Solano-Flores, Guillermo & Milbourn, Tamara (2016). Assessment Capacity, Cultural Validity and Consequential Validity in PISA. *RELIEVE*, *22*(1), M12. DOI: http://dx.doi.org/0.7203/relieve.22.1.8281

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Assessment Capacity, Cultural Validity and Consequential Validity in PISA

Capacidad evaluativa, validez cultural y validez consecuencial en PISA

Solano-Flores, Guillermo [(1)] **& Milbourn, Tamara** [(2)]

(1) Stanford University. (2) University of Colorado Boulder.

Abstract

International student assessments have played an increasing important role in educational policy. These international test comparisons generate valuable information about each participating country's student performance and the social and contextual factors. A complex picture of the cultural, economic, and social factors that shape PISA participation is emerging. We aim to understand the relationship between national assessment capacity and how countries participate in international test comparisons. We propose a framework for examining assessment capacity as key to addressing two aspects of validity -cultural and consequential. Also, we discuss the multiple facets of assessment capacity as conditions for addressing cultural validity and consequential validity in international test comparisons

Reception Date
2016 April 04

Approval Date
2016 June 20

Publication Date:
2016 June 21

Keywords:

PISA; assessment capacity; cultural validity; consequential validity

Resumen

Las evaluaciones internacionales de estudiantes han desempeñado un papel cada vez más importante en la política educativa. Estas comparaciones internacionales basadas en pruebas generan información valiosa sobre el rendimiento del estudiantado de cada país participante y los factores sociales y contextuales asociados. Una imagen compleja de los factores culturales, económicos y sociales que dan forma a la participación de PISA empieza a emerger. Nuestro objetivo es entender la relación entre la capacidad evaluativa nacional y la forma en que los países participan en estas comparaciones internacionales. Proponemos un marco conceptual para examinar la capacidad evaluativa como clave para abordar dos aspectos de la validez: cultural y consecuencial. Asimismo, se discuten las múltiples facetas de la capacidad evaluativa como condiciones para abordar la validez cultural y validez consecuencial en comparaciones internacionales.

Fecha de recepción
04 Abril 2016

Fecha de aprobación
20 Junio 2016

Fecha de publicación
21 Junio 2016

Palabras clave:

PISA; capacidad evaluativa; validez cultural; validez consecuencial

International student assessments, such as the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), and the Progress in International Reading Literacy Study (PIRLS) have played an increasing important role in educational policy.

These international test comparisons generate valuable information about each participating country's student performance and the social and contextual factors that may account for achievement differences (e.g., infrastructure, characteristics of the teachers and the curriculum). Potentially, this

Autor de contacto / Corresponding author

Solano-Flores, Guillermo. Stanford University Graduate School of Education. 485 Lasuen Mall. Stanford, CA 94305-3096. United States. gsolanof@stanford.edu

information can help participating jurisdictions to inform their educational policies. Indeed, the Organization for Economic Co-operation and Development (OECD)—PISA's organizing agency—encourages participating countries and partner economies to develop new education policies based on the strengths and weaknesses identified:

> PISA offers insights for education policy and practice, and helps monitor trends in students' acquisition of knowledge and skills across countries and in different demographic subgroups within each country. The findings allow policy makers around the world to gauge the knowledge and skills of students in their own countries in comparison with those in other countries, set policy targets against measurable goals achieved by other education systems, and learn from policies and practices applied elsewhere (OECD, n.d., p. 8)

While analyses of results of international test comparisons have focused on the link between student performance and factors such as the organization of the curriculum and the national expenditures on education (see Suter, 2000), more attention needs to be paid to the factors that shape how countries use information from their participation in international test comparisons. Indeed, in order for a country to properly "set policy targets against measurable goals achieved by other education systems, and learn from policies and practices applied elsewhere" (see the quote above), proper interpretations of scores need to be made according to the specific national context. The reason is simple: Policies and practices that appear to be successful (as reflected by PISA outcomes) in certain countries may not be successful or may not be easy to implement in other countries.

The impact of PISA on policies varies considerably across countries (Breakspear, 2012) and can be shaped by misinterpretation and misuse of score rankings (Ercikan, Roth, & Asil, 2015). Moreover, the impact of PISA outcomes on national policies does not necessarily warrant a corresponding impact on countries' teaching and assessment practices (Teltemann & Klieme, 2016). Clearly, a complex picture of the cultural, economic, and social factors that shape PISA participation is emerging. Understanding the interaction of these factors is critical to meeting the assessment program's goals and ensuring that countries truly benefit from participating in this international test comparison.

This paper addresses an important consideration underlying the inception of international test comparisons in the 1960s—that appropriate human and institutional resources are needed to properly participate (see Husén, 1983). We aim to understand the relationship between national assessment capacity and how countries participate in international test comparisons. Assessment capacity is especially important for Latin American countries, as many of them have engaged recently in national assessment programs and, in many cases, with scant experience with large-scale assessment programs (Ferrer, 2006; Ravela, 2001).

We first discuss how the assessment capacity of a country largely influences the extent to which it can benefit from participating in PISA and make sound interpretations and use of PISA information. Then, we propose a framework for examining assessment capacity as key to addressing two aspects of validity—cultural and consequential. The former refers to the extent to which the multiple ways in which cultural experience influences students' interpretations of test items are considered throughout the entire process of assessment development (Solano-Flores, 2011; Solano-Flores & Nelson-Barber, 2001); the latter refers to the use and consequences of inferences based on test scores (Messick, 1989; Shepard, 1997). Finally, we discuss the multiple facets of assessment capacity as conditions for addressing cultural validity and consequential validity in international test comparisons.

Assessment Capacity

The United Nations Development Programme defines capacity as "the ability of individuals, institutions and societies to perform functions, solve problems, and set and achieve objectives in a sustainable manner" (Capacity Development Group, 2007, p. 3). We adopt and expand the three aspects of assessment capacity identified by Clarke (2012) in the context of international test comparisons:

1) an enabling context that supports or is conductive to assessment activities—the extent to which a country has developed or is able to develop and use technically sound assessment instruments;

2) the alignment of assessment activities and instruments with other components of the education system—the extent to which a country has created or is able to create and sustain assessment systems, and;

3) the psychometric quality of the instruments generated—the extent to which a country is able to use the information provided by those instruments and assessment systems to inform their policies and practices.

The OECD recognizes the importance for countries to build assessment capacity in order to effectively assess student learning and the need for countries to systematically analyze their capacity to participate in international test comparisons, as "many countries may not have a good understanding of the scope of assessment operations or of the detailed project management involved (Lockheed, Prokic-Bruer, & Shadrova, 2015, p. 60). Accordingly, limited financial resources, a short history of participation in large-scale testing, an inexperienced culture of evaluation and accountability, and restricted access to skilled assessment specialists with formal psychometric training, are factors that may weaken a country's national assessment capacity (Ercikan & Solano-Flores, 2016; Solano-Flores, 2008).

We contend that a limited assessment capacity makes it difficult for a country to properly implement international assessment programs' procedures and may prevent it from obtaining the maximum benefit from its participation. An indication of the potentially serious impact of limited assessment capacity is the fact that some participating countries may not have access to a sufficient number of in-country assessment experts (Kamens & McNeely, 2010). This limited access to assessment experts may constitute a serious challenge, even in cases in which the professionals involved at the country level with international assessments also work on national assessments (Gilmore, 2005).

There is no easy way of knowing the number of assessment experts in a country, if we focus on one of the many aspects of assessment capacity. However, information from PISA 2012 and from the directory of the International Test Commission (ITC) (Illescu, personal communication, November 11, 2012) helps to appreciate that there is tremendous unbalance in the assessment expertise of countries participating in international test comparisons. Fifteen countries that participated in PISA in 2012 (24 percent) did not have any individual or organization with membership in the ITC. These numbers suggest that a number of participating countries could be lacking an adequate number of assessment experts.

Conceptual Framework of Assessment Capacity and Validity

Figure 1 shows our conceptual framework connecting assessment capacity and validity in international test comparisons. For the purposes of this paper, participation in an international test comparison can be thought of as comprising four stages:

Stage 1: Test Development. Test items are developed by participating countries and selected for their inclusion in the assessment according to format and content criteria specified by the corresponding organizing

agency in documents such as the assessment framework and a set of item specifications.

Stage 2: Test Translation. Test items are translated and adapted according to translation guidelines provided by the organizing agency with the intent to reflect the characteristics of the culture and the language used in national curricula.

Stage 3: Test Administration and Analysis of Test Results. Tests are administered and the organizing agency collects, analyzes, and reports data.

Stage 4: Use of Assessment Data. Participating countries enact new educational policies based on the analysis of their national test results, often in comparison to the performance of other countries, and presumably with the intent to identify areas of improvement in their policies and practices.

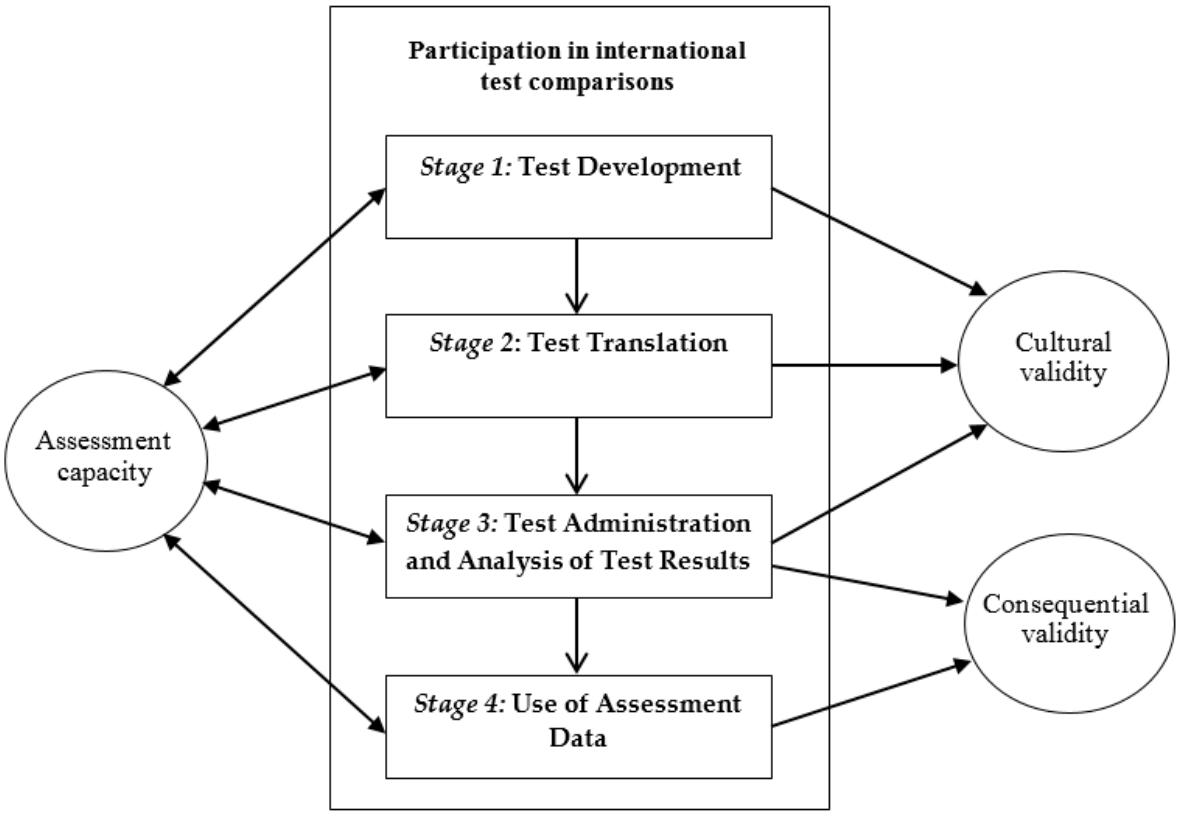


Figure 1. Framework for national assessment capacity and participation in international test comparisons

The double arrows in Figure 1 represent the synergistic relationship between a country's national assessment capacity and its participation in international test comparisons. For example, assessment capacity influences the fidelity with which a country implements PISA procedures. In turn, participating in PISA may contribute to building the country's assessment capacity.

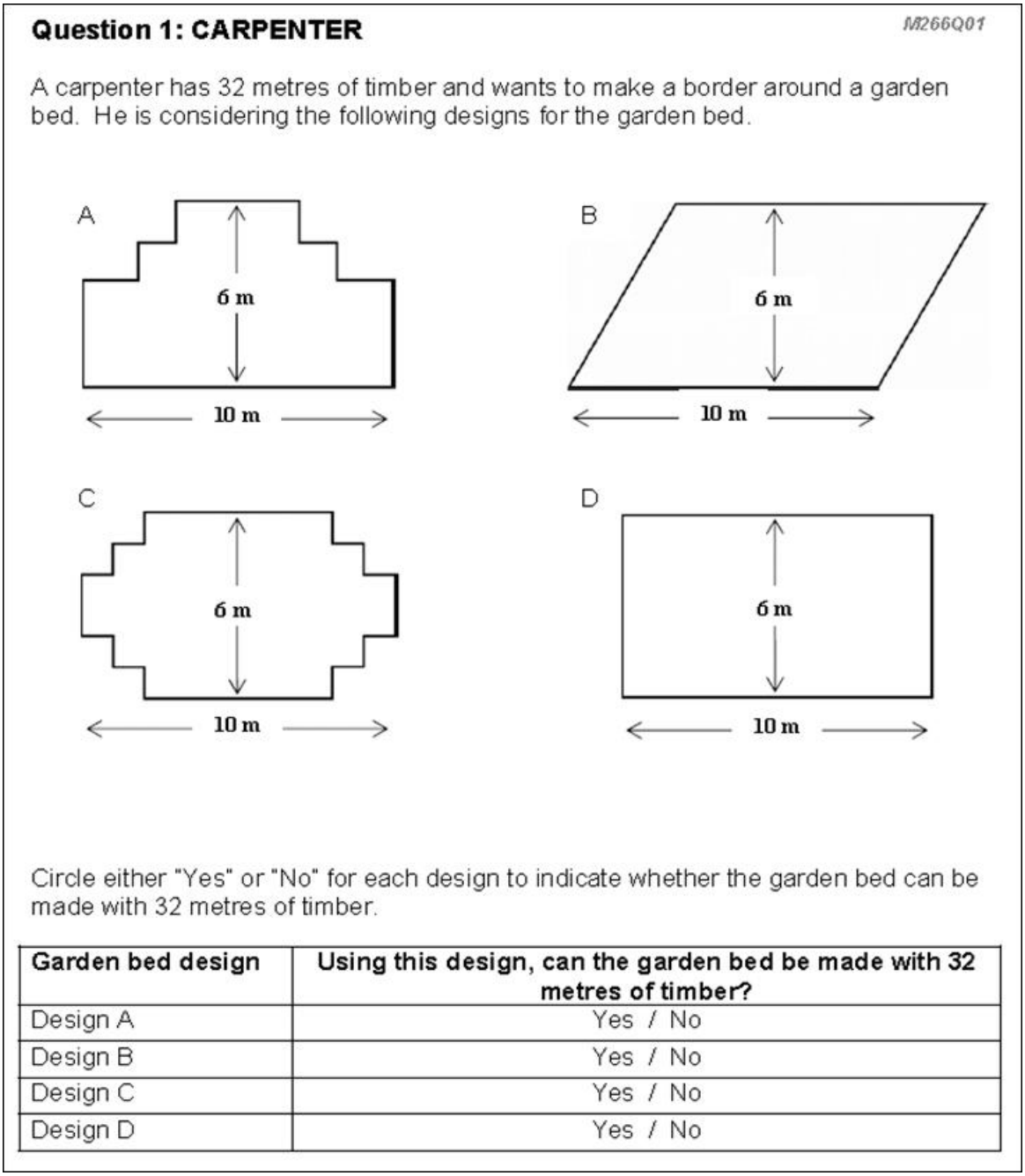
The single arrows represent the impact on validity of the synergy between assessment capacity and the activities performed at the four stages listed above. At Stages 1 and 2 (respectively, Test Development and Test Translation), this synergy mainly concerns cultural validity. At Stages 3 and 4 (respectively, Test Administration and Analysis of Test Results and Use of Assessment Data), this synergy mainly concerns consequential validity. At Stage 3, this synergy also concerns cultural validity.

Cultural Validity

According to our framework, successful participation in the Test Development and Test Translation stages contributes to cultural validity in the development and interpretation of the tests administered in a country. For the purposes of this paper, culture is defined as the shared history and the set of shared experiences, practices, values, and views of a social group mediated by implicit and explicit forms of communication and socialization. This broad definition encompasses multiple aspects of the life of a society, such as economic and environmental circumstances, language and dialect, traditions, legislation, policy, learning styles, teaching practices, epistemologies, and ways of representing information (among many others) (see Solano-Flores & Nelson-Barber, 2001).

Contemporary theories in cultural anthropology and language acquisition maintain that culture and language significantly influence how people make meaning of their experiences and how they construct knowledge (e.g., Bialystok, 2002; Vygotsky, 1978; Wertsch, 1985). As cultural artifacts, tests are not an exception—culture permeates all aspects of assessment (Basterra, 2011; Bonnet, 2002; Dogan & Circi, 2010; Hamano, 2011; Solano-Flores, Contreras-Niño, & Backhoff, 2006; Wuttke, 2007). The content that items are intended to assess is based on explicit assumptions about what students should know at a given grade—assumptions formalized, for example, in PISA's assessment framework. In contrast, the format of those items and the contextual information they provide may be largely based on implicit assumptions about the test takers' cultural experience.

The item shown in Figure 2 illustrates this. While the format in which students are asked to provide their answers (i.e., by circling a Yes or a No for each of a series of categories) may be formalized in an Item Specifications document, the degree of the students' familiarity with this format may vary across countries, depending on the characteristics of the curriculum, the formal or informal teaching practices, and the students' societies' practices concerning the ways in which information is represented. This is not to imply that students would be completely unable to understand and respond to items provided in this format if they are unfamiliar with it. Yet, there is evidence that students from different cultural groups make sense of items based on different sets of types of cultural experiences inside and outside school (Solano-Flores & Li, 2009). Cognitively, the lack of familiarity with the format of a task may increase cognitive load—the effort made in working memory (Sweller, 1994). Potentially, students who are unfamiliar with the format may need to use more working memory than students who are familiar with it to make sense of the task they have to complete and to figure out how to provide their responses. This difference in the time needed to process information could be a source of cultural bias.



Question 1: CARPENTER M266Q01

A carpenter has 32 metres of timber and wants to make a border around a garden bed. He is considering the following designs for the garden bed.

Circle either "Yes" or "No" for each design to indicate whether the garden bed can be made with 32 metres of timber.

Garden bed design	Using this design, can the garden bed be made with 32 metres of timber?
Design A	Yes / No
Design B	Yes / No
Design C	Yes / No
Design D	Yes / No

Figure 2. The Carpenter item. Source: OECD (2006)

Of course, international assessment organizing agencies' procedures take cultural differences into consideration. For example, test item creation is a collaborative process in which country representatives interact with expert advising committees, have the opportunity to take actions intended to address the cultural validity of the test, and are allowed some leeway in adapting specific items to better meet their national cultural contexts (e.g., OECD, 2010). Indeed, countries are advised to be "particularly aware of issues related to nationality, culture, ethnicity, and geographic location" (Mullis & Martin, 2011, p. 8). However, in spite of these efforts, there is still much to do and learn regarding cultural validity in international test comparisons. For example, teams of experts charged with reviewing items (e.g., see Mullis, Martin, Ruddock, O'Sullivan & Preuschoff , 2009) do not typically include experts in disciplines related to language and culture. Yet there is

evidence from experience in the field of testing linguistic minorities that, while necessary, ensuring the participation of individuals from different cultures in assessment development and assessment review teams is not sufficient to properly address culture (Solano-Flores & Gustafson, 2013). In the absence of specialized, formal training, these professionals may not be totally aware of the subtle ways in which culture may influence students' interpretations of items or the extent to which lack of familiarity with the contextual information provided by test items may hamper the ways in which students make sense of them.

According to our framework, successful activities in the Test Administration and Analysis of Test Results stage contribute also to cultural validity. One of these activities is the use of item response theory in the detection of items that are culturally biased. An item that is biased (or differentially functioning) if two different groups, the reference group and the focal group (e.g., respectively, the group tested with the original version of a test and the group tested with a translation of the test) have different probabilities of responding correctly to that item after controlling for differences between the groups on the overall performance on a test (Camilli, 2006; Camilli & Shepard, 1994). While the analysis of differential item functioning is frequently invoked in discussing bias in multicultural assessment contexts (e.g., van de Vijver, 2016), its effectiveness is limited by the extent to which this technique is used with substantial numbers of items and at stages in which items detected as biased can be discarded or modified timely. Needless to say, limited assessment capacity may prevent countries from using this technique properly or sufficiently.

Consequential Validity

Limited assessment capacity may affect the extent to which a country benefits from information obtained from assessment activities. While many countries use international assessment results to initiate education reform (Gilmore, 2005; Stachelek, 2010), most countries primarily use their rankings as evidence of their education failure or success. Several scholars have warned that rankings should be treated cautiously and used only with a full understanding of their construction (Figazzolo, 2009; Hamano, 2008/2011; Sjøberg, 2007; Stachelek, 2010; Tatto, 2006; Wuttke, 2007). However, the strategic and sound use of this information is not likely to occur under limited assessment capacity.

According to our framework, successful participation in the Analysis of Test Results and Use of Assessment Data stages (Figure 1) contributes to addressing consequential validity. Messick (1989) defined test validity in general as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and *appropriateness* of *inferences* and*actions* based on test scores or other modes of assessment" (p. 13, italics in the original). The *actions* part of this definition is what Messick (1989) refers to as consequential validity. Shepard (1997) includes the use and consequences of inferences based on test scores as part of consequential validity.

While some debate exists over including the actions taken based on test scores in the discussion of validity, Messick (1995) elaborated on his assertion of the importance of considering the social use of tests. He argued that "to appraise how well a test does its job, one must inquire whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but also at the same time consistent with other social values" (p. 744).

These important notions, which are critical to fairly and validly assessing culturally diverse populations, are insufficiently addressed in the context of national assessments (see Kane, 2006) and have not been addressed at all in the context of international comparisons. Among the many issues that need to be examined are: *How does assessment capacity influence the fidelity with*

which a country implements international assessment programs' procedures? How can participation in international test comparisons contribute to build a country's assessment capacity? How can education leaders and policy makers be confident that the inferences they are making from their results are valid and the reforms they enact are not having negative, unintended consequences? Unfortunately, while adequate assessment systems are assumed in PISA participating countries, no formal evaluations have been conducted to examine how countries develop their assessment capacity through participation in PISA (Lockheed et al., 2015).

An important aspect of the relationship between consequential validity and assessment capacity is that the answers to questions like those shown above and the ways in which the issues are to be optimally addressed may vary considerably across countries, as there are compelling indications that different societies ascribe different meanings and sets of consequences and possibilities to assessments. The range of perceptions is wide—from assessments as tools for social oppression to assessments as opportunities for social promotion (e.g., see Gebril, 2016; Kennedy, 2016; Lingard & Lewis, 2016).

Conditions for Cultural and Consequential Validity

Based on the reasoning above, it is possible to identify the main conditions that contribute to attaining cultural and consequential validity in international test comparisons. Table 1 provides an initial list, which includes, with adaptations, components from previous efforts oriented to examining issues of cultural validity, consequential validity, and assessment capacity in both national (see Ad-Hoc Technical Committee on the Development of Technical Criteria for Examining Cultural Validity in Educational Assessment, 2015; Martínez-Rizo, 2015) and international assessment contexts (Ercikan & Solano-Flores, 2016; Solano-Flores, 2008). The conditions are grouped into three categories, Assessment Program, Participation, and Practices.

Table 1 - Conditions for Cultural Validity and Consequential Validity

ASSESSMENT PROGRAM
Assessment Framework. The development of the assessment's framework addresses the fact that, because assessments are cultural artifacts, the content of the assessment and the performance of students in the assessment are mediated by socio-cultural experience and language.
Population Specification and Sampling. The procedures used to define and draw samples of the population of students are sensitive to the fact that different countries have different forms of cultural and linguistic diversity and different sets of social and school contexts.
Item Specifications. In addition to being based on the characteristics of the knowledge and skills assessed, decisions concerning the different item formats used in the assessment address the notion that students from different cultures may not be equally familiar with the contextual information provided by items and with their linguistic and graphic features.
Correction Mechanisms. The process of assessment development stipulates the actions to be taken timely with items that are inadequate according to evidence on cognitive and cultural validity obtained from different sources, including expert reviews, differential item functioning analyses, and generalizability studies.
PARTICIPATION
Research and Practice Agenda: The country is able to use information and outcomes from its participation in the assessment program to generate new knowledge on areas of national relevance and to improve practice in those areas.

Sustainability, Stability, and Continuity: The country is able to sustain long-term programs and activities related to its participation in the assessment program efforts regardless of financial and political uncertainty.

Human Resources: The country has or is able to develop or increase in a reasonable time a critical mass of qualified professionals in the field of educational measurement and related areas in connection with its participation in the assessment program.

Financial Resources and Infrastructure: The country has the minimum infrastructure and is able to allocate the financial resources needed to perform activities related to its participation in the assessment after that participation has ended.

Systemic Congruence: The country coordinates its participation in the assessment program with key components in the educational system

Decision Making: The country devotes resources and time to make careful interpretations of the outcomes of the assessment (e.g., beyond simply country rankings) and makes sound decisions concerning its educational system that are within the scope of the information that the assessment is able to produce.

Implementation and Enrichment: In addition to implementing with fidelity the procedures established by the assessment program, the country is able to add activities to those procedures in ways that enable it to meet specific needs.

ASSESSMENT PRACTICES

Conceptual Foundation. The procedures used to address linguistic, cultural, and socio-economic diversity are theoretically defensible.

Timelines. The calendar of the process of assessment development allocates reasonable amounts of time to take all actions concerning cultural diversity.

Development Team. In addition to content experts, educators, and test developers, the teams in charge of developing assessment items include professionals with specialties in the field of culture and language (e.g., anthropology and scociolinguistics) as well as representatives of diverse cultural and linguistic groups.

Representation of Diversity in Population Samples. Samples of students from multiple cultural, linguistic, and socio-economic subgroups are included at all stages of the process of assessment development (e.g., piloting stages).

Cognitive Interviews. Cognitive interviews are conducted to examine whether students from different cultural, linguistic, and socio-economic subgroups interpret in the same ways.

Expert Review Process. A process of review exists in which external teams of experts examine the items and the potential challenges irrelevant to the constructs measured they may pose to students from multiple cultural, linguistic, and socio-economic groups.

Differential Item Functioning Analysis. Representative samples of items are examined to check for differential functioning with special focal groups (e.g., low-income students, students from specific cultural and linguistic groups).

Generalizability Studies. A series of generalizability studies are conducted with different test composites and different student population samples to examine if similar generalizability (reliability) coefficients are obtained for different cultural, linguistic, and socio-economic groups.

Data Disaggregation. The technical properties of tests are examined separately for each main cultural, linguistic, and socio-economic group.

Assessment Program

The conditions under the Assessment Program category refer to activities, procedures, or products that already exist, at least in principle, but which should address aspects of culture and diversity more explicitly. We assume that addressing cultural validity largely depends on the extent to which culture is addressed through the process of development of norming documents and through the planning of the assessment program activities.

Existing international assessment frameworks are not naive about the fact that culture may influence how students interpret items; they typically provide some discussion of culture. However, these documents could make a more substantial contribution to addresing cultural validity if they incorporate culture as a component that cuts across knowledge, rather than treat it as something to think about when most of the assessment framework has been developed.

Other Assessment Program conditions that are seemingly trivial may actually be critical to properly addressing cultural validity. Such is the case of *Timelines* and *Correction Mechanisms*. Typically, actions intended to address issues related to culture (e.g., the review of translation or the procedures for examining cultural sensitivity or potential sources of bias) are evaluated summatively (at the end of the process of assessment development), not formatively (during the process of assessment development). A potential, undesired consequence of this practice may be that any delays in the completion of different stages of the process of assessment development are carried over, eventually impacting the amount of time available to address issues of culture.

Participation

The conditions under the Participation category refer to the characteristics of the actions taken by a country beyond simply implementing the assessment program's procedures, and to the country's ability to insert these actions as part of a broader national plan. We assume that addressing both cultural validity and consequential validity largely depends on the extent to which a country's participation is driven by a clear idea of what needs to be improve and what it takes, financially and in terms of human resources, to make any changes needed. Being able to make sense of data from an international assessment, beyond simply jumping to conclusions based on country rankings requires from a country to make proper allocations of time and human, material, and financial resources. Properly using information from an international test comparison in ways that are not driven by political pressure requires a minimum of institutional stability and system congruence in a country's educational system. Also, it may require deep policy transformations, such as those concerning the structure of allocation of funds in support of high-need schools (Darling-Hammond, 2014). Properly using information from an international test comparison also involves interpreting test results as reflective of deep social inequalities (Carnoy, 2015).

Special attention to *Implementation and Enrichment* allows understanding the extent to which countries can maximize the benefit from participating in international assessment programs. PISA test items are created first in English and in French, and then translated from these two source languages into the participating countries' languages. PISA participating countries are provided with guidelines for test translation (e.g., Hambleton, 2005; National Project Managers' Meeting, 2010) and assessment programs have well established mechanisms for reviewing the accuracy of the translated test items. However, there is evidence that, due to the fact that languages encode experience in different ways, even impeccable translation cannot prevent translation error from occurring. Translation error due to the cultural specificity of the contextual information provided by items, and related to syntactic complexity, semantics, and the alteration of the constructs measured can increase item difficulty (Solano-

Flores, Backhoff, & Contreras-Niño, 2009; 2013). An implication of this finding is that, while translation guidelines provide a valuable standard, general procedure for test translation, countries need to devise approaches for implementing the translation procedures and internally reviewing their own translated items formatively in ways that are sensitive to their cultures and language varieties. This implementation may vary tremendously across countries depending on its resources and their ability to assemble adequate test translation review teams.

Practices

The conditions under the Practices category refer to the assessment activities that a country is able to perform not only when it partakes in an international assessment program but also as part of its own assessment programs. We assume that addressing both cultural validity and consequential validity largely depends on the extent to which countries' practices and experience with assessment are sensitive to issues of culture in their own national contexts.

As with the conditions under the Assessment Program category, most of the conditions listed under this category exist already—in principle. For example, as mentioned above, techniques for detecting differentially functioning items can be used to examine cultural bias (e.g., van de Vijver, 2016). However, the ability of this technique is not a guarantee of use. It is unclear to what extent all PISA participating countries examine differential item functioning routinely with substantial numbers of items or the actions they take with items detected as biased. Moreover, while necessary, the analysis of differential item functioning may not detect biased items when cultural and linguistic heterogeneity in focal groups is not taken into consideration. There is evidence that the more heterogeneous the focal group is linguistically or culturally, the more likely differentially functioning items are to go undetected (Ercikan, Roth, Simon, Sandilands & Lyons-Thomas, 2014). Limited assessment capacity may hamper the ability of some countries to examine differential item functioning for important focal groups along with properly modeling heterogeneity.

Generalizability Studies deserves a special consideration. Generalizability studies are not currently part of established practices related to addressing culture, though research shows it is relevant to examining cultural validity (Solano-Flores & Li, 2006, 2009). Generalizability (G) theory (Brennan, 2001; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson & Webb, 1991) is a theory of measurement error and also a theory of sampling of observations (Kane, 1982). Unlike differential item functioning approaches, G theory-based approaches focus on overall test scores on random-assumed test composites (see Solano-Flores, 2016). Research on the use of G theory with linguistic and cultural minorities indicates that the dependability of the scores on the same test may vary across cultural groups. The implication of this finding brings to another Practices condition—*Disaggregation of Data*. Disaggregating data by cultural group and comparing groups in terms of the dependability of the scores they obtain in a test is a more rigorous approach than comparing groups as to their mean test scores (Solano-Flores & Li, 2013).

Final Remarks

We have examined the relationship between assessment capacity, cultural validity, and consequential validity in PISA. We contend that a critical level of assessment capacity is needed for a country to benefit from participating in international test comparisons. At the same time, participation in international test comparisons is a good opportunity for countries to increase their assessment capacities, as long as they pay proper attention to cultural validity and consequential validity.

Some of the considerations on cultural and consequential validity discussed should impact procedures used by international test comparison organizing agencies (e.g., those concerning the development of assessment framework and item specification documents). Others involve readiness or the responsibility

of each participating country to address the challenges that are specific to the cultural makeup of its student population.

Limited assessment capacity may hamper countries' effectiveness in addressing cultural validity and certainly raises issues concerning consequential validity. This is a fundamental issue of equity at the international level. PISA participating countries need to ensure they have a minimum assessment capacity if their participation in international test comparisons is to accurately inform their education reform efforts. Otherwise, PISA countries' participation needs to be accompanied by solid national assessment capacity building programs. Most importantly, organizing agencies should take responsibility in supporting countries in developing their assessment capacity.

References

Ad-Hoc Technical Committee on the Development of Technical Criteria for Examining Cultural Validity in Educational Assessment. (2015). Promoting and evaluating cultural validity in the activities performed by the National Institute for Educational Evaluation (INEE). Submitted to the National Institute for Educational Evaluation. Mexico City, Mexico, January 16.

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

Bialystok, E. (2002). Cognitive processes of L2 users. In V. J. Cook (Ed.), *Portraits of the L2 user* (pp. 145-165). Buffalo, NY: Multilingual Matters.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice*, *9*(3), 387-399.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Paper Number 71*. Retrieved from OECD website: http://www.oecd-ilibrary.org/education/the-policy-impact-of-pisa_5k9fdfqffr28-en

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer Verlag.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221-256). Westport, CT: American Council on Education and Praeger.

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

Capacity Development Group (2007, May). *Capacity assessment methodology: User's guide*. Bureau for Development Policy, United Nations Development Program. New York, September 2005. Retrieved from the United Nations Development Programme website: https://www.unpei.org/sites/default/files/PDF/institutioncapacity /UNDP-Capacity-Assessment-User-Guide.pdf

Carnoy, M. (2015). International test score comparisons and educational policy. Carnoy, M. (2015). *International Test Score Comparisons and Educational Policy: A Review of the Critiques*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/international-test-scores

Clarke, M. (2012). What matters most for student assessment systems: A framework paper. Retrieved from the World Bank website: https://openknowledge.worldbank.org/bitstream/handle/10986/17471/682350WP00PUBL0WP10READ0web04019012.pdf?sequence=1

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Darling-Hammond, Linda (2014). What can PISA tell us about U.S. education policy? *New England Journal of Public Policy*: *26*(1), Art. 4. Retrieved from http://scholarworks.umb.edu/nejpp/vol26/iss1/4

Dogan, E., & Circi, R. (2010). A blind item-review process as a method to investigate invalid moderators of item difficulty in translated assessment. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* (Vol. 3) (pp. 157-172). Hamburg: IERI.

Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about uses of international assessments. *Teachers College Record*, *117*(1), 1-28.

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

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

Ferrer, G. (2006). *Educational assessment systems in Latin America: Current practice and future challenges*. Washington, DC: PREAL. Retrieved from http://www.uis.unesco.org/Education/Documents/Ferrer.pdf

Figazzolo, L. (2009). *Impact of PISA 2006 on the education policy debate*. Retrieved from http://download.ei-ie.org/docs/IRISDocuments/Research%20Website%20Documents/2009-00036-01-E.pdf

Gebril, A. (2016). Educational assessment in Muslim countries: Values, policies, and practices. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Gilmore, A. (2005). The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries: An evaluation of the value of World Back support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS). Retrieved from http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/Gilmore_Impact_PIRLS_TIMSS.pdf

Hamano, T. (2011).The globalization of student assessments and its impact on education policy [English version]. *Proceedings, 13*, 1-11. (Originally appeared in Japanese in 2008 in the *Annual Bulletin of JASEP (Japan Academic Society for Educational Policy), 15*, 21-37). Retrieved from http://teapot.lib.ocha.ac.jp/ocha/bitstream/10083/51418/1/Proceedings13_01Hamano.pdf

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

Husén, T. (1983). *An incurable academic: Memoirs of a professor*. Oxford, UK: Pergamon Press.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, *54*(1), 5-25. doi: http://dx.doi.org/10.1086/648471

Kane, M. T. (1982). A sampling model of validity. *Applied Psychological Measurement, 6*, 125-160. doi: http://dx.doi.org/10.1177/014662168200600201

Kane, M. T. (2006). *Validation*. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kennedy, K. J. (2016). Exploring the influence of culture on assessment: The case of teachers' conceptions of assessment in Confucian-heritage societies. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Lingard, B., & Lewis, S. (2016). Globalization of the American approach to accountability: The high price of testing. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Martínez-Rizo, F. (2015). Las pruebas ENLACE y EXCALE: Un estudio de validación [The ENLACE and EXCALE assessments: A validation study]. Retrieved from http://publicaciones.inee.edu.mx/buscadorPub/P1/C/148/P1C148.pdf

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

Messick, S. (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: http://dx.doi.org/10.1037/0003-066X.50.9.741

Mullis, I. V. S., & Martin, M. O. (2011). *TIMSS 2011 item writing guidelines.* Retrieved from http://timssandpirls.bc.edu/methods/pdf/T11_Item_writing_guidelines.pdf

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011: Assessment frameworks.* Retrieved from http://timssandpirls.bc.edu/timss2011/downloads/TIMSS2011_Frameworks.pdf

National Project Managers' Meeting (2010, October). Translation and adaptation guidelines for PISA 2012. Doc: NPM10104e. PISA Consortium. Budapest, Hungary. Retrieved from https://www.oecd.org/pisa/pisaproducts/49273486.pdf

Organisation for Economic Co-operation and Development (OECD). (n.d.). *Programme for international student assessment (PISA): Results from PISA 2012, Country note: United States.*Retrieved from http://www.oecd.org/pisa/keyfindings/PISA-2012-results-US.pdf

Organisation for Economic Co-operation and Development (2006). *PISA released items: Mathematics*. Retrieved http://www.oecd.org/pisa/38709418.pdf

Organisation for Economic Co-operation and Development (2010). *Translation and adaptation guidelines for PISA 2012.* Retrieved on from http://www.oecd.org/pisa/pisaproducts/49273486.pdf

Lockheed, M., Prokic-Bruer, T., & Shadrova, A. (2015). *The experience of middle-income countries participating in PISA 2000-2015* (PISA series). Washington, D.C. & Paris: The World Bank & OECD Publishing. doi: http://dx.doi.org/10.1787 /9789264246195-en

Ravela, P. (Ed.). (2001). Los próximos pasos: ¿Hacia dónde y cómo avanzar en la evaluación de aprendizajes en América Latina? [The next steps: Where and how to advance the evaluation of learning in Latin America?] Document No. 20. Working Group on Assessment and Standards. Santiago: PREAL. Retrieved from http://campus-oei.org/calidad/grade.PDF

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13. doi: http://dx.doi.org/10.1111/j.1745-3992.1997.tb00585.x

Sjøberg, S. (2007). PISA and "real life challenges": Mission impossible. In S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *According to PISA—Does PISA keep what it promises?* Berlin: LIT Verlag.

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

Conference of the International Test Commission, Liverpool, UK.

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

Solano-Flores, G. (2016). Generalizability. In L. E. Suter, D. Wyse, E. Smith, & N. Selwyn (Eds.), *The BERA/SSAGE Handbook of Educational Research* (chap. 47). London: Sage.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2006). Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y otras comparaciones internacionales [Test translations and adaptation: Lessons learned and recommendations for countries participating in TIMSS, PISA, and other international comparisons*].Revista Electrónica de Investigación Educativa (REDIE) [Electronic Journal of Educational Research], 8(*2). Retrieved from http://redie.uabc.mx /redie/article/download/143/246

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L.A. (2009). Theory of test translation error. *International Journal of Testing, 9*, 78-91.

Solano-Flores, G., Contreras-Niño, L.A., & Backhoff, E. (2013). The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research in the context of the programme for international student assessment* (pp. 71-85). Springer Verlag.

Solano-Flores, G., & Gustafson, M. (2013). Assessment of English language learners: A critical, probabilistic, systemic view. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large scaleassessment in education: Theory, issues, and practice* (pp. 87-109). New York: Routledge.

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13-22.

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

Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation, 19*(2-3), 245-263. doi: http://dx.doi.org/10.1080/13803611.2013.767632

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 533-573. doi: http://dx.doi.org/10.1002/tea.1018

Stachelek, A. J. (2010). Exploring motivational factors for educational reform: Do international comparisons dictate educational policy? *Journal of Mathematics Education at Teachers College, 1*,52-55.

Suter, Larry E. (2000). Is student achievement immutable? Evidence from international studies on schooling and student achievement. *Review of Educational Research, 70*(4), 529-545. doi: http://dx.doi.org/10.3102/00346543070004529

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295-312. doi: http://dx.doi.org/10.1016/0959-4752(94)90003-5

Tatto, M. T. (2006). Education reform and the global regulation of teachers' education, development and work: A cross-cultural analysis. *International Journal of Educational Research, 45*, 231-241. doi: http://dx.doi.org/10.1016/j.ijer.2007.02.003

Teltemann, J., & Klieme, E. (2016). The impact of international testing projects on policy and

practice. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment* (Chap. 21). New York: Routledge.

van de Vijver, F. J. R. (2016). Assessment in education in multicultural populations. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*, (Chap. 25).New York: Routledge.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.

Wuttke, J. (2007). Uncertainties and bias in PISA. In S. T. Hopmann, G. Brinek, and M. Retzl (Eds.), *According to PISA – Does PISA keep what it promises?* Berlin: LIT Verlag.

Author / Autor

To know more / Saber más

Solano-Flores, Guillermo (gsolanof@stanford.edu).

Is the corresponding author for this article. He is professor of education at the Graduate School of Education, Stanford University, US. He specializes in educational assessment and the linguistic and cultural issues that are relevant to both international test comparisons and the testing of cultural and linguistic minorities. His contributions to the field of educational assessment include the theory of test translation error, the use of generalizability theory—a psychometric theory of measurement error—in the testing of linguistic minorities, the formalization of the concept of cultural validity, and the design of a methodology for designing and analyzing illustrations used in test items.

Milbourn, Tamara (tamara.milbourn@colorado.edu)

Is a Ph.D. Candidate in Educational Foundations, Policy and Practice at the University of Colorado Boulder with a Master's Degree in Applied English Linguistics from the University of Wisconsin-Madison, US. Her work examines the experiences of international students on American campuses. She is interested in issues in education related to mono/multi-lingualism, with an emphasis on issues of equity as connected to language practices and academic norms. She has worked in Taiwan, Japan and Benin and is currently teaching educational policy and linguistics courses in the University of Colorado System

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

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]