Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Impact of large-scale tests in contexts of weak technical tradition: Experience of Mexico and the Iberoamerican Group for PISA

Impacto de las pruebas en gran escala en contextos de débil tradición técnica: Experiencia de México y el Grupo Iberoamericano de PISA

Martínez-Rizo, Felipe
Universidad Autónoma de Aguascalientes

Resumen

El artículo inicia con un breve repaso del desarrollo de la psicometría, que destaca la brecha entre la situación de Estados Unidos y la de otros países, sobre todo de menor desarrollo, en cuanto a la existencia de personal calificado en temas técnicos complejos. En seguida se hacen consideraciones sobre la noción de validez y su importancia, en especial en cuanto a la validez de consecuencias. Luego se describe el impacto de las evaluaciones, con especial atención a las de gran escala y las internacionales, con el caso de PISA, considerando en especial el caso de países con una tradición psicométrica débil. En la conclusión se discute el tema, a partir de la experiencia de México y del Grupo Iberoamericano de PISA.

Fecha de recepción
23 Abril 2016

Fecha de aprobación
22 Junio 2016

Fecha de publicación
23 Junio 2016

Palabras clave:

Pruebas en gran escala; Psicometría; Validez; Validez de consecuencias; PISA.

Abstract

The paper begins with a brief review of the development of psychometry, which highlights the gap between the situation in the United States and in other countries, especially less developed ones, as to the existence of qualified personnel in complex technical issues. Next, considerations are made on the notion of validity and its importance, especially concerning consequential validity. The impact of the tests is then described, with special attention to those of large and international scales, with the case of PISA, and mainly concerning countries with weak psychometric tradition. In the conclusion the whole issue is discussed, from the experience of Mexico and the Latin American Group of PISA.

Reception Date
2016 April 23

Approval Date
2016 June 22

Publication Date:
2016 June 23

Keywords:

Large-scale tests; Psychometry; Validity; Consequential Validity; PISA

THE UNEQUAL DEVELOPMENT OF PSYCHOMETRY

The atipical situation in the United States

The discipline specialized in measurement in human sciences was developed in the United States since the end of the 19th century and throughout much of the 20th century. Cattell invented the word *test* with the 1890 book *Mental Tests & Measurements*. The intelligence tests that Binet developed in France were adapted by Terman in Stanford (1917) and were extended to be used by the American Armed Forces. (De Landsheere, 1996: 56-71). In New York 1904, Edward L. Thorndike published his seminal book *An Introduction to Theory of Mental and Social Measurement*. (Martinez-Arias, 1995).

The *College Board* (*College Entrance Examination Board*) was created in 1900 in order to rationalize the selection and admission processes in northeastern American universities by means of essay questions. Since it was difficult to accurately, rapidly and

Corresponding autor / Autor de contacto
Martínez-Rizo, Felipe. Universidad Autónoma de Aguascalientes (Mexico): felipemartinez.rizo@gmail.com

effectively grade these kind of questions, the College Board developed "objective" standardized tests with multiple choice questions. The *Scholastic Aptitude Test* (SAT) was administered for the first time in 1926. An important advancement in standardized tests since 1941 was the equalization of different test forms for the stability of tests results over time (Donlon, 1984).

In 1948, Princeton University constitued the *Educational Testing Service* (ETS). In the following decade, the University of Iowa also introduced a standardized test, the *American College Test* (ACT). (De Landsheere, 1996:150, note 4). By 1950, with the publication of Gulliksen *Theory of Mental Tests*, Classical Test Theory was considered complete. (Martínez Arias, 1995)

In the second half of 20th century, the process continued with an increasing number of tests administered every year. This trend was connected with the concern for the quality of the American school system brought to light by the Coleman Report (1966) and the decline in average SAT results. As early as 1957, the launch of Sputnik served as undeniable evidence that the USSR was superior to the United States in the Space Race because Russians had better scientists and engineers, and better education in mathematics and science. (Mathison & Ross, 2008). Several states made regular assessment of students' educational progress an obligation, by means of tests designed to demonstrate a minimum level of performance. In 1982, 42 out of 50 states had *minimum competencies tests*, but they were frequently deficient and did not meet expectations. As a result, there were many lawsuits that questioned them for beeing discriminatory, biased and unreliable (Baker & Choppin, 1990).

In 1969, US government started the *National Assessment of Education Progress* (NAEP) to evaluate the quality of education at a national level. In 1983, NAEP entrusted its operation to ETS (Walberg, 1990). In the same year, the report *A Nation at Risk* was published at the request of President Ronald Reagan. With this report, the *educational standards movement* began (Mathison & Ross, 2008). In 1989 at the Charlottesville educative summit, the 50 states governors endorsed a common goal stating that by the year 2000, American students would rank first in the world in the 4th, 8th and 12th grades with high competency levels in English, Mathematics, Science, History, and Geography (Mathison & Ross, 2008).

In 2002, President George W. Bush signed the *No Child Left Behind* (NCLB) Act, enforcing all American federal states to have education standards and create assessment systems that include annual English, Math, and Sciences tests for all students from the 4th to 8th grade. The participation in NAEP tests became mandatory in order to have access to federal funds for programs to improve educational quality. With the NCLB Act, tests became high stakes since their results were used as criteria to decide whether a school would receive federal funds or needed to be closed. With the law that substituted NCLB (Every Student Succeeds Act, ESSA), signed by President Barack Obama on December 10th, 2015, the weight attributed to tests has been reduced, but it is too early to assess its impact.

The situation in other countries and the international assessments

Until World War II, there were no advances in psychometry outside the United States. The difference was so large that, in 1931 when Thorndike heard participants at an international conference considering tests as typically American, he reacted by saying that *for the sake of science and our well-being, it is better that tests are not called American.* (Joncich, 1968, De Landsheere, 1986: 68, nota 24)

In the second half of 20th century, the changes of societies and education systems, as well as psychometric methodology advances, brought rapid diffusion of large-scale tests. NAEP served as a reference for systems for monitoring educational quality in countries like Australia (ACER) and Holland (CITO).

Events that increased concern for educational quality in the US, such as Sputnik, contributed also to the development of international tests. As the differences in curriculum and traditional ways of assessing learning progress prevented comparing student achievement in different countries, a group led by Torsten Husén proposed in 1958 the development of a test that would provide comparable results at international level. In 1959, the group organized a pilot study, and in 1964 the first mathematics study was conducted. In 1966, the *International Association for the Evaluation of Educational Achievement* (IEA) was created, which completed other studies in the sixties and seventies. In the 1980s, the organization conducted a second mathematics study and others about sciences and writing. Until the mid-1990s, the IEA conducted other studies, specifically the Third International Mathematics & Science Study (TIMSS). Later on, IEA conducted studies on civic education and other areas. TIMSS (now *Trends in Mathematics and Science Study*) adopted a four-year cycle, and the reading study a five-year cycle (*Progress in International Reading Literacy Study*, PIRLS). (De Landsheere, 1994; Husén & Neville-Postlethwaite, 1996; Postlethwaite, 1985).

In 2000, another well-known international project started, promoted by the Organization for Cooperation and Economic Development (OECD): *Program for Institutional Student Assessment* (PISA).

Today, large-scale assessments exist in almost all the European Union and other highly developed countries like Japan, South Korea, Singapore and Israel. They began to be implemented in Arabic countries and in Africa. With support from UNESCO's International Institute for Educational Planning, there are also projects in Africa's French speaking countries and English Speaking, which form the *South African Consortium for the Monitoring of Educational Quality*, SACMEQ (Ross, 1994; SACMEQ, 1995).

Situation in Ibero-American area

Latin American countries, like other middle and low development ones, don't have a strong history with large scale testing. In the sixties, tests started to be used to determine students' access to universities. In primary education, with a few exceptions like Chile, Latin American countries did not apply large scale international and national tests until the 1990s.

In 1994, the Latin American Laboratory for the Assessment of Educational Quality (in Spanish LLECE) was established, which since that time organized three *regional comparative and explanatory studies*, whose results became known in 1997, 2008 and 2015. Starting with Chile in the eighties, most Latin American countries have implemented large-scale testing systems since the nineties. Recently, and in addition to Chile, there are census-based tests in Mexico, Brazil, Columbia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Peru, and Uruguay.

In addition to Spain and Portugal, Mexico, as an OECD member, has been active in PISA since 2000. Brazil voluntarily followed Mexico´s footsteps and later on, at different times, so did Argentina, Chile, Peru, Uruguay, Columbia, Costa Rica, Panama and Dominican Republic.

The weak tradition of large scale testing common in Ibero-American countries –Spain and Portugal, Brazil and Spanish speaking countries of Latin- America— explain that. There are only a few specialists with training to use psychometric models like those of Item Response & Generalizability Theory, as well as Hierarchical Linear & Structural Equation Models, and others, for analyzing databases with large scale tests results. There is also a shortage of people who are able to design and carry out validation processes that take into account different dimensions of validity.

QUALITY AND IMPACT OF TESTS

Validity

Until the middle of 20th century, the notion of validity focused on the prediction of a

criterion: *in a general sense, a test is valid for anything in which it is correlated with (Gilford according to Messik, 1989: 18).* Afterward, the distinction between content, criterion (predictive-concurrent) and construct validity was introduced with a growing weight of the latter, as it was considered that the construct included content and criterion validity. For Messik, validity *is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (1989: 13).* Later the focus was on the interpretation of scores obtained with a measurement instrument. Emphasis on construct validity remained, as the core of a unified concept synthetized by Cronbach's dictum: *all validation is one* (1980: 90).

In a way similar to the definition used since 1998, the most recent version of *Standards for Educational and Psychological Testing* defines validity as *the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests.* (AREA-APA-NCME, 2014: 11) This definition coincides with the vision of Messick (1989) and Kane (2006) in the sense that the validation process should focus on *interpretation* and *use* of scores obtained through a measurement tool. Validity isn´t an attribute of the tool that is used to collect the information, but rather of the interpretations and uses that are made of the results.

To validate an interpretive inference is to check the degree in which it is sustained through multiple types of evidence, while alternative inferences are less supported. To validate an action inference not only requires supporting the meaning of a certain score from a measurement tool, but also to value that action's implications and results, especially assessing the relevance and utility of scores from a test for specific purposes, such as the social consequences for using scores to support decisions. *Although there are different sources and mixes of evidence for supporting score-based inferences, validity is a unitary concept. Validity always refers to the degree to which empirical evidence and the theoretical rationale support the adequacy and appropriateness of interpretations and actios based on test scores*. (Messick, 1989: 13)

Today, there are contrasting postures, as can be seen in Borsboom, Mellenbergh and van Heerden, 2004, or in the first 2013 issue of *Journal of Educational Measurement.* A presentation by Newton (2013) lists 143 meanings for the term validity, and there are even proposals to abandon the concept.

More elaborate versions of the concept include the idea of validity as an argument (Kane, 2006, 2013). Another dimension is that of *cultural validity*, which highlights the importance of taking care since the test design and development, the way in whch different cultural, linguistic and social factors from constructs that are measured can influence the way in which subjects interpret the content of ítems and how they respond to them (cfr. Basterra, Trumbull & Solano-Flores, 2011).

Consequential Validity

Looking at the evolution of the notion of validity, with construct dimension becoming omnipresent in validation process, it is possible to see that the only source of evidence not explicitly incorporated to construct validity is that which evaluates social impact. Messick considers that it´s ironic, as far as validity was initially conceived in functional terms: how well a test works for what it was designed to do. Consequential validity appeared in 1999 AERA-APA-NCME standards and introduced a complexity that, for some people, made the idea more confusing instead of clarifying it, and gave rise to many discussions. Others point to the necessity of including the new dimension due to the fact that testing moved from psychometry to the policy arena. Therefore, the appraisal of an assessment can't be reduced to technical issues, but must also include its consequences.

Today, the most important definitions of validity, starting with 2014 version of AERA,

APA and NCME standards, include consequences--individual or social , desired or not, planned or unplanned—that come with the use of the test. (Kane, 2013; Moss, 2008; Sireci, 2013)

Furthermore, validity is a matter of degree, not all or none, and evidence strengthens or weakens through new findings. This includes that anticipation of testing possible social consequences change with new evidence about real consequences and with changing social conditions. Validity is an evolving property, and validation a continual process. (Messick, 1989: 13)

IMPACT OF ASSESSMENT, AND IN PARTICULAR OF LARGE SCALE TESTINGE

In relation to students' assessment and its consequences, and in particular of classroom assessment, Rick Stiggens points out that until recently it was considered normal that only a few students achieve learning goals, while an important number didn´t, and assessment role was only to reliably distinguish one group from another. Therefore, the criteria to assess the quality of a test were validity and reliability. Today, it is expected that schools make all students reach high competence levels to be able to live in a knowledgable society. For that reason, we must reflect on the appropriate way to assess students' achievement in this new context. Stiggens says:

> The most valid and reliable assessment in the world that has the effect of causing students to give up in hopelessness cannot be regarded as productive because it does far more harm than good. Quality control frameworks of the past have not taken into account impact on the learner. The vision of excellence in assessment framed herein places this criterion of quality at center stage. (2008: 2 and 3)

The role of large-scale tests now have implications for consequential validity, similar to those that Stiggens points out for classroom assessment. In the past, large scale tests were low stakes, as their results didn´t influence decisions for individual students, teachers or schools. In the United States this began to change in 1980, and was accentuated in 1990, to comply with NCLB Act, and large-scale tests acquired a weight without precedents in decisions about individual students, teachers, and schools.

La aplicación de la ley NCLB evidenció deficiencias y consecuencias contraproducentes, tanto para los maestros al asociarse decisiones fuertes para ellos con base en los resultados de sus alumnos, incluso con Modelos de Valor Agregado, como para las escuelas que no pudieron cumplir las metas de la ley sobre Avance Anual Adecuado (*Adequate Yearly Progress*), y debieron enfrentar consecuencias que podían llegar hasta su cierre. Las metas nada realistas de la ley NCLB promovieron prácticas fraudulentas por parte de maestros en lo individual, pero también de escuelas e incluso distritos y estados completos (Oakes & Lipton, 2007), lo que justificó los cambios de la ley ESSA.

With the implementation of NCLB, its limits and counter-productive consequences became apparent for teachers when tenure and promotion decisions were made based on their students' test results, even with Value-Added Models, as well as for schools that didn't reach the legal goals about Adequate Yearly Progress and had to face consequences that could result in closure. The unrealistic goals of NCLB promoted fraudulent practices from some teachers, but also from schools, districts and complete states (Oakes & Lipton, 2007), which justified changes of the new ESSA Act.

In many other countries something similar occurred. Massive application of tests, and the diffusion of their results by means of rankings or league tables of schools based on students' scores, without considering context, implied tests being transformed into high stakes. That impact resulted by the very fact of the ranking diffusion, even without legal dispositions about consequences for schools according to their position in the ranking. If there are such dispositions the situation is worse, but even

without that, the impact on schools, teachers, and finally on students can be very strong, as tests tend to be corrupt and negative practices appear like teaching to the test, narrowing of curriculum, or the alteration of results through more openly dishonest strategies.

People with little or no trainning in these topics ignore that the precision of any test results is limited, and therefore rankings based on them are deceitful. At education system level results can be more precise and stable, but those of individual schools can change from one application to another. Furthermore, results of a test only partially depend on schools and teachers, as family and social context factors have an important role as well. Statistical techniques to control the effect of those factors, such as *Value-Added Models*, neither have precision nor reliability necessary to sustain consistent rankings over time.

Some public opinion sectors support strong policy decissions based on census-based test results, as they believe that achievement levels will be largely improved in a short time thanks to *rankings* that, when diffused, create a competition between schools and teachers that presses them to try harder. The implicit assumption of the argument is that improving poor students' performance is an easy job, and that it is not achieved simply because of teachers' negligence or students' lack of effort, which coud be corrected easily with strong policy measures:

> Test-based accountability systems are based on the belief that public education can be improved through a simple strategy: require all students to take standardized achievement tests and attach high stakes to the tests in the form of rewards when test scores improve and sanctions when they do not. (Hamilton, Stecher & Klein, 2002, p. iii)

One extremely unfortunate consequence of overevaluating large-scale tests is to lose sight of the importantance of classroom assessments carried out daily by every teacher with his/her students. Large-scale tests, at least in the present state of psychometry technology, must be composed by multiple choice or similar items that can be graded automatically. For this reason, they can hardly measure more complex parts of the curriculum, including entire areas of it, such as oral and written expression, or attitudinal and valule-related aspects. Progress in these aspects only can be assessed with the precision and frecuency necessary to guide teaching and learning practices by means of the teachers' classroom work, that any large-scale test can substitute.

PISA possibilities and limits

PISA can contribute valuable elements to improve educational systems, by allowing comparison between countries (and regions within countries with subnational samples, like Mexico and Brazil, in addition to Spain), enabling analysis of results that consider school and social context factors, exploring levels of social equity.

With PISA data it is possible to identify two kinds of problems: those present when there is a high proportion of low performance students, mostly poor, and those that arise when there are very few students at high performance levels. In the first case the challenge is to make all youngsters have a sufficient competence level for a satisfying life as workers, but also as citizens. In the second case, the challenge is training the future elites of engineers and scientists that will occupy executive positions in business and political sectors. From 2009 PISA application, with option of applying a booklet with low-difficulty items, it is possible to know what students below Level 1 as defined by PISA scales are or are not able to do. This new option was promoted by Ibero American countries. The cognitive results, with data from context questionaires and items that explore attitudes and other non-cognitive aspects, can contribute useful elements for exploring factors that influence students' competence levels.

However, due to matrix design and sample-based application, PISA cannot provide results at a school level, and even less for individual students, but only about the whole education system. So, results are mostly relevant for

those responsible for public policies, and only indirectly for school principals and teachers, as PISA cannot offer sufficient information to guide teaching and learning practices. PISA by itself is also insufficient as a base for policy decisions, that requiere to incorporate many other context data and educational indicators, not derived from large-scale tests, to have enough information to support decision making. The work of Axel Rivas (2015) is an excellent example in this respect.

CONCLUSION

A positive consequence of PISA and large-scale tests is that they call society´s attention to the importance of undertaking efforts to improve achievement levels in schools. Another positive impact is the consolidation of national institutions specialized on assessment, whose outlook is very different today than it was 15 years ago. The technical capacity has advanced a lot in Chile, in a State agency and an university measurement center; in Mexico with the National Institute for Educational Evaluation created in 2002 and with full autonomy since 2013. In Uruguay, with a unit for measurement of educational results within the Ministry and since 2013 with its own National Institute; with consolidation of older agencies in Brazil (Institute for Studies and Educational Research Anisio Teixera, INEP) and Colombia (Institute for Higher Education Development, in 2010 Institute of Evaluation). Participation of these countries in PISA was a factor that contributed to this consolidation.

In the first round of PISA (2000), the participation of Mexico and other Ibero-American countries was limited to the basics: to translate items produced by the consortium in charge of the tests; to pilot and administer them to the minimum sample; to rate open response items, send the results to the consortium and wait for international analysis, without participating in the planning of tests and item design, nor making an analysis of national results. Around 2003, Mexico used an extended sample in order to have results for each federal state and prepared a national report that was disseminated at the same time as the international report (Vidal & Díaz, 2004), which implied training of Mexican specialists on Item Response Theory, Hierarchical Linear Models, and other workshops. Chile and Uruguay also prepared national reports, and Uruguay produced a Newsletter to disseminate particular points.

Since the start of the preparation of PISA around 2006, Mexico proposed to increase the level of participation for two reasons. First, because assessments in which participate many countries, languages, cultures and levels of development involve a high risk of cultural bias, to avoid which it is important that a country is not limited to the administration of instruments developed elsewhere, but actively engage in item design and analysis of results; and second, because participating in PISA next to countries with a stronger psychometry tradition is a learning opportunity for technical groups with less experience. Shortly thereafter, an age of intense collaboration started between technical teams in charge of PISA in Ibero-American countries, with the formation of the Ibero-American Group of PISA (GIP), lead by Mexico and Spain. (Martínez Rizo & Roca, 2009)

Initially, the collaboration consisted in sharing the translations of original versions of PISA instruments and manuals from English and French to Spanish and Portuguese. After the 2006 application, National Project Managers (NPM) supported each other for coding open response items and database cleaning. The exchange of experiences between NPM of Ibero American countries was also useful for preparing PISA 2006 national reports. With support from OECD, secretariate meetings and training workshops were organized.

Collaboration included preparing reading item units for PISA 2009, with a training workshop conducted by specialists from the consortium in charge of PISA, and with a wide exchange over several months, in which NPM from Ibero American countries exchanged

item units developed before sending them to the international consortium; as a result, for the first time PISA 2009 tests had item units developed in Ibero-America. GIP also promoted the option of applying low-difficulty units that don't decrease whole test difficulty level, nor prevent comparison with earlier rounds results. This makes it easier to accurately describe students' competencies that do not reach the lowest level measured so far by instruments than in the past. An outstanding aspect of collaboration was the preparation of an international report of PISA 2006 results in the Ibero-American countries participating, as well as in Spain autonomous communities, and in Brazil and Mexico federal states.

PISA and large-scale tests, however, can also have negative consequences. In Mexico, two large-scale tests stand out. The case of national test called ENLACE is clearest: its massive annual application to all students at the last four grades of primary, and the three of secondary school (4° to 9° of international classification CINE), and its association with important economic stimulus for teachers and advantages for schools, caused corrupt practices to proliferate, such as teaching to the test, narrowing of curriculum, and falsification of results. This led to the cancellation of the tests as of 2014.

In the case of PISA, despite the fact that its sample-based application and matrix design prevented results by individual schools and students, the tests' international visibility, and the weight of OECD, led Mexican government to establish as the main goal of Education Sector Planning for 2007-2012 term to have better results in PISA, with refers to distortion and narrowing of educational policy.

The excessive attention paid to these two tests brought consequences that have been summarized in the following points (Martínez Rizo, 2010b):

- Banalization of public debate about the quality of education, reduced to superficial discussions of rankings, losing sight of the complexity of the topic.
- Deceptive marketing of schools, mainly private, to attract students based on rankings.
- Impoverishment of curriculum, through the tendency to teach to the test, neglecting aspects that will not be assessed, even though they are important.
- Fatigue and discouragement in schools that, despite their efforts, do not achieve results that can be compared with schools in more favorable conditions, and the students' negative attitude towards an education focused on preparing them for tests.
- Impoverishment of public policies that tend to look for easy solutions to complex problems, neglecting essential aspects as fairness.

The situation of Ibero-American countries in the use of large-scale tests, and especially in relation to PISA, has been similar to that of Mexico. 15 years ago, these countries started to venture in this field, but with the exception of SIMCE in Chile, the impact of tests was low and even null due to limited dissemination of results (Martínez Rizo & Roca, 2009). In the second decade of 21st century, in addition to Chile, census-based tests were applied in Uruguay, Mexico, Brazil, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Peru, and Dominican Republic, and their impact is increasing, similar to the one described earlier for Mexico´s case. 15 years ago, the results of the few assessments were low stakes. Today tests proliferate, attract attention, and become an important reference for educational policy, but risk of practices stemming from inadequate comprehension of tests scopes and limits leads to expecting almost miraculous results for schools by the sole application of tests, without recognizing their true reaches.

The combination of positive and negative consequences of large scale tests make notion of consequential validity having special relevance. In a recent study about Mexico's main large-scale tests (Martínez Rizo, 2015), consequential validity is specified with the following points that authorities in charge of

test development and/or use of results should attend:

- Inform users about the test purpose and characteristics, specifying what may or may not be measured, and the intended uses and consequences, with theoretical arguments and empirical evidence to support both, warning about uses for which there is not sufficient evidence of validity, trying to identify the most likely.
- Report results in reasonable time to the interested parties with clear and precise language, without unnecessary technical jargon, with information to minimize the possibility of incorrect interpretations and using labels that do not stigmatize.
- Provide a normative framework for assessing the performances of the examinees; describe the profile and characteristics of the reference population.
- Support institutions and users to develop the necessary capacity for the adequate interpretation and use of results.
- Document and assess the extent to which the anticipated and/or desirable consequences of test are produced, as well as existence of unforeseen uses or consequences, whether adequate or inadequate; if there is evidence of inappropriate uses they are investigated, and if they persist, users are informed, and an attempt is made to apply corrective measures.

Even though PISA psychometric quality is high, and the technical documentation shows, in general, its scopes and limits, it seems possible to argue that consequential validity has not been sufficiently taken care of, and that there is a need to specify the five points just quoted.

It can be argued that, rather than to the technical bodies in charge of test development and analysis of results at international level, this kind of care would correspond to national technical bodies, as well as to agencies responsible of each country education system. Accepting this argument, the weakness of the specialized agencies in many medium or low development countries, with the workload that many of these agencies have, explains their limits in the care of consequential validity.

Because of previously mentioned impact of PISA in educational policy of many countries, OECD should be more careful in addressing these aspects. In countries with the greatest technical capacity, it may be less important, but in countries that, because of low development, have fewer elements to prevent or correct inappropriate uses, and where the risk of negative consequences is greater, OECD's role is much more important

In order to maximize the positive potential and minimize the negative of the powerful tools that are standardized tests, it is essential that technical competencies of specialized agencies to be consolidated, and especially to address consequential validity, involving both the authorities in charge of education systems and civil society. The international organisms that manage large-scale tests could contribute more to such consolidation.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, Authors.

Baker, E. & Choppin, B. (1990). Minimum competency testing. In Walberg, H y Haertel, H. (Eds.). *The International Encyclopedia of Educational Evaluation*. New York: Pergamon Press, pp. 499-502.

Basterra, M. Rosario, Trumbull, E. & Solano, G. (eds.) (2011). *Cultural validity in assessment: Addressing linguistic & cultural diversity*. New York: Routledge.

Borsboom, D., Mellenbergh G. J. & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061-1071. doi: http://doi.org/10.1037/0033-295X.111.4.1061

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer, & H. Braun, (Eds.), *Test validity* (pp. 3–17). Princeton: IEA.

De Landsheere, G. (1994). *Le pilotage des systèmes d'éducation*. Bruxelles: De Boeck.

De Landsheere, G. (1996). *La investigación educativa en el mundo*. Mexico: Fondo de Cultura Económica [French original edition: 1986].

Donlon, T. (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.

Ferrer, G. (2006). *Educational Assessment Systems in Latin America: Current Practice and Future Challenges*. Washington: PREAL.

Grupo de Trabajo Sobre Estándares y Evaluación. (2007-2008). Evaluaciones nacionales. In *Observatorio regional de políticas de evaluación educativa*. Santiago: PREAL.

Hamilton, L., Stecher, B. & Klein, S. (2002). *Making Sense of Test-Based Accountability in Education*. Santa Mónica: RAND.

Husén, T. & Postlethwaite, T. S. N. (1996). A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education: Principles, Policy & Practice*, *3*(2), 129-141. doi: http://doi.org/10.1080/0969594960030202

Joncich-Clifford, G. (1968). *The Sane Positivist: A Biography of Edward L. Thorndike*. Middletown: Wesleyan University Press.

Kane, M. (2006). Validation. En R. Brennan (ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport: American Council on Education & Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1): 1–73. doi: http://doi.org/10.1111/jedm.12000

Martínez-Arias, R. (1995). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Martínez-Rizo, F. (2008). Las evaluaciones educativas en América Latina. In Instituto Nacional para la Evaluación de la Educación (Coord.), *Cuadernos de investigación*. Mexico: INEE.

Martínez-Rizo, F. (2010a). Assessment practice in policy context: Latin American countries. In P.Peterson, E. Baker & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 479-485). New York: Elsevier-Academic Press.

Martínez-Rizo, F. (2010b). Usos y abusos de la evaluación. *Este País*, agosto (232), 24-27.

Martínez Rizo, F. (Coord.). (2015). *Las pruebas ENLACE y Excale. Un estudio de validación*. Mexico, INEE.

Martínez Rizo, F. & Roca, E., (Coords.) (2009). *Iberoamérica en PISA 2006*. Madrid. Santillana.

Mathison, S. & Ross, E. (2008). *The Nature and Limits of Standards-Based Reform and Assessment*. New York: Teachers College Press.

Messick, S. (1989). Validity. En R. L. Linn, ed. *Educational Measurement* (3rd ed., pp. 13-103). New York, American Council on Education & Macmillan.

Michell, Joel (2000). Normal Science, Pathological Science and Psychometrics. *Theory & Psychology*, *10*(5): 639-667. doi: http://doi.org/10.1177/0959354300105004

Moss, P. (2008). A critical review of the validity research agenda of the NBPTS at the end of its first decade. In L. Ingvarson & J. Hattie (Eds.), *Assessing teachers for professional certification: the first decade of the National Board for Professional Teaching Standards* (pp. 257–312). Oxford, Elsevier.

Newton, Paul E. (2013). *Does it matter what 'validity' means?* Presentation at Department of Education, Oxford University, February 4.

Postlethwaite, T. S. N. (1985). International Association for the Evaluation of Educational Achievement. In T. Husén & T. S. N. Postlethwaite (Eds.). *The International Encyclopedia of Education*. New York: Elsevier, pp. 2645-2646.

Rivas, Axel (2015). *América Latina después de PISA: Lecciones aprendidas de la educación en siete países (2000-2015)*. Buenos Aires: Fundación CIPPEC.

Ross, K. (1994). *The Establishment of a Southern Africa Consortium for the Monitoring of the Quality of Education*. Paris: IIEP.

Sireci, S. G. (2013). Agreeing on Validity Arguments. *Journal of Educational Measurement*, *50*(1): 99–104. doi: http://doi.org/10.1111/jedm.12005

Southern Africa Consortium for Monitoring Educational Quality (1995). *Southern Africa Consortium for Monitoring Educational Quality*. París: IIPE.

Stiggins, R. (2008). *Assessment Manifesto: A Call for the Development of Balanced Assessment Systems*. Portland: ETS-ATI.

Vidal, R. & Díaz, M. A. (2004). *Resultados de las pruebas PISA 2000 y 2003 en México. Habilidades para la vida en estudiantes de 15 años*. Mexico: INEE.

Walberg, H. (1990). National assessment of educational progress: retrospect and prospect. In H. Walberg & G. Haertel (Eds.), *The International Encyclopedia of Educational Evaluation*. (Pp. 435-440). Oxford, New York: Pergamon Press.

Wolff, L. (2004). Educational Assessments in Latin America: The State of the Art. *Applied Psychology: An International Review*, *53*(2), 192-214. doi: http://doi.org/10.1111/j.1464-0597.2004.00168.x

Author / Autor | **To know more / Saber más**

Martínez-Rizo, Felipe (felipemartinez.rizo@gmail.com).

Professor at the University Autónoma de Aguascaliente (1974-2016), where he has been Rector. He has written 57 books and over 190 articles or chapters. A member of Mexico's National System of Researchers and Mexican Academy of Sciences.The founder and first Director General of Mexico's National Institute for the Evaluation of Education from 2002 to 2008. He has an Honorary PhD by the University of Valencia (Spain).

RELIEVE

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

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]