Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Intended and unintended interpretations and uses of PISA results: A consequential validity perspective

Interpretaciones no intencionadas e intencionadas y usos de los resultados de PISA: Una perspectiva de validez consecuencial

Taut, Santy [(1)]; Palacios, Diego [(2)]

(1) Pontificia Universidad Católica de Chile (2) Pontificia Universidad Católica de Chile

Resumen

Este artículo explora la relevancia de considerar las consecuencias de las pruebas como parte de las discusiones acerca de la validez, la investigación sobre validación, en el contexto del Programa Internacional para la Evaluación de Estudiantes de la OCDE, PISA. Lo primero que describe la concepción moderna de validez como un aspecto fundamental de la calidad de los ensayos y sistemas de pruebas, es que evoluciona en torno a las interpretaciones propuestas y usos de las puntuaciones de las pruebas: "La validez se refiere al grado en el cual la evidencia y la teoría apoyan las interpretaciones de las puntuaciones de las pruebas sobre las propuestas de su uso en los test. La validez es, por tanto, la consideración más fundamental en el desarrollo y evaluación de las pruebas". (AERA, APA & SNEM, 2014, p. 11). En particular, nos centramos en el papel que han jugado sus consecuencias en la literatura sobre validez de la prueba y validación. Así como a continuación, introducimos PISA y sus interpretaciones y usos previstos como base para el examen de su validez. Esto es seguido por un resumen de los estudios empíricos existentes sobre los usos y consecuencias de PISA. Finalmente, el documento presenta piezas que faltan en la evidencia de validez en relación con las consecuencias y se analiza la importancia de una agenda pro-activa en estos temas por parte de los grupos de interés de PISA a nivel internacional y nacional

Fecha de recepción
1 Abril 2016

Fecha de aprobación
22 Junio 2016

Fecha de publicación
22 Junio 2016

Palabras clave:

PISA; validez; usos puntuaciones de test; validación; construcción de pruebas

Abstract

This paper explores the relevance of considering the consequences of testing as part of discussions about the validity, and validation research, in the context of the OECD Programme for International Student Assessment, PISA. We first describe the modern conception of validity as a core aspect of quality of tests and testing systems, evolving around the proposed interpretations and uses of test scores: “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests.” (AERA, APA & NCME, 2014, p. 11). In particular, we focus on the role that consequences have played in the literature on test validity and validation. We then introduce PISA and its intended interpretations and uses as the basis for examining its validity. This is followed by summarizing existing empirical studies on the uses and consequences of PISA. Finally, the paper presents missing pieces in the validity evidence related to consequences and discusses the importance of a pro-active agenda on these topics by the PISA stakeholders at international and national levels.

Reception Date
2016 April 1

Approval Date
2016 June 22

Publication Date:
2016 June 22

Keywords:

PISA; validity; uses of test scores; validation; developing tests.

Corresponding author / Autor de contacto

Taut, Sandy. Pontificia Universidad Católica de Chile, Escuela de Psicología, Centro de Medición MIDE UC, Avda Vicuña Mackenna 4860, Macul, Santiago (Chile). staut@uc.cl

Consequential validity: An introduction

The Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014), hereafter referred to as "the Standards," represent a professional consensus on criteria for judging the quality of educational and psychological tests in the United States. Although they were developed in the United States, they have become an influential reference used by the measurement communities outside of the U.S. as well. The Standards define three central aspects of quality testing: validity, reliability/precision and errors of measurement, and fairness. Among these three, validity occupies center stage and will be the focus of this paper. Specifically, the paper focuses on one particular type of validity evidence the Standards recommend to include in test validation: evidence based on consequences of tests.

Among other sources of validity evidence the Standards mention consequences and distinguish between consequences that follow directly from "interpretations and uses of test scores intended by test developers", "claims made about test use that are not directly based on test score interpretations" and "consequences that are unintended" (2014, pp. 19-20). Those interpretations and uses intended by the test developer must be validated by the test developer, providing relevant theoretical and/or empirical evidence for each one. However, the Standards point out that "evidence about consequences is relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components. Evidence that cannot be so traced is not relevant to the validity of the intended interpretations of test scores" (p. 21) (also see Standard 1.25, pp. 30-31). Therefore, following Messick (1989), test invalidity occurs only when consequences are due to flaws in the test, but not if they are external to features of the test. Cronbach (1988), however, argued for a more central role of consequences in test developers' obligations, if not under the umbrella of validation studies, then as another kind of (social) obligation in order to evaluate the legitimacy of test use. Much in line with Cronbach's (1988) orientation, a group of international scholars charged with reviewing validation evidence for the Mexican student achievement testing systems ENLACE and Excale derived a set of criteria of good practice in validation, somewhat similar to the Standards (AERA, APA & NCME, 2014) but formulated in a more concrete, operational style (Martinez Rizo et al., 2015). The criteria related to consequential validity are found in Appendix A.

In cases where the test user proposes interpretations and uses that differ from those supported by the test developer, then the user has the responsibility to present the corresponding validity evidence (AERA, APA & NCME, 2014, p. 13). Validity evidence on consequences must be presented for the assumptions that underlie the theory of action for each specific use of the test scores. For example, if the claim was made that PISA scores could be used to monitor the impact of a curricular reform in a certain country, then evidence must be produced that PISA is a valid measure of curricular reform impact in that country. This, however, is not the responsibility of the test developer but of the national policy makers who use the test for this purpose. However, if the claim was made that PISA results, in connection with other educational indicators measured through PISA questionnaires, could be used to describe the extent of educational equity achieved by a certain country – which corresponds to a claim actually made by the OECD, the PISA test developer – then the test developer must present evidence that PISA (as including the questionnaires and the tests) in fact offers valid information on educational equity in the participating countries (see AERA, APA & NCME, 2014, p. 20).

Michael Kane (2006, 2013) further developed such an argument-based approach to validation. In this approach, validation corresponds to evaluating an interpretive

argument that consists of a set of assumptions that must be met so that intended interpretations and uses are valid. With regard to consequences, Kane points out that many testing programs have moved beyond the "traditional monitoring role, to the use of testing as the engine of reform and accountability in education" (2006, p. 55). This objective makes them subject to validation that must be similar to program evaluation of educational interventions, which includes intended and unintended outcomes. In this conception of validation, according to Kane test developers must take on the examination of intended semantic interpretations of test scores, as well as test uses that they explicitly or implicitly recommend (Shepard, 1997), while test users must play a large role in examining the actual consequences of test use in their respective context, population, and procedures. Users of testing programs must explicate the underlying assumptions leading to desired consequences of test use, and these assumptions should be credible across different stakeholder groups. Then, test users must evaluate these consequences much as other new educational interventions are evaluated, in terms of their effectiveness, cost and benefit, weighing positive and negative consequences against each other (Kane, 2006). Linn (1998) includes the measurement community – in a role he refers to as "test evaluators" – among those who have responsibility in this regard.

However, there are also those who have questioned whether consequential validity should in fact be discussed at all as part of validity and validation. For example, Popham (1997) argued that the consequences of testing were important and should be examined by both test developers and test users, but it should not be considered part of validity. To avoid confusion by test users, these two concepts – validity and consequences – should remain clearly separated. Likewise, Mehrens (1997) posits that consequences of testing are beyond the scope of the term "validity": "Let us reserve the term for determining the accuracy of inferences about (and understanding of) the characteristic being assessed, not the efficacy of actions following assessment." (p. 18).

The issue of consequences as validity evidence is further complicated by the fact that "different decision makers may make different value judgments about the impact of consequences on test use." (Kane, 2006, p. 21) (also see Mehrens, 1997). This means that there is often no agreement among test users, and no objective truth, regarding what are considered unintended uses and consequences and what can be legitimate uses that go beyond the intended uses proposed by the test developer.

Today, in many societies the interpretations, uses and consequences of educational testing receive strong attention by educational stakeholders, including teachers and parents. Likewise, the consequential aspect of validity in scholarly writings and Standards frameworks now seems here to stay, so test developers must include its investigation in their validation efforts, at least as far as intended interpretations and uses are concerned. But not only test developers have a key role in this regard: researchers can support their efforts by studying uses and consequences of testing in different policy contexts. Likewise, policy makers and other test users must develop their own capacity to fully understand the intended and unintended interpretations and uses of test scores, and they should better resist the political pressures that surround testing and test use in many countries, often resulting in misuses. This also applies to the uses of PISA results, which have served the political agendas of education authorities and opposition forces, as well as particular interest groups, to support a variety of arguments in a way not always justified or appropriate from a validity point of view or as intended by the test developer.

What do PISA results look like?

The OECD Programme for International Student Assessment, PISA, is an international comparative student assessment instrument implemented by the Organisation for

Economic Cooperation and Development (OECD) every three years, starting in the year 2000. It assesses 15-year-olds' knowledge and skills in areas including mathematics, reading, science and problem-solving, depending on the year of implementation. About two-thirds of the test – which takes two hours to complete – contains open-ended questions asking children to apply their knowledge and skills to novel problems, while one-third of the test contains multiple-choice items. The assessment takes a literacy perspective and focuses on the ability to use knowledge and skills to meet real-life challenges, instead of the mastery of specific school curricula. In 2012, over 500,000 students in 65 countries and economies completed the assessment, with a particular focus on mathematics. PISA also includes questionnaires for students and school principals.

PISA results are communicated in a variety of formats, with "researchers, policy makers, educators, parents and students" being its main audiences (see OECD 2014, p. 5). First of all, the OECD publishes the international PISA results in a number of different formats, including full reports, policy briefs, databases, videos, and presentations. The main international results are published in a 50-page report that synthesizes the results and policy lessons based on the data analyses performed by the OECD. This report contains descriptions of the main findings, tables showing results by country listed in descending order based on their PISA scores, graphical displays of descriptive and correlational analyses, and policy-relevant interpretations of these analyses such as "Nurturing top performance and tackling low performance need not be mutually exclusive" or "The gender gap in student performance can be narrowed considerably as both boys and girls in all countries and economies show that they can succeed in all three subjects" (OECD 2014, p. 9).

In addition, a large number of thematic reports analyze specific topics, such as factors related to low performance across countries, possible origins of gender differences, educational policies characterizing high performing or strongly improving countries, among others. The OECD itself also offers detailed analyses of particular countries (for PISA 2012 these included the United States, Korea and Japan). Finally, each participating country with its academic community often publishes more in-depth country-level and comparative results in books, articles, reports and media releases (see, for example, Instituto Nacional de Evaluación Educativa, 2014; Instituto Nacional para la Evaluación de la Educación, 2013; Ministerio de Educación de Chile, 2012, 2014; Prenzel, Sälzer, Klieme & Köller, 2013).

Intended interpretations of PISA results from the standpoint of the test developers

Validity is closely tied to the purposes a test is set out to fulfill and the intended interpretations of the scores a testing system produces. In fact, the Standards state, "Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use" (AERA, APA & NCME, 2014, p. 11). Therefore, it is necessary that we give attention to these intended interpretations in the context of PISA. As some researchers have shown, educational stakeholders diverge in their perceptions of the purposes of a test and its intended uses, which makes it difficult to assume a single set of intended interpretations that can guide the way toward appropriate uses (Linn, 1998; Taut et al., 2010). However, according to the Standards (AERA, APA & NCME, 2014), the *test developer* has the main responsibility in communicating clearly and publicly what are the intended interpretations that the particular test is designed to offer (see Standards 1.1., 1.2, 1.5 and 1.6), and can be held responsible for presenting evidence that these intended interpretations and uses are in fact valid, and that claims underlying them count with empirical backing. In the case of PISA, the Organization for Economic Cooperation and Development (OECD) takes

this role, and the remainder of this section presents what key documents and internet pages communicate in this regard.

Communication materials by the OECD, for example the current flyer on PISA, stress the intended interpretations and uses for participating countries' *educational policy making*, in terms of (a) whether the education system equips young people with important skills for life, also in comparison to other countries; (b) whether the education system is fair; (c) whether young people have potential for lifelong learning (motivation, self-beliefs, etc.); and (d) how performance evolves over time, setting policy targets and assessing the impact of education policy decisions. A video prominently placed on the PISA website explains that the aims of PISA are to show countries where they stand, also in relation to other countries, in how effectively they educate their children, and to track their progress over time. It points out that success in education also includes how equally distributed educational achievement is in relation to students' backgrounds. Finally, the video talks about analyzing characteristics of successful education systems, showing what is possible, as well as similarities and differences across countries, thus helping countries review their education policies and designing new and better ones. Finally, the Frequently Asked Questions section of the PISA website states that PISA responds to "member countries' demands for regular and reliable data on the knowledge and skills of their students and the performance of their education systems", which allows them "to track their progress in meeting key learning goals."

The PISA 2012 Technical Report (OECD, 2014, p. 24) states that when linking student achievement data with contextual information from the questionnaires, PISA provides information to analyze differences between countries regarding the following topics:

- relationships between student-level factors and achievement;
- relationship between school-level factors and achievement;
- proportion of variation between and within schools;
- extent to which schools moderate the relationship between individual-level factors and achievement;
- relationship between education systems and national context and achievement;
- changes in the above-mentioned relationships over time.

In addition to score points on a scale with a mean of 500 and a standard deviation of 100, PISA also reports results in terms of six proficiency levels, in an interest to be able to describe students' literacy in a more meaningful way. Summary descriptions of the six proficiency levels on the mathematical literacy scale, and each subscale, are provided in the technical report (OECD, 2014, pp. 297-301).

The PISA 2012 Technical Report (OECD, 2014) is not structured according to the Standards for Educational and Psychological Testing (2014) but instead, is based on specifically developed Technical Standards for PISA 2012 (see Annex F). According to these standards, "valid cross-national inferences" (p. 447) depend on consistency, precision, generalizability, and timeliness. The standards are further divided into data standards, management standards and national involvement standards.

Although only tangentially related to the main topic of this article, we should mention that cross-cultural validity is one type of validity evidence that is explicitly addressed in the PISA 2012 Technical Report. The national involvement standards are said to "ensure that the internationally developed instruments are widely examined for cross-national, cross-cultural and cross-linguistic validity" (OECD, 2014, p. 448). Furthermore, the cross-cultural comparability of measures in the PISA Context Questionnaires is given special attention. The report points out that "cross-cultural differences in response styles have been considered to represent a serious source of bias in international surveys that use Likert

items" (OECD, 2014, p. 53), and further explains that PISA 2012 strove to address this threat to validity by introducing new item formats (anchoring vignettes, signal detection debiasing based on the overclaiming technique, forced choice items, and Situational Judgment Tests). This issue is elaborated in more detail in the chapter about scaling procedures and construct validation of context questionnaire data.

In summary, PISA intends its results (both test-based and questionnaire-based) to be used in at least three distinguishable ways by its main audience, namely, participating countries'*educational policy-makers*:

(1) as *diagnostic information at country level* (in terms of proficiency in tested areas, equity of the education system, other factors at individual, school and system levels that are related to learning outcomes);

(2) for *comparisons over time within each country* (which allows to track progress and evaluate the impact of policy decisions);

(3) for *comparisons with other countries* (for benchmarking and to learn from their successes and failures).

It is thus important to stress that both test scores *and questionnaire-based measurement of relevant constructs* must be included in validating intended interpretations and uses of PISA as an international testing program.

Unintended interpretations of PISA results from the standpoint of the test developers

Besides communicating intended interpretations of test scores, test developers have the duty to explicitly warn against unintended interpretations and uses in case it is possible due to prior experience to anticipate and proactively address them (AERA, APA & NCME, 2014, p. 19). Again, we present what the OECD as the test developer communicates in key documents and internet pages in this regard. The PISA 2012 Technical Report contains no mention of possible unintended interpretations. However, the PISA explanatory video previously mentioned does mention unintended uses on two occasions: (a) PISA does not say "This policy created this effect"; (b) PISA does not aim to create competition among systems by ranking them in terms of PISA performance. In the context of the rankings, the Frequently Asked Questions on the PISA website clarify that it is not possible to assign a single exact rank in each subject to each participating country or jurisdiction. It further states that there is statistical uncertainty involved in sampling students and extrapolating to a population, and that "it is therefore only possible to report the range of positions (upper rank and lower rank) within which a country can be placed. For example, in PISA 2003 Finland and Korea were widely reported as ranking 1st and 2nd in PISA, when in fact we can only say that, among OECD countries, Finland's rank was between 1st and 3rd and Korea's was between 1st and 4th." Further with regard to the use of rankings and in response to an open letter from Heinz-Dieter Meyer and Katie Zahedi (Meyer & Zahedi, 2014) the OECD points out, "less than 1% of the PISA reporting is devoted to league tables. The view of the OECD is that it should be up to individual countries to decide to what extent they wish to be compared internationally …" (see PISA website, FAQ section).

Despite these initial efforts, it still seems fair to say that in OECD communication products about PISA little can be found about unwarranted interpretations and potentially harmful uses of PISA results in participating countries and at international level, the only exception being the rankings.

Existing empirical evidence on the consequential validity of PISA

The literature[i] about uses and consequences of PISA is mainly focused on the three first PISA waves (2000, 2003 and 2006). This body of literature includes some studies that have revised the policy effects of PISA in different countries, especially in European countries. For instance, Baird and

collegues (2011) reviewed the policy response to PISA in six case countries/regions, where high-performing participants (Canada and Shanghai-China) were contrasted with European countries that generally performed towards the average (England, France, Norway and Switzerland), but in which there had been interesting policy impacts of PISA. Another example was an external evaluation of the policy impact of PISA commissioned by the PISA Governing Board (OECD, 2008), which used both quantitative and qualitative approaches to evaluate the relevance, effectiveness and sustainability, as well as unexpected impacts, of PISA. In the quantitative strand, a set of stakeholders (policy makers, local government officials, school principals, parents, academics, and media representatives) of 43 countries and economies were surveyed via email. In the qualitative strand, different stakeholder groups took part in interviews and focus groups. This study also included case studies in five countries and economies (Canada, Hong Kong-China, Norway, Poland and Spain), considering their differences in terms of PISA performance and policy impact, equity, and government structure. Finally, a few studies (Breakspear, 2012; Martens, Nagel, Windzio, & Weymann, 2010) interviewed representatives and experts from OECD countries, and also analyzed policy documents, in order to highlight the diverse national responses to the release of PISA results.

These studies highlighted a series of findings: (a) the impact of PISA was greater at the national level than at regional or school level; (b) policy-makers were identified as the most significant stakeholder group; (c) countries increasingly valued the skills assessed in PISA; (d) PISA was commonly used for monitoring a country's performance as well as equity; (e) PISA's influence on policy seemed to be increasing over time; (f) PISA has the potential to 'define' challenges in educational policy and set the agenda for policy debate at the national and state levels; (g) the majority of countries initiated some kind of policy reform or initiative - to varying extent, mostly depending on their level of performance - in direct response to PISA at some point across the survey rounds (Baird et al., 2011; Breakspear, 2012; OECD, 2008).Based on the previously mentioned literature as well as additional sources, evidence of PISA's consequential validity was organized with respect to the intended and unintended interpretations and uses listed in the preceding sections.

Based on the previously mentioned literature as well as additional sources, evidence of PISA's consequential validity was organized with respect to the intended and unintended interpretations and uses listed in the preceding sections.

(1a) Diagnostic information at country level

In this group, we will focus on the examples of France and Germany. In the French case, its reading literacy performance was consistently around the OECD average (Urteaga, 2010). However, the wide spread of students' score distribution in reading literacy aroused the concern of policy makers and politicians. In other words, the results showed that there are large proportions of students in the top band of performance as well as in the lowest band, and, moreover, this pattern was more pronounced in 2009 than in 2000. In response, the French government announced a series of reforms in the primary school curriculum such as introducing a strategy to fight illiteracy, but also personalized learning assistance throughout the system in order to help lower-achievement students, complemented by more school autonomy with schools being able to manage their own budgets (Baird et al., 2011).

German results in PISA 2000 showed that German students' science achievement was significantly below the OECD average, and their reading and mathematics performance was similar. These results were much lower than expected ('PISA-Schock') and thus represented devastating news for the German education system, previously considered one of the best in the world. As a result, important changes were introduced in the German education system. First of all, changes in the

political discourse were accompanied with a wide-ranging reform agenda including a number of initiatives (e.g. programs to improve instructional quality and increased funding for schools), but most importantly, the introduction of the National Educational Standards (NES). The core of the reform was the notion of skills and competencies over the traditional German notion of education. For example, the NES describe scientific competencies students are expected to have acquired at the end of their lower secondary education (Neumann, Fischer, & Kauertz, 2010). Additionally, in terms of curriculum development processes, outcome control and external assessment acquired additional relevance. Finally, the academic discourse changed its direction towards a greater emphasis on the empirical research of pedagogic practices (Ertl, 2006).

(1b) Comparisons over time within each country

In this category we find the cases of Poland and Hong Kong (China). In the first case, the significant gains in Poland's PISA scores have been related to two significant improvement periods (2000-2003 and 2009-2012) (Amoroso et al., 2015). This positive scenario, especially the improvement in the 2000-2003 period, has been attributed to educational reforms in Poland during the 1990s, although separating the precise impact of each reform element is difficult. This reform was designed to increase educational opportunities for all students, but also to advance the educational system inherited from the communist period. The mentioned reform included structural changes such as delaying the selection between general and vocational tracks by one year, and introducing three years of comprehensive, mandatory lower-secondary education (gimnazjum), with positive effects on achievement results (Jakubowski, 2015). One suggested hypothesis about the association between the Polish educational reform and its improvement in PISA is that the changes in curriculum and student assessment led to improvement of cognitive skills, where pupils tested in 2012 had already completed three years of lower secondary education under the new curriculum (Amoroso et al., 2015; Jakubowski, 2015). In summary, although PISA was not the driver of change in Poland's education policies, it was used as a monitoring tool for checking the progress of students' scores in terms of policy impacts.

As in the case of Poland, in Hong Kong (China) PISA 2003 and 2006 results were not identified as the key driver for reforms, but served as an important monitoring tool, with improved performance being attributed to the educational reform program (OECD, 2008). This reform included a new curriculum in 2002 and later a new student tracking system. Furthermore, the OECD study (2008) suggests that Hong Kong used PISA as the guide for the construction of its new educational objectives, moving away from encouraging students to acquire subject knowledge towards developing their comprehension, problem solving, reasoning and strategic thinking. Finally, Hong Kong introduced a broad range of achievement testing (at grades 6, 9 and for 15-year-olds) and international benchmarking regarding school drop out rates, upper secondary completion rates and life-long learning participation.

(1c) Comparisons with other countries

An important consequence of international studies such as PISA is that politicians and policy makers have to respond to their country's position in league tables. This is particularly relevant when a country's performance is worse than expected, either by sliding down the league table or by doing worse than neighboring countries (Stobart & Eggen, 2012). An interesting example of the latter is the case of Norway. In 2000 and 2003, Norwegian PISA results were below the OECD average and, importantly, also below its Scandinavian neighbors (Sweden, Denmark and Finland), despite a well-funded and self-confident education system (OECD, 2008). Consequently, the unfavorable comparison with its peers on PISA has had a significant impact on educational policy in Norway,

leading to a series of reforms in terms of both assessment and curriculum policies (Baird et al., 2011; Chung, 2016; Elstad, 2010). The 'Norwegian PISA shock' became a driving force for reforming the education system, which included changes at both primary and secondary levels: a new curriculum with more emphasis on measurable outcomes; comprehensive government projects promoting formative assessment; a new national quality assessment system with national testing; new regulations for examinations and teacher reporting of overall grades (Tveit, 2013).

(2) Unintended consequences

Based on survey responses from different stakeholder groups from 43 countries and economies, the OECD's (2008) external evaluation described the following unexpected effects of PISA, both positive and negative. In terms of positive unexpected impacts, findings include particularly high levels of public interest and debate in light of PISA results; more value assigned to the skills evaluated by PISA and aligning national assessments in this regard; increased collaboration between stakeholder groups for improving their country's results and education system; and increased interest in empirical educational research. On the negative side, the report mentions national discussions that seek to place responsibility for poor performance on particular groups (e.g., teachers), resulting in a 'culture of blame;' and use of PISA for legitimation of educational reforms that would otherwise be more openly discussed and contested.

In terms of country-specific evidence we found references about Turkey, Japan, Spain and Chile. In the case of Turkey, its results in PISA 2003 and 2006 were lower than the OECD average. The reaction of educators, policy makers, and journalists was to focus on the poor performance in comparison to other countries (league tables). For instance, newspapers were mostly interested in rankings, thus ignoring other relevant information that the PISA results revealed and what they imply for improving the educational system (Gür, Çelik, & Özoğlu, 2012). Gür and colleagues (2012) examined public documents (e.g. official reports and news bulletins published by the Education Ministry) and concluded that the authorities had already decided to introduce a new education reform much before the PISA 2003 results were published. However, government officials used PISA results to justify the need for a reform of the education system without a careful examination of the results and what they mean for the system as a whole.

Similarly, although Japan was a top performer in PISA 2000, its PISA 2003 results were interpreted by the press as a downward trend, resulting in a perceived 'crisis' that encouraged significant public and political debate on education reform. In response to the decline in scores, the Japanese government reversed a contentious low-pressure curriculum policy in favor of national assessment practices (Takayama, 2008). However, from an objective point of view the results in PISA 2003 were not statistically different from PISA 2000 in mathematics; there was only a statistically significant decline in reading literacy, which represents a long-identified weak point of Japanese students. Additionally, the Japanese press' interpretation about rankings did not mention that top performers in 2003 (The Netherlands and Hong Kong) had not been included in PISA 2000 (Takayama, 2008).

Likewise, in 2013 Spain introduced a series of educational reforms explicitly inspired and justified by poor PISA results in 2012 (Choi & Jerrim, 2015; OECD, 2014). Particularly, the latest and most important initiative was the Organic Law for the Improvement of Educational Quality (LOMCE, in Spanish), which included initiatives such as greater autonomy for schools, new preventive diagnostic testing in primary education (year 6), more vocational pathways starting in the final years of lower secondary education, and exit exams in lower and upper secondary education (OECD, 2014). However, some

authors have argued that the interpretations of PISA results used to legitimate this reform have been incorrect, in particular related to Spain's ranking position and comparisons with European neighbors (Choi & Jerrim, 2015; Bonal & Tarabini, 2013). This has been true for politicians (Jornet, 2013, 2016c) and the press (Carabaña, 2008). According to these authors, this highlights the negative consequences that come from the exclusive and inaccurate use of rankings for educational policy making (Jornet, 2016a, 2016b).

Finally, Chile experimented a significant improvement in reading literacy during the period 2000-2009, especially among low-income students, thus reducing the socio-economic achievement gap (OECD, 2010). Nevertheless, educational policy makers and the press chose to ignore this improvement and exclusively highlighted the glass half empty, emphasizing Chile's below-average ranking position compared to mainly OECD countries and stressed the need for further educational reform (Ravela, 2011). Even though in international comparisons over time Chile is one of the strongest improving countries during the past decade, at national level what should have been good news was misinterpreted for political purposes.

In summary, according to the literature focusing on different case examples, PISA results triggered significant mobilizing reactions and intended uses in some countries. These generally seem to be countries where PISA diagnosed a performance that stayed strongly below national expectations – which could be based either on prior high performance, or ambitious aspirations not met, or unfavorable comparison with "peer countries". Sometimes, PISA outcomes on different subjects were complemented by information on the distribution across performance categories and indicators of educational equity within these countries. Also, trends over time served as monitoring information to gauge progress regarding certain kinds of skills and topics. In terms of unintended uses, these seem to be undertaken by national political players, for political reasons, sometimes supported by the media, and mostly related to the use of PISA rankings to generate a sense of urgency and to legitimate educational reforms, establishing direct causal links between certain national policies and PISA results, and unjust comparisons. This has had negative consequences in terms of diagnosing failures of educational policies on the one hand, or pushing through unwarranted reforms, on the other hand.

Pending research on the consequential validity of PISA

According to the Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014), the test developer – in this case, the OECD – is responsible for presenting "logical or theoretical arguments and empirical evidence" (p. 24) that support those interpretations and uses of PISA results that the test developer explicitly or implicitly suggests. The Standards go so far as to include any indirect benefit that is anticipated on the grounds of the testing program (see Standard 1.6). These intended interpretations and uses correspond to the ones presented in the respective section above. Although the Technical Report does contain a wealth of information that is relevant to building a validity argument for PISA, the report does not present this information organized according to the validity claims, or intended interpretations and uses, nor does the website contain other documents that specifically address the validation of PISA, or present validity rationales for each of its intended interpretations and uses. As mentioned above, there was an external evaluation regarding the policy impact of PISA (OCDE, 2008), but this study was not technically (i.e., measurement) oriented and did not present comprehensive arguments and evidence for each intended interpretation and use. It did provide evidence of intended uses of PISA at country level, it also explicitly studied unintended effects of PISA; the results of this study have been presented above. However, comprehensive

validation research for PISA that is driven by its intended interpretations and uses, as demanded by widely accepted measurement standards (AERA, APA & NCME, 2014) seems to remain a pending task, at least according to publicly and readily available documentation on the PISA website. However, PISA does devote considerable attention to cross-cultural validity, an aspect not included in other standards frameworks cited above.

Likewise, although some caution is provided as to the use of rankings, information on unintended or unsupported interpretations and uses is virtually absent from the PISA website and the documentation contained therein. As the Standards (AERA, APA & NCME, 2014) point out, the test developer cannot be held responsible for unintended uses and consequences unless they are due to flaws in the test itself (construct underrepresentation or construct-irrelevant variance). Additionally, local PISA coordinating institutions bear more responsibility in this regard, but even they cannot prevent policy makers and the media from reaching unsound conclusions based on PISA results. However, as also stated in the recommendations of the external evaluation of PISA (OECD, 2008), the OECD and its partners could do more to call attention to unsound practices, provide support for developing assessment literacy, and be active in promoting intended uses while being transparent about, and cautioning against, unintended ones.

This scenario regarding pending validation is surprising given the substantial financial resources PISA requires and the high level of technical expertise that is regularly involved in its development and analysis. Participating countries may demand such evidence in the future (for examples, see Martinez Rizo et al., 2015; Schafer, Wang & Wang, 2009; Taut, Santelices & Stecher, 2012), and the OECD might dedicate a particular chapter in the Technical Report (and section on the website) to presenting evidence supporting proposed interpretations and uses, as well as in other, equally visible places, cautioning against unintended or unsupported ones.

Conclusions

"Validity theory is rich, but the practice of validation is often impoverished" (Brennan, 2006, p. 8). This conclusion, so often stated in educational measurement circles, also seems to apply to the PISA testing system, and particularly to the consequential aspect of validity. Above all else, test developers must be held responsible for validating intended interpretations, uses and consequences, and such documentation must be publicly available in a timely manner. Such evidence can hardly ever be complete and definitive, but an explicit effort involving substantial resources should be visible.

Test developers have decreasing responsibility as test use keeps moving farther away from the test scores (and accompanying questionnaire data) they produce. While they can be held accountable for the kinds of interpretations they support, as well as for clearly communicating what should and should not be uses of test scores, actually preventing inappropriate uses and negative consequences is clearly out of their sphere of influence. However, test developers can play an advocacy role in educating assessment users in appropriate test use and to call public attention to foreseeable misinterpretations, and actual cases of data misuse. It is beyond the scope of this paper to judge how much PISA has done, and how much responsibility it holds, to prevent any misuses that have occurred at national level in the past. In fact, the external evaluation of PISA's intended and unintended policy impact (OECD, 2008) included two final recommendations in this regard: (a) "At a minimum, PISA should produce guidelines of dissemination for those who participate in the program"; and (b) "PISA should consider, at a minimum, the creation of a policy group for countries that request its advice on policy formation and better use of the PISA results" (p. 9). A good example constitute the guidelines for uses of PISA-based tests for schools (OECD, 2013).

In any case, national players such as testing agencies, Ministries and academics also play a key role in exerting influence so that PISA results are adequately interpreted and used in their respective countries.

Referencias

American Educational Research Association, American Psychological Association & National Council for Measurement in Education [AERA, APA & NCME] (2014). *The Standards for Educational and Psychological Testing*. *Washington*, D.C.: AERA.

Amoroso, J. M., Moreno, J. M., Gortazar, L., Herrera Sosa, K. M., Kutner, D., & Bodewig, C. (2015). *Poland - Skilling up the next generation : an analysis of Poland's performance in the program for international student assessment*, 1–21. Retrieved from http://documents.worldbank.org/curated/en/2015/12/25518729/poland-skilling-up-next-generation-analysis-poland%E2%80%99s-performance-program-international-student-assessment

Baird, J.-A., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *Policy effects of PISA*. Retrieved from http://research-information.bristol.ac.uk/en/publications/policy-effects-of-pisa(833739c4-7e0a-4c18-b249-a3f12120065f).html

Bonal, X., & Tarabini, A. (2013). The role of PISA in shaping hegemonic educational discourses, policies and practices: The case of Spain. *Research in Comparative and International Education*, *8*(3), 335–341. http://dx.doi.org/10.2304/rcie.2013.8.3.335

Breakspear, S. (2012). The policy impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance. *OECD Journals*, (71), 1–32. DOI: http://dx.doi.org/10.1787/19939019

Brennan, R. (2006). Perspectives on the Evolution and Future of Educational Measurement. In R. Brennan (ed.), *Educational Measurement*, 4th ed., pp. 1-16. Westport, CT: Praeger.

Carabaña, J. (2008). *Las diferencias entre países y regiones en las pruebas PISA*. Madrid: Colegio Libre de Eméritos

Choi, A., & Jerrim, J. (2015). The Use (and Misuse) of PISA in Guiding Policy Reform: The Case of Spain. *SSRN Electronic Journal*, 1–16. DOI: http://dx.doi.org/10.2139/ssrn.2580141

Chung, J. (2016). The (mis)use of the Finnish teacher education model: "policy-based evidence-making"? *Educational Research*, *58*(2). DOI: http://dx.doi.org/10.1080/00131881.2016.1167485

Cronbach, L. (1988). Five perspectives on the validity argument. In H. Wainer & H. Braun (eds.), *Test validity*, pp. 3-17. Hillsdale, NJ: Lawrence Erlbaum.

Elstad, E. (2010). P*isa Debates and Blame Management Among the Norwegian Educational Authorities*: Press Coverage and, *48*, 10–22.

Ertl, H. (2006). Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32*(5), 619–634. DOI: http://dx.doi.org/10.1080/03054980600976320

Gür, B. S., Çelik, Z., & Özoğlu, 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

Instituto Nacional de Evaluación Educativa. (2014). *PISA 2012 Resolución de problemas de la vida real. Resultados de matemáticas y lectura por ordenador. Informe Español. Versión preliminar*. Instituto Nacional de Evaluación Educativa. Retrieved from http://www.mecd.gob.es/dctm/inee/internacional/pisa2012-resolucionproblemas/pisaresoluciondeproblemas.pdf?documentId=0901e72b8198bee8

Instituto Nacional para la Evaluación de la Educación. (2013). *México en 2012*. Mexico: Instituto Nacional para la Evaluación de la Educación. Retrieved from http://www.inee.edu.mx/images/stories/2013/principal/PISA2013/PISA_2012041213web1.pdf

Jakubowski, M. (2015). Opening up opportunities: education reforms in Poland, (January).

Jornet, J. (2013, January 30). Cuestionados los supuestos malos datos españoles del informe Pisa. *Comunidad Valenciana*. Valencia. Retrieved from http://ccaa.elpais.com/ccaa/2013/01/30/valencia/1359572336_318312.html

Jornet, J. (2016a). *España en PISA*. Valencia: Ateneo Mercantil de Valencia.

Jornet, J. (2016b, January 26). La educación no está tan mal; el informe PISA. *Levante. El Mercantil Valenciano*, p. 10. Valencia.

Jornet, J. (2016c, January 26). Cómo desmontar el informe PISA. *Las Provincias*. Valencia. Retrieved from http://www.lasprovincias.es/comunitat/20160126/como-desmontar-informe-pisa-20160126001834-v.html

Kane, M. (2006). Validity. In Brennan, R. (ed.), *Educational Measurement*, 4th ed., pp. 17-64. Westport, CT: Praeger.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448-457.

Linn, R. (1998). Partitioning responsibility for the evaluation of the consequences of use. *Educational Measurement: Issues and Practice*, 17(2), 28-30.

Martens, D. K., Nagel, A.-K., Windzio, M., & Weymann, A. (2010). *Transformation of Education Policy*. Basingstoke: Palgrave.

Martinez Rizo, F. et al. (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.

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

Messick, S. (1989). Validity. In R. Linn (ed.), *Educational measurement* (3rd ed.), pp. 13-103. Washington, D.C.: American Council on Education.

Meyer, H.-D. & Zahedi, K. (2014, May 4). *Open Letter to Andreas Schleicher*, OECD, Paris. Retrieved from http://www.globalpolicyjournal.com/blog/05/05/2014/open-letter-andreas-schleicher-oecd-paris

Ministerio de Educación de Chile. (2012). *Evidencias para Políticas Públicas en Educación: Selección de Investigaciones Concurso Extraordinario FONIDE-PISA*. Santiago de Chile: Ministerio de Educación de Chile. Retrieved from https://s3.amazonaws.com/archivos.agenciaeducacion.cl/documentos-web/Estudios+Internacionales/PISA/Evidencias_para_Politicas_Publicas_en_Educacion_FONIDE_PISA.pdf

Ministerio de Educación de Chile. (2014). *Informe Nacional Resultados Chile Pisa 2012*. Santiago de Chile: MINEDUC. Retrieved from https://s3.amazonaws.com/archivos.agenciaeducacion.cl/documentos-web/Estudios+Internacionales/PISA/Informe_Nacional_Resultados_Chile_PISA_2012.pdf

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: the impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, *8*(3), 545–563. http://doi.org/10.1007/s10763-010-9206-7

Organisation for Economic Co-Operation and Development (2008). *External evaluation of the policy impact of PISA*, (November), 3–5. Retrieved from http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/PISA/GB(2008)35/REV1&docLanguage=En

Organisation for Economic Co-Operation and Development (2010). PISA 2009 Results: Executive Summary. *Executive Summary*, 1–

21. Retrieved from http://www.oecd.org/pisa/pisaproducts/46619703.pdf

Organisation for Economic Co-Operation and Development (2013). *General guidelines for the availability and uses of the PISA-based test for schools*. Retrieved on May 15, 2016 from https://www.oecd.org/pisa/aboutpisa/PISA-based-test-for-schools-guidelines.pdf

Organisation for Economic Co-Operation and Development (2014, April). *Education Policy Outlook Spain*. Retrieved May 1, 2016, from http://www.oecd.org/edu/EDUCATION%20POLICY%20OUTLOOK%20SPAIN_EN.pdf

Organisation for Economic Co-Operation and Development (2014). *PISA 2012 Technical Repor*t. Retrieved on May 1, 2016, from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

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

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (eds.) (2013). *PISA 2012. Fortschritte und Herausforderungen in Deutschland (PISA 2012. Improvements and challenges in Germany)*. Münster: Waxmann.

Ravela, P. (2011). ¿Qué hacer con los resultados de PISA en América Latina? *PREAL. Programa de Promoción de La Reforma Educativa En América Latina y El Caribe*, *58*.

Schafer, W., Wang; J. & Wang, V. (2009). Validity in action: State assessment validity evidence for compliance with NCLB. In R. Lissitz (ed.), *The concept of validity*, pp. 173-193. Charlotte, NC: Information Age Publishing.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-8.

Stobart, G., & Eggen, T. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 1–6. DOI: http://dx.doi.org/10.1080/0969594X.2012.639191

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

Taut, S., Santelices, V. & Stecher, B. (2012). Validation of a national teacher assessment and improvement system. *Educational Assessment Journal*, *17*(4), 163-199. DOI: http://dx.doi.org/10.1080/10627197.2012.735913

Taut, S., Santelices, V., Araya, C. & Manzi, J. (2010). Theory unterlying a national teacher evaluation program. *Evaluation and Program Planning*, *33*, 477-489. DOI: http://dx.doi.org/10.1016/j.evalprogplan.2010.01.002

Tveit, S. (2013). Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 221–237. DOI: http://dx.doi.org/10.1080/0969594X.2013.830079

Urteaga, E. (2010). Los resultados del estudio PISA en Francia. *Revista Complutense de Educación*, *21*, 231–244.

Notes

[i] The present paper used various search tools and databases, including Web of Science, Scopus, ScienceDirect and Google Scholar to search for potentially relevant studies. The keywords for this search were *PISA* and its combinations with the following terms*: interpretations, effects, uses, consequences, decision making, validity, validation, consequential validity*. Next, we included the combination of *PISA results* with the same set of terms. Additionally, we repeated the same processes in Spanish.

Autores / Authors | **To know more / Saber más**

Taut, Sandy (staut@uc.cl).
Degree in Psychology from the University of Cologne (Germany) and her Ph.D. in Education from the University of California Los Angeles (UCLA, USA). She is an assistant professor at the School of Psychology at Pontificia Universidad Católica de Chile and an associate researcher at the Measurement Center MIDE UC. She is de Corresponding autor for this article. Her address is: Pontificia Universidad Católica de Chile, Escuela de Psicología, Centro de Medición MIDE UC, Avda Vicuña Mackenna 4860, Macul, Santiago (Chile).

Palacios, Diego (dfpalaci@uc.cl).
Associate researcher at the Measurement Center MIDE UC. at Pontificia Universidad Católica de Chile. His postal address is: Pontificia Universidad Católica de Chile, Escuela de Psicología, Centro de Medición MIDE UC, Avda Vicuña Mackenna 4860, Macul, Santiago (Chile).

RELIEVE

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

[ISSN: 1134-4032]