis technique

PISA has gradually included the significance (in the purely statistical rather than the substantive sense) of inter- and intra-country differences along with other population sub-groups and moderating variables, on the basis of variance components in truly Fisherian style. Similarly, the differences between successive three-yearly editions continue relying on the "sacrosanct level of salvation" of $\alpha = 0.05$, taking no account of the power of the test and the good-enough effect size; there is no evidence of "how much difference makes the difference" (Fernandez-Cano, 2009; p. 101-102). It is, after all, easy to obtain statistically- significant differences; we just need to increase the sample size, which is already high in large-scale surveys such as PISA.

It is astonishing that such reports, generated by such expensive program, continue to be dragged down by the simplicity and irrelevance of adhering to the use of gross difference in average scores merely because that difference is statistically significant, according to correlational or inferential tests using a significance statistic which is not always stated (whether *z*, *F*, *t* or χ^2).

The significance of the results is in all cases entirely statistical, and is omnipresent in the various reports. By equating statistical significance with substantive significance, the basic research underlying this program is corrupted, as are the applied research it generates and the studies derived from PISA, which continue to place their reverential trust in statistical significance.

Absence of substantive significance statistics as magnitudes of the effect

It must remember that from its fourth edition, the American Psychological Association Publication Manual (1990) emphasized that *p* values are not acceptable indices of effect and `encouraged' effect-sizes reporting.

The manifold inter-group comparisons set out in PISA should at least have used a statistic for the size of the parametric effect calculated on the basis of the difference between standardized means (i. e. Glass' ubiquitous *d*, 1977) or some other statistic of a correlational nature such as the R_I intraclass correlation coefficient, which determines the percentage of variance explained; for conceptualizations and formulae regarding effect magnitude, see Fernandez-Cano (2009, p. 99-124).

Rather than the old trick of obtaining statistical significance by increasing the size of the groups being compared, we need criteria-based values to indicate how much difference makes the difference, and to determine the appropriate sample size by considering both the level of significance (α) and the power of the test (1-β). A strange value difference of nine points in the CEEB scale (500; 100) could be considered significant even though supported by no evidence and certainly by no author, except perhaps Wikipedia (2014): "… a difference of nine points is sufficient to be considered significant" [sic].

On the basis of the PISA data, however, certain effect sizes can be calculated. Lynn and Mikk (2009), using data from PISA 2000, 2003 and 2006 and data from PIRLS 2001 and 2006, reported an average sex effect size ($d = 0.42$) in favor of girl students in Reading; this figure is similar to the 0.44 obtained from PISA 2009 data by Reilly (2012), while in Mathematics male students performed better than female students with $d = 0.22$.

Problematic presentation of findings

PISA reports tend to be excessively technical-economic, written with a bureaucratic logic and full of extremely lengthy tables and graphs; in example, the 2012 Technical reports (OECD, 2014b) exposes 282 tables and 75 figures. PISA abounds in the arguably excessive use of rankings and various statistics that are barely intelligible even for a qualified reader from the educational field, thereby leaving its findings open to misinterpretation. In this line, Berliner (2015) comments that trying to understand PISA is analogous to the parable of the blind men and the elephant where the most important of all the issues associated with PISA is discussed, namely the interpretation of scores across nations. Takayama (2008) spoke about non-contextualized "PISA international league tables". Lee (2014) clarifies that publicly released rankings in PISA change to some extent when the rankings are reevaluated by taking other factors into consideration; so then, this implication should be allowed for in interpreting the results of international assessment and the relative rankings of participating countries.

Questionable implications of findings for educational policies and practice

Attempts have been made to draw conclusions from the PISA findings in order to steer educational policies and regulations. However, it is difficult to draw consistent evaluative inferences covering an entire school system using cross-sectional data limited to the level of the 15 year-old students supposedly sampled by PISA. Its counterpart organization, NAEP, works with three cohorts: fourth, eighth and twelfth-grade students, a wider and more diverse population.

Attention is also drawn to the perverse distortion of aggregation (the ecological fallacy) in interpretations that label all the subjects, even the able ones, as included in the most disadvantaged group and especially the consideration of current moral panics surrounding the underachievement of boys (Smith, 2009). Similarly, comparisons between national education systems are strained.

Despite clamors, blown up out of proportion, in the media and in political circles, some authors (i. e. Yore, Anderson and Chiu, 2010) argue that the strictly educational impact of PISA on curriculum development, teacher training, the specific evaluation of students, teachers and schools and educational policies has been limited, being restricted merely as a basic goal to policy analyses[6].

Conclusions and recommendations

PISA is undoubtedly an evaluative undertaking that has generated a wealth of research but PISA needs to exercise greater methodological rigor, and state its methods clearly in future technical reports, above all in the national reports issued periodically. PISA needs to pay increasing attention to the development of longitudinal data, something that is acknowledged as a target in PISA 2012. To this end, before drawing up comparisons between two applications (OECD, 2013a, p. 8) it should identify longitudinal trends such as time-series in order to chart the evolution of educational systems against a comparative international background; but this would have to be on condition that the instruments used in any given edition were equal or comparable to those used in earlier editions, rather than—as has hitherto been the case—discarding some items, introducing new items and modifying existing ones.

PISA is only an international evaluative macro-study; it is therefore essential to exercise the utmost caution in drawing unfounded conclusions, supposedly derived from it, with regard to its total lack of diagnostic value in terms of the individual performance of students, curricular changes, non-contextualized reductionist comparisons or teaching practices. Notwithstanding, PISA has powerful and stimulating impact with remarkable positive consequences; therefore its study merits deeper and sustained investigation of meta-evaluative character. A way to contribute to the improvement of a project is to hear their criticisms which are not always easy to formulate.

References

Adams, R. J. (2003). Response to 'Cautions on OECD's recent educational survey (PISA)'. *Oxford Review of Education*, *29*(3), 377-389. doi: http://dx.doi.org/10.1080/0305498032000120319

Adams, R., Berezner, A. & Jakubowski, M. (2010). *Analysis of PISA 2006 preferred items ranking using the percent correct method*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2006/44919855.pdf

Adams, R. J., Wu, M. L. & Carstensen, C. H. (2007). Application of multivariate Rasch models in international large-scale educational assessments. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 271–280). New York: Springer.

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: APA.

Bank, V. (2012). On OECD policies and the pitfalls in economy-driven education: The case of Germany. *Journal of Curriculum Studies*, *44*(2), 193-210. doi: http://dx.doi.org/10.1080/00220272.2011.639903

Berliner, D. C. (2011). The context for interpreting PISA results in the USA. Negativism, chauvinism, misunderstanding, and the potential to distort the educational systems of nations. In M. Pereira, H-G. Kotthoff & R. Cowen (Eds.), *PISA under examination: Changing knowledge, changing tests, and changing schools* (pp. 77-96). Rotterdam: Sense Publishers.

Berliner, D. C. (2015). The many facets of PISA. *Teachers College Record*, *117*(1), 20.

Bollen, K., Paxton., P. & Morishima, R. (2005). Assessing international evaluations. An example from USAID's democracy and governance program. *American Journal of Evaluation*, *26*(2), 189-203. doi: http://dx.doi.org/10.1177/1098214005275640

Brown, G., Micklewright, J., Schnepf, S. V. & Waldmann, R. (2007). International surveys of educational achievement: How robust are the findings? *Journal of the Royal Statistical Society Series A-Statistics in Society*, *170*(3), 623-646. doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00439.x

Carnoy, M. & Rothstein, R. (2013). *International tests show achievement gaps in all countries, with big gains for U.S. disadvantaged students*. Economic Policy Institute: Washington, DC. Retrieved from http://www.epi.org/blog/international-tests-achievement-gaps-gains-american-students/

Couso, D. (2009). Y después de PISA, ¿qué? Propuestas para desarrollar la competencia científica en el aula de ciencias [And after PISA, what? Proposals to develop the scientific competence in the Science classroom]. *Enseñanza de las Ciencias*, especial issue, 3547-3550.

DeSeCo Project (2008). *Definition and selection of competencies: Theoretical and conceptual foundations*. Paris: OECD. Retrieved from http://www.deseco.admin.ch/

De Witte, K. & Kortelainen, M. (2013). What explains the performance of students in a heterogeneous environment? Conditional efficiency estimation with continuous and discrete environmental variables. *Applied Economics*, *45*(17), 2401-2412. doi: http://dx.doi.org/10.1080/00036846.2012.665602

Dolin, J. & Krogh, L. B. (2010). The relevance and consequences of PISA science in a Danish context. *International Journal of Science and Mathematics Education*, *8*(3), 565-592. doi: http://dx.doi.org/10.1007/s10763-010-9207-6

Drechsel, B., Carstensen, C. & Prenzel, M. (2011). The role of content and context in PISA interest scales: A study of the embedded interest items in the PISA 2006 science assessment. *International Journal of Science Education*, *33*(1), 73-95. doi: http://dx.doi.org/10.1080/09500693.2011.518646

Ehmke, T., Drechsel, B. & Carstensen, C. H. (2008). Klassenwiederholen in PISA-I-Plus: Was lernen sitzenbleiber in mathematik. [Grade repetition in PISA-I-Plus: What do students who repeat a class learn in mathematics?]. *Zeitschrift fur Erziehungswissenschaft*, *11*(3), 368-387. doi: http://dx.doi.org/10.1007/s11618-008-0033

Ercikan, K., Roth, W-M. & Asil, M. (2015). Cautions about inferences from international assessments: The case of PISA 2009. *Teacher College Records*, *117*(1), 1-28.

Fernández-Cano, A. & Fernández-Guerrero, IM. (2009). *Crítica y alternativas a la significación estadística en el contraste de hipótesis*. Colección Cuadernos de Estadística, nº 37. Madrid: Arco Libros-La Muralla.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*(1), 351-379.

Grisay, A., De Jong., Gebhardt, E., Berezner, A. & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249-266

Gur, B. S., Celik, Z. & Ozoglu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy*, *27*(1), 1-21. doi: http://dx.doi.org/10.1080/02680939.2011.595509

Hanberger, A. (2014). What PISA intends to and can possibly achieve: A critical programme theory analysis. *European Educational Research Journal*, *13*(2), 167-180. doi: http://dx.doi.org/10.2304/eerj.2014.13.2.167

Hartig, J. & Frey, A. (2012). Validity of a standard-based test for mathematical competencies. Relations with the competencies assessed in PISA and variance between schools and school tracks.

Diagnostica, *58*(1), 3-14. doi: http://dx.doi.org/10.1026/0012-1924/a000064

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E. & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation*, *17*(6), 497-509. doi: http://dx.doi.org/10.1080/13803611.2011.632668

IEA (2012). *The International Association for the Evaluation of Educational Achievement*. Retrieved from http://www.iea.nl/

Jerrim, J. (2011). *England's" plummeting" PISA test scores between 2000 and 2009: Is the performance of our secondary school pupils really in relative decline* (Nº. 11-09). London: Department of Quantitative Social Science-Institute of Education of University of London.

Judkins, D. R. (1990). Fay's method of variance estimation. *Journal of Official Statistics*, *6*, 223-239.

Kjaernsli, M. & Lie, S. (2011). Students' preference for science careers: International comparisons based on PISA 2006. *International Journal of Science Education*, *33*(1), 121-144. doi: http://dx.doi.org/10.1080/09500693.2011.518642

Knipprath, H. (2010). What PISA tells us about the quality and inequality of Japanese Education in Mathematics and Science. *International Journal of Science and Mathematics Education*, *8*(3), 389-408.

Kreiner, S. & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*(2), 210-231. doi: http://dx.doi.org/10.1007/s11336-013-9347-z

Kubinger, K. D., Hohensinn, C., Hofer, S, Khorramdel, L, Freborta, M, Holocher-Ertl, S., Reif, M. & Sonnleitner, P. (2011). Designing the test booklets for Rasch model calibration in a large-scale assessment with reference to numerous moderator variables and several ability dimensions. *Educational Research and Evaluation*, *17*(6), 483-495. doi: http://dx.doi.org/10.1080/13803611.2011.632666

Lafourcade, P. (1971). *Evaluación de los aprendizajes* [Learning evaluation]. Buenos Aires: Kapelusz.

Lee, J. (2014). An attempt to reinterpret student learning outcomes: A cross-national comparative study. *Peabody Journal of Education*, *89*(1), 106-122. doi: http://dx.doi.org/10.1080/0161956X.2014.862476

Lu, Y. & Bolt, D. M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-scale Assessments in Education*, *3*(2). doi: http://dx.doi.org/10.1186/s40536-015-0012-0

Lynn, R. & Mikk, J. (2009). Sex differences in reading achievement. *Trames-Journal of the Humanities and Social Sciences*, *13*(1), 3-13. doi: http://dx.doi.org/10.3176/tr.2009.1.01

Meyer, H-D. & Benavot, A. (Eds.). (2013). *PISA, power, and policy. The emergence of global educational governance*. Providence, RI: Symposium Books.

Ministerio de Educación, Cultura y Deporte [MECD] (2010). *PISA 2009. Programa para la Evaluación Internacional de los Alumnos. OCDE. Informe español* [PISA 2009.The Spanish report]. Madrid: Instituto de Evaluación. Retrieved from http://www.educacion.gob.es/dctm/ministerio/horizontales/prensa/notas/2010/20101207-pisa2009-informe-espanol.pdf?documentId=0901e72b806ea35a

National Center for Education Statistics (2012). *National Assessment of Educational Progress*. Retrieved from http://nces.ed.gov/nationsreportcard/

Olsen, RV. & Lie, S. (2011). Profiles of students' interest in science issues around the world: Analysis of data from PISA 2006.

International Journal of Science Education, *33*(1), 97-120. doi: http://dx.doi.org/10.1080/09500693.2011.518638

Organisation for Economic Co-operation and Development (2009). *PISA 2009 key findings*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2009/pisa2009keyfindings.htm

Organisation for Economic Co-operation and Development (2012). *PISA 2009 Technical report*, PISA. Paris: OECD Publishing. doi: http://dx.doi.org/10.1787/9789264167872-en

Organisation for Economic Co-operation and Development (2013a). *PISA 2012 Results in focus. What 15-year-olds know and what they can do with what they know*. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf

Organisation for Economic Co-operation and Development (2013b). *PISA 2012 results*. What make schools successful? Resources, policies and practices. Vol. 4. Paris: OECD. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-IV.pdf

Organisation for Economic Co-operation and Development (2014a). *SPAIN – Country note –Results from* PISA *2012 problem solving*. Retrieved from http://www.oecd.org/spain/PISA-2012-PS-results-eng-SPAIN.pdf

Organisation for Economic Co-operation and Development (2014b). *PISA 2012 technical report*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Organisation for Economic Co-operation and Development (2015a). *School governance, assessments and accountability*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/keyfindings/Vol4Ch4.pdf

Organisation for Economic Co-operation and Development (2015b). *PISA 2012 results*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results.htm

Prais, S. J. (2003). Cautions on OECD'S recent educational survey (PISA). *Oxford Review of Education*, *29*(2), 139-163. doi: http://dx.doi.org/10.1080/0305498032000080657

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *Plos One*, *7*(7). doi: http://dx.doi.org/10.1371/journal.pone.0039904

Rutkowski, L. (2014). Sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education*, *27*(2), 115-132. doi: http://dx.doi.org/10.1080/08957347.2014.880440

Rychen, D. S. & Salganik, L. H. (2003). *Key competencies for successful life and a well-functioning society*. Göttinga: Hogrefe & Huber.

Scriven, M. (2011). *Evaluating evaluations:A meta/evaluation checklist*. (6th ed.). Retrieved from http://michaelscriven.info/images/EvaluatingEvals-Checklist.pdf

Smith, E. (2009). Underachievement, failing youth and moral panics. *Evaluation & Research in Education*, *23*(1), 37-49.

Strietholt, R., Rosén, M. & Bos, W. (2013). A correction model for differences in the sample compositions: the degree of comparability as a function of age and schooling. *Large-scale Assessments in Education*, *1*(1). doi: http://dx.doi.org/10.1186/2196-0739-1-1

Stufflebeam, D. (2011). Meta-evaluation. *Journal of MultiDisciplinary Evaluation*, *7*(15), 99-158.

Takayama, K. (2008). The politics of international league Tables: PISA in Japan's

achievement crisis debate. *Comparative Education*, *44*(4), 387-407. doi: http://dx.doi.org/10.1080/03050060802481413

Wang, M.C., Haertel, G.D. & Walberg, H.J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, *63*(3), 249-294. doi: http://dx.doi.org/10.3102/00346543063003249

Wikipedia (2014). *Informe* PISA [PISA Report]. Retrieved from http://es.wikipedia.org/wiki/Informe_PISA

Yarbrough, D.B., Shulha, L.M., Hopson, R.K. & Caruthers, F.A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Yore, L. D., Anderson, J. O. & Chiu, M. H. (2010). Moving PISA results into the policy arena: Perspectives on knowledge transfer for future considerations and preparations. *International Journal of Science and Mathematics Education*, *8*(3), 593-609.

Zuckerman, G. A., Kovaleva, G. S. & Kuznetsova, M. I. (2013). Between PIRLS and PISA: The advancement of reading literacy in a 10-15-year-old cohort. *Learning and Individual Differences*, *26*, 64-73. doi: http://dx.doi.org/10.1016/j.lindif.2013.05.001

Notes

1. Specifically Latin American countries (Mexico, Argentina, Brazil and Uruguay), Egypt or Thailand whose compulsory schooling ends at 14 years. Other countries where, despite the end of schooling is postponed to 15 or 16 years, child labor and high drop-out rates are usual.

2. The limitations of resampling techniques are manifold. Three are set out here: First, they represent a poor and potentially confusing substitute for the real thing since, even if these analyses are based on independent subsamples extracted from the original sample, they will always be restricted to the characteristics of the original sample; therefore they are not a substitute for a true replication. They do not take into account potential hidden deviations and biases associated with all transversal studies, based on a single sample of participants generating a population for which certain blindly-assumed assumptions are made. They also yield inflated evaluations because they operate with small samples of data that are poorly representative and highly dependent.

3. The present author would have opted for bootstrapping; considering so many configurations of subjects, in which the same case may be represented various times or none at all, it would be possible to indicate to what extent the results are stable and generalizable via different types of subject. Moreover, parametricity assumptions do not need to be verified in the generating sample, either if the interval measure is maintained and/or if it is not possible to assume any sort of model of population distribution, since it is feasible to carry out non-parametric bootstrapping.

4. It does not seem appropriate to consider NAEP a "counterpart" to PISA because they are not affiliated and serve different purposes. Notwithstanding, PISA could learn from NAEP its longitudinal design and curricular emphasis.

5. An inverse case is set out by Hartig and Frei (2012), in which PISA 2006 data are the “criterion” variable to indicate the curricular concurrent validity of the test based on standards of mathematical competence used in the German education system. This test correlates with PISA Mathematics (r = 0.94), PISA Reading (r = 0.75) and PISA Sciences (r = 0.85). Clearly, the authors recognize the incomplete picture of the validity of this test based solely on the derived correlations.

6. PISA covers all bases (OECD, 2012) by declaring that, "PISA examines how well students are prepared to meet the challenges of the future, rather than how well they master particular curricula" (p. 3) and "it looks at their ability to use their knowledge and skills to meet real-life challenges" (p. 22). PISA induces a "standard" student considering that thorough student-level evaluations are an unnecessary damning component of a national assessment and consequently the PISA assessment does not generate scores for individuals, nor does it pretend, but the ecological fallacy is ever present in manifold report, PISA's reports include, confusing group (the students) with case (one student).

Author / Autor

To know more / Saber más

Fernandez-Cano, Antonio (afcano@ugr.es).

PhD. Chairman and Professor at the Department of Research Methods and Diagnostics in Education, Faculty of Educational Sciences, University of Granada, Spain. His interest areas are methodologies for research and evaluation, program evaluation, educational scientometrics and research. Her postal adderess is: Universidad de Granada. Facultad Ciencias de la Educación. Departamento de Métodos de Investigación y Diagnóstico Educativo. Despacho 216.1 Campus de Cartuja. 18071- Granada (España)

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]