Understanding validity issues in test-based models of school and teacher evaluation

Beatrice L. Bridglall
Secondary and Special Education, Montclair State University, Montclair, New Jersey, USA

Jade Caines
Department of Education, University of New Hampshire, Durham, New Hampshire, USA, and

Madhabi Chatterji
Organization and Leadership, Teachers College, Columbia University, New York, New York, USA

Abstract

Purpose – This policy brief, the second AERI-NEPC eBrief in the series "Understanding validity issues around the world", focuses on validity as it applies to test-based models of evaluation employed for schools, instructional programs, and teachers around the world. It discusses validity issues that could arise when data from student achievement test administrations and other sources are used for conducting personnel appraisals, program evaluations, or for external accountability purposes, suggesting solutions and recommendations for improving validity in such applications of test-based information.

Design/methodology/approach – This policy brief is based on a synthesis of conference proceedings and review of selected pieces of extant literature. It begins by summarizing perspectives of an invited expert panel on the topic. To that synthesis, the authors add their own analysis of key issues. They conclude by offering recommendations for test developers and test users.

Findings – The authors conclude that systematic improvement and transformation of schools depends on thoughtfully conceptualizing, implementing, and using data from testing and broad-based evaluation systems that incorporate multiple kinds of evidence. Evaluation systems that are valid and fair to students, teachers and education leaders need all three of the following: assessment resources and training for all participants and evaluation users; knowledgeable staff to continuously monitor processes and use assessment results appropriately to improve teaching and learning activities; and a strengths-based approach to make improvements to the education system based on relevant data and reports (as opposed to a deficits-based one in which blame or punishment is leveled at individuals or groups of workers when gaps in performance are observed).

Originality/value – To improve validity in interpretations of results from test-based teacher and school evaluation models, the authors provide recommendations for measurement and evaluation specialists as well as for educators, policy makers, and public users of data. Standardized test use in formative and more "high stakes" educational accountability contexts is rapidly spreading to various regions of the world. This eBrief shows that understandings of validity are still uneven among key stakeholders. By translating complex information pertinent to current validity issues, this policy brief attempts to address this need, and also bridge knowledge and communications gaps among different constituencies.

Keywords Educational accountability, Formative evaluation, School evaluations, Teacher evaluations, Test-based school accountability

Paper type Research paper

Quality Assurance in Education
Vol. 22 No. 1, 2014
pp. 19-30
Emerald Group Publishing Limited
0968-4883
DOI 10.1108/QAE-12-2013-0053

Introduction

Today, students' standardized test scores serve as important indicators of performance in evaluation systems designed for judging school or teacher effectiveness in different ways[1]. Some actions tied to such uses of test-based information are formative. That is, they are intended for improving instruction, student learning, or other functions of educational institutions. Other actions are more summative and associated with "high stakes" accountability-related actions and policies. The latter involve final pronouncements of teacher or school quality, often with public sanctions or rewards tied to the results obtained from large-scale student testing efforts.

Student testing programs are usually designed to fulfill particular purposes and to measure particular domains in particular populations – for example, math problem-solving ability in fifth graders for the purposes of program improvement (Chatterji, 2013). Sometimes, however, the results of a given testing program are called upon to serve multiple purposes, some of which may conflict. When this happens, to what degree are the interpretations and actions based on test scores and other data valid, and what can we do to improve validity and use of evaluation results?

This policy brief – eBrief in short – focuses on validity as it applies to test-based models of evaluation employed for schools, instructional programs, and teachers around the world. It discusses validity issues that could arise when data from student achievement test administrations and other sources are used for conducting personnel appraisals, program evaluations, or for external accountability purposes, suggesting solutions and recommendations for improving validity in such applications of test-based information.

In this eBrief, we reference school and teacher evaluation systems in The Netherlands, the US, Denmark, Malaysia, and other nations to try to address these issues. The eBrief also explores several open questions on validity and related issues in measuring and evaluating school-based constructs.

Who and what this eBrief speaks to

The validity issues discussed here have application to a wide variety of audiences. Measurement and evaluation specialists or researchers with similar interests will find insights about the following:

- How should we go about designing and operating sound test-based evaluation systems for schools and teachers for formative or summative decision-making?
- What could we do to improve validity when using tests and evaluations in such decision-making contexts?

Decision-makers and educational leaders will find insights about issues like these:

- How we should go about preparing teachers and other school-based staff to use evaluation information in more valid and appropriate ways, consistent with the purposes of the assessment and evaluation systems (whether formative or summative).

Educators, media and public stakeholders at large will find information here about:

- The appropriate uses of test-based information (with or without other kinds of data) in performance evaluation systems typically found in schools and school systems.

A summary of main themes

Adrie Visscher's main ideas

The Dutch Inspectorate of Education in The Netherlands assesses the quality of primary schools, teachers and student learning with the chief goal of continuous improvement. In describing this evaluation system, Visscher (2013) stresses the salience of its formative mission. Noting the national education policy context in The Netherlands, he points to the prerequisites for continually improve teaching, learning and assessment practices in primary classrooms and schools.

The Dutch Inspectorate of Education (DIE) employs a proportional supervision of schools for evaluation, identifying "strong" versus "weak" primary schools on public information sites. In 2007, the Dutch government also launched its Achievement Oriented Work (AOW) policy with a goal that by 2018, 90 percent of all Dutch primary schools will work in a "results-oriented way" – meaning that teachers would continuously shape their day-to-day teaching and learning processes based on formative use of results of centrally developed student achievement tests in core subject areas (Visscher, 2013).

Outside the expectation that teachers use test scores and employ AOW processes, other criteria employed in evaluating schools are:

- annual student achievement gains based on test scores;
- average student achievement at the end of their primary education; and
- classroom teaching and learning processes and levels of student care.

Currently, however, Visscher (2013) notes that only 24.5 percent of 7,000 primary schools meet the AOW standards. This is notwithstanding various initiatives, such as the establishment of a student-monitoring system at the national level for facilitating AOW implementation in schools.

Schools that are doing well are visited by the DIE at least once every four years. However, schools at risk are investigated in depth and, if required, monitored intensively through visits to the school; meetings with the school board; interviews with principals, teachers and sometimes students; classroom observations; and the analysis of school documents. These data sources are guided by the DIE's supervision framework and standards. If student-achievement data are not available, the school is assessed on whether it knows the educational needs of its students, particularly whether it annually evaluates student achievement, the instructional processes, school improvement activities, and the quality of the educational processes. The goal is to improve school performance as well as evaluations of the school's efforts. This strategy has resulted in a decline in the number of schools labeled as "very weak" (Visscher, 2013).

Visscher (2013) indicates that the AOW policy, which is anchored in the research literature, espouses the importance of feedback, setting goals, and improving instruction using data from student assessments. This rational, goal-oriented approach to schooling relies heavily on evaluation as a central mechanism for promoting the use of systematically gathered data. It approaches school effectiveness and teacher performance from the point-of-view of student learning, aided by formative use of data.

All is not, however, going smoothly in The Netherlands, according to Visscher. Teachers' activities do not adequately mirror AOW's goals. Evaluation and assessment results are often used in minor ways that do not reflect clear and detailed analyses of educational needs of students in a classroom. Teachers are often not aware of where students are in their learning trajectory and cannot define clear and explicit instructional goals or choose suitable approaches to help increase levels of student learning. In noting difficulties in gauging how well the AOW processes are implemented today, Visscher points to various likely influences and factors. These, in particular, include resource availability and a recent limited-scale performance-pay policy tied to students' test scores, that is also being tried out in The Netherlands (Visscher, 2013).

Visscher's recommendations

Given these dynamics, Visscher (2013) identifies certain pre-conditions for AOW to be implemented more effectively in Dutch primary schools and at the regional education system levels. Seven key needs are:

(1) Motivation to transform the school culture into one that is performance oriented.

(2) Provision of timely analyses by school leaders to teachers, discussion of results at the organizational and faculty levels, and monitoring of decisions regarding the implementation of organizational and pedagogical improvement efforts.

(3) Development of an AOW culture that values evaluation use, including use of test-based data, that provides time for school teams to analyze and discuss results, set goals and make decisions regarding instructional improvements for student success.

(4) Cooperation and exchange of AOW task-relevant information between and within schools.

(5) The use of a high-quality student testing and monitoring system that yields reliable and valid information.

(6) A good match between the tests administered to students and learning materials used in classrooms – a factor that will contribute to the validity of information on school and classroom performance.

(7) A clear division of labor concerning data collection, analyses, interpretation of results and subsequent follow up AOW processes.

Reactions to Visscher: main ideas

In their respective discussions of Visscher's (2013) perspective, Aaron Pallas, Drew Gitomer, Jakob Wandall and Haniza Yon recognize the formative goals of The Netherlands' AOW policy as aiming to help educators use evaluation data in unique ways to improve student achievement. Pallas (2013), for instance, applauds the Dutch government's goal to use of data to monitor and improve school quality, not simply to identify or shut down underperforming schools – as is the current policy in the New York City school system, which he has examined closely in recent years. The AOW and DIE system is not punitive to teachers at present, in his view.

He observes, however, that in any evaluation system we should recognize that data are infused with values – even when generated from seemingly "objective" sources like standardized student achievement tests. He contrasts the underlying values and potential repercussions of the DIE evaluation system with that the more summative "value-added" New York City teacher evaluation model. He discusses a teacher-evaluation case from New York, showing how errors can easily arise with limited evaluation designs that overlook limitations of test-based data.

Also noting his involvement with a university-based assessment committee that collects data simultaneously for summatively-orientated accreditation decisions coupled with more formative and continuous program improvement processes, Pallas acknowledges the practical challenges in keeping the formative and summative processes relevant and separate. He believes that the summative documentation of program results for accreditation can easily overshadow the use of those results formatively to improve a program's activities (Pallas, 2013).

Gitomer (2013) agrees that it is important for school and teacher accountability systems to represent a set of shared societal values. He believes that current teacher evaluation systems in the US that utilize students' standardized test scores and are tied to high-stakes actions are unsatisfactory on several counts. They do not accurately reflect teachers' real contributions to student learning and achievement. Better methodologies for measuring teacher performance are still being devised, yet policy-makers seem to believe that existing data systems are "good enough" for implementing more summative evaluation systems right away. In comparison, he applauds the Dutch evaluation system, which strives for continuous improvement of education, regardless of the nation's high performance on the Programme of International Student Assessment (PISA) and other international large scale assessment programs.

Gitomer (2013) appreciates the AOW policy's focus on teaching and learning. He also, however, expresses concerns that while AOW has a lofty theory of action, it does not adequately consider contextual factors that can impede implementation. For instance, teacher and staff training in AOW processes is prerequisite to its success. As such, Gitomer asserts, schools and educators may be ill-prepared to implement the AOW. It would also be beneficial, he suggests, to mesh the summative goals of the DIE's school evaluation framework with the more formative AOW policy for classrooms.

Wandall (2013) notes some key historical, cultural, and value-based differences among the primary education systems in Denmark (his native country), The Netherlands (Visscher's homeland) and the US. At the primary education level, Denmark has a decentralized education system that emphasizes formative and relatively informal evaluations of student learning by teachers. This, Wandall believes, is because in Denmark the overall goals of basic education are to promote non-cognitive skills like, motivation for education and personal development, initiative, democratic and participatory skills. Academic skills and knowledge are important, but primarily as means to achieve these overall non-cognitive goals. This ingrained value offers a contrast to more achievement-oriented competition valued in the US culture and public education system, which emphasizes test performance and accountability. In Wandall's view, there are strong similarities between Dutch and Danish values, but The Netherlands' system falls somewhere in between the US and Denmark in terms of educational goals and values (Wandall, 2013).

In 2010, prodded by results of international assessments like the PISA, Denmark finally developed an internet based adaptive National Testing System that enable the Ministry of Education to monitor the municipalities, the municipalities to monitor schools and the school to monitor student performance. However, results are supposed to be used in significantly different ways from countries such as The Netherlands and the US. Not only are the tests designed for formative purposes, but only the schools have access to students test data and only the relevant teacher must know the test results in details. It is unlawful in Denmark to publicize test results of municipalities, schools, classes and students (Wandall, 2013).

The challenge with Denmark's decentralized approach is that many schools and teachers are not aware of how to use the student assessment data to improve instructional practices and student learning. Commenting on The Netherlands' evaluation system, Wandall (2013) believes that the real problem may not be monitoring school quality, but developing a culture that supports the appropriate use of test results for improving instruction. Schools cannot be changed from the outside, he asserts, but must be transformed from the inside (Wandall, 2013).

Yon (2013) similarly points to the need for a complete transformation of schools in order for the AOW-related assessment reforms to be successful. Specifically, she outlines other considerations – and possibly unforeseen consequences – for the Dutch as they attempt to implement the AOW policy and school evaluation system. She contends that policies related to performance and pay levels of teachers should be re-examined given the mixed research findings on the effects of teacher monetary rewards. Since teachers are administering the exams to students themselves as a part of the AOW system, issues related to cheating and "teaching to the test" should also be addressed. There are also some key characteristics of schools and teachers that can influence the success of the AOW policy, including additional resources for teachers, the creation of knowledge-sharing communities, and attitude shifts of both school principals and teachers.

Yon (2013) also compares the Dutch system to Malaysia's national testing system. She discusses the benefits of AOW-like assessment and educational reforms there, particularly in the area of teacher empowerment. She believes that continuously monitoring these new test-based evaluation systems in The Netherlands, Malaysia and beyond will be critical for establishing and upholding validity standards.

Stakeholder views

Audience concerns

Audience members expressed some concerns worth noting here. For example, a representative from a US teachers union had concerns that the Dutch were investigating performance pay for teachers, especially in light of evidence since the 1950s in the US, demonstrating that teacher performance pay has not been clearly beneficial. Visscher explained that policy makers were pursuing the notion of fairness in rewarding teachers for helping students move beyond their starting achievement levels. But he acknowledged that the authorities were likely not aware of the technical and conceptual difficulties of measuring "value added" by teachers to student learning, nor were they aware of recent research studies suggesting that rewarding teachers with bonus pay does not increase student achievement.

Interestingly, Dutch teachers themselves were, however, aware of key hurdles faced by such a policy. Visscher recounted that in a poll of teachers in The Netherlands, 85 percent believed that it was not possible to accurately measure the impact of their performance on student achievement. As a result, many did not believe in the performance-based pay policy. Visscher noted that he and many others in The Netherlands hope that efforts around teacher performance-based pay policies would not be implemented on a large scale.

Continuing this discussion, a US parent with four children in the local New York City school system wondered how to preserve and protect what cannot be quantified in education. Referring to the high-stakes, "value-added" teacher evaluation system that gives great weight to students' performance on standardized tests in New York City, she described the widespread fear generated by this accountability policy. She said that teachers are stressed, and there is a marked flight of good teachers away from City schools. Students and parents are also stressed. "How do we preserve learning communities that may be squashed by micromanagement (resulting from too much testing and accountability)? How do we prevent the fracturing of the community and prevent the sense of stress on the part of educators on how students and teachers are evaluated?" she inquired.

The panelists agreed that such high-stakes accountability systems are influencing education in a way that is not productive. To mitigate the damage, there is a need to demystify the technical aspects of evaluation processes and prepare teachers to understand assessment. This form of training must be embedded in the course of their careers or in their pre-service training, panelists asserted. We can and should do a better job of helping teachers, educators and educational policy-makers understand the technical aspects of assessments and accountability policies that are used for taking high-stakes actions that are consequential for schools, teachers, leaders and students and their families.

A higher-education representative from China observed that students from Shanghai excelled on PISA tests in 2010 (referred to as the "Shanghai Phenomenon" by the media), but this reality is not reflective of the status of education and students from other Chinese provinces. He attributed the Shanghai Phenomenon to students having greater opportunities to learn in Shanghai. Given this situation, he asked, how can we make assessments and evaluations based on international testing efforts more fair and valid for students from other regions who do not have adequate opportunities to learn? Do the comparative PISA results provide a fair evaluation of education systems? Visscher noted that the Shanghai Phenomenon reminds him of the situation in Belgium, where PISA received much public attention. For improving valid interpretations, Visscher recommended separation of student performance data by region and evaluating results in the individual regional contexts, rather than by comparison with each other.

Conclusions

Our thoughts

Visscher's (2013) thoughts and the resulting discussion of issues by panelists and various stakeholders illustrate that systematic improvement and transformation of schools depends on thoughtfully conceptualizing, implementing, and using data from evaluation systems. This is a moral imperative in our connected world, where students need to have the background and preparation to be successful in a global economy.

Responsive teaching and assessment practices that accommodate different student backgrounds and learning styles and support development are necessary. Prioritizing more responsive teaching and learning approaches can help to close opportunity gaps and therefore achievement gaps in education.

In this vein, it is crucial that educators are informed about and trained in collaborative assessment and evaluation efforts that use multiple kinds of evidence, including student test scores. This inclusive stance recognizes the unique role educators play in whether students learn and learn how to learn. Towards this end, educators need the support of their school leaders and policy makers, who themselves must gain expertise as instructional leaders with a basic knowledge of assessment and evaluation.

Evaluation systems that are valid and fair to students, teachers and education leaders need all three of the following:

(1) assessment resources and training for all participants and evaluation users;

(2) knowledgeable staff to continuously monitor processes and use assessment results appropriately to improve teaching and learning activities; and

(3) a strengths-based approach to make improvements to the education system based on relevant data and reports (as opposed to a deficits-based one in which blame or punishment is leveled at individuals or groups of workers when gaps in performance are observed).

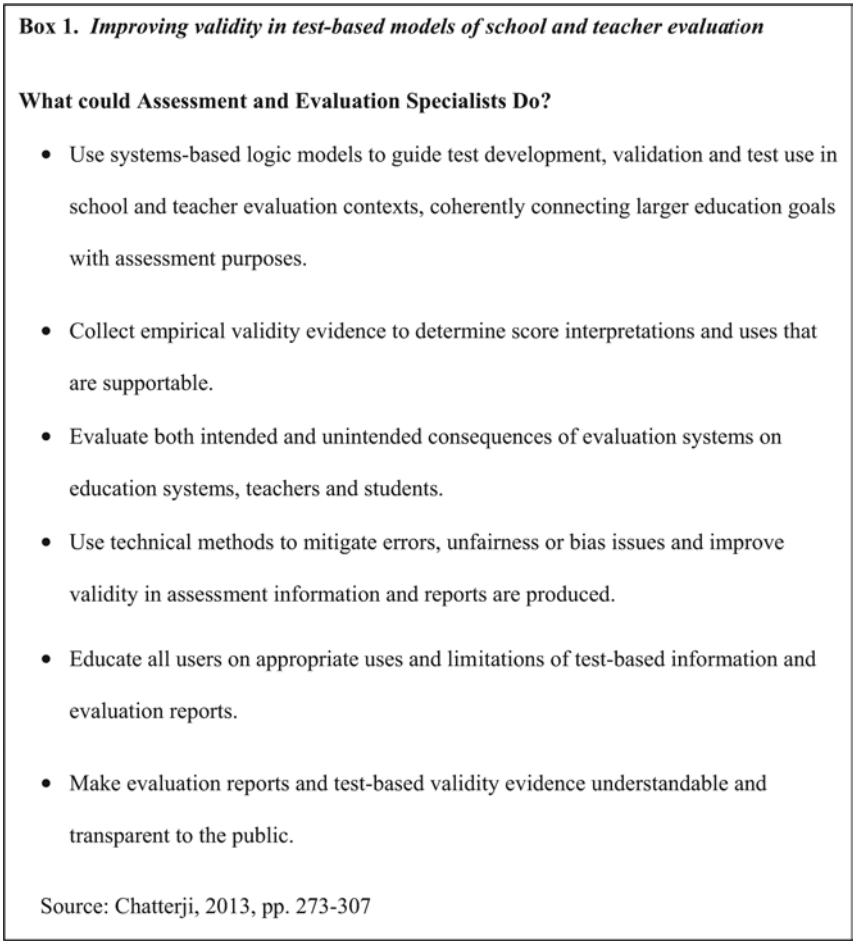
A strengths-based focus of this type enables a cultural shift in organizations that reduces blame and encourages both formative and summative decision-making with assessment and evaluation results that are compliant with professional standards of practice (AERA, APA and NCME, 1999; Yarborough *et al.*, 2010).

Recommendations

As we work towards these goals, we are reminded that Messick's (1989) idea of validity as "an integrated evaluative judgment" that informs "the adequacy and appropriateness of inferences and actions based on test scores," has important implications for designing and conducting evaluations of teachers and schools in education (p. 13). To apply this principle, Chatterji (2013) summarizes collective insights from all the contributors to the conference, making observations and recommendations of her own that audiences of this eBrief may find useful (see Figures 1-2).

To improve validity of interpretations of results from test-based teacher and school evaluation models, for example, Chatterji (2013) recommends that measurement and evaluation specialists use "systems-based logic models" to guide validation procedures and improve communications with stakeholders to forestall unintended consequences that could clash with well-intentioned goals of education systems.

When high stakes punishments or rewards are tied to test-based evaluation results for schools and teachers, a commonly found consequence is that teachers narrow their instruction by "teaching to the test" rather than focusing on the overall curriculum. This practice shortchanges the richness of what students learn in school. Yet test scores typically go up, becoming a misleading and invalid indicator of school or teacher effectiveness! Logic models – a flow-charting method – could help map and improve understandings of how particular policy actions with test scores or assessment reports could lead to classroom-level behaviors that could prove to be either beneficial or harmful to the originally-set goals of schooling (see Figure 1). Insights gained from logic



Box 1. ***Improving validity in test-based models of school and teacher evaluation***

What could Assessment and Evaluation Specialists Do?

- Use systems-based logic models to guide test development, validation and test use in school and teacher evaluation contexts, coherently connecting larger education goals with assessment purposes.
- Collect empirical validity evidence to determine score interpretations and uses that are supportable.
- Evaluate both intended and unintended consequences of evaluation systems on education systems, teachers and students.
- Use technical methods to mitigate errors, unfairness or bias issues and improve validity in assessment information and reports are produced.
- Educate all users on appropriate uses and limitations of test-based information and evaluation reports.
- Make evaluation reports and test-based validity evidence understandable and transparent to the public.

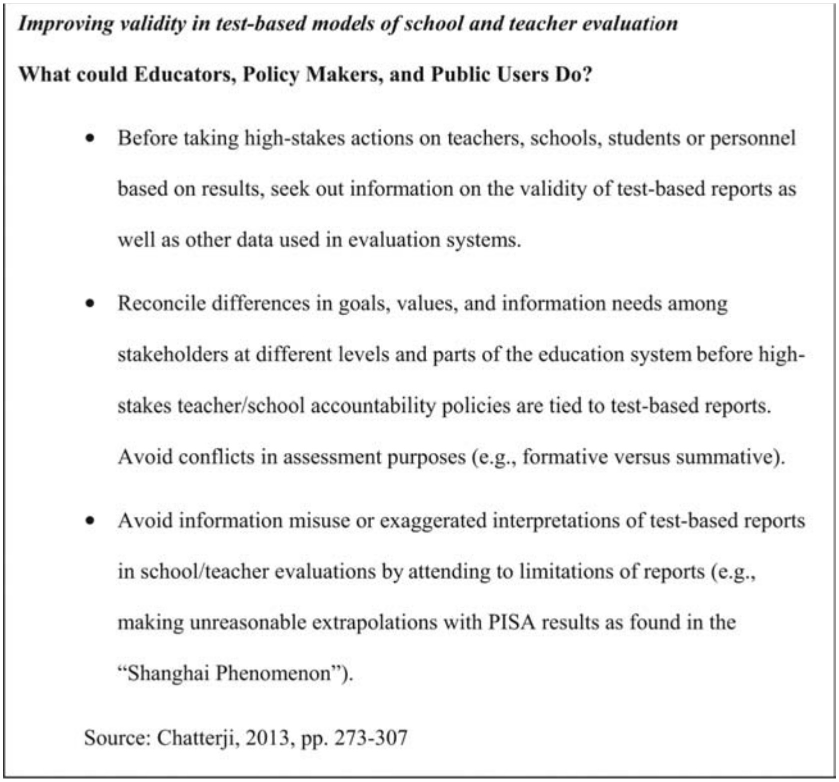
Source: Chatterji, 2013, pp. 273-307

Figure 1.
How can test developers improve validity?

modeling could broaden the validation agenda, allowing the collection of appropriate kinds of evidence. Technical methods could also be employed to mitigate potential errors in interpretation and test score misuse that could arise down the road.

Secondly, when there are multiple purposes for the same assessment data and reports (formative versus summative; see Figure 2), a consequence is that the test-based results may be sufficiently valid for one use and interpretation, such as identifying learning needs of students and improving instruction, but have inadequate levels of validity for other uses, such as ranking schools and teachers on their performance levels and quality. Validity evidence should be acquired at a sufficient level for all intended actions and interpretations, before test-based evaluation systems are made operational in accountability policy contexts.

To public test users, including the media, Chatterji (2013) cautions against making exaggerated claims and misusing test-based data beyond the information reports



Improving validity in test-based models of school and teacher evaluation

What could Educators, Policy Makers, and Public Users Do?

- Before taking high-stakes actions on teachers, schools, students or personnel based on results, seek out information on the validity of test-based reports as well as other data used in evaluation systems.
- Reconcile differences in goals, values, and information needs among stakeholders at different levels and parts of the education system before high-stakes teacher/school accountability policies are tied to test-based reports. Avoid conflicts in assessment purposes (e.g., formative versus summative).
- Avoid information misuse or exaggerated interpretations of test-based reports in school/teacher evaluations by attending to limitations of reports (e.g., making unreasonable extrapolations with PISA results as found in the "Shanghai Phenomenon").

Source: Chatterji, 2013, pp. 273-307

Figure 2.
What could policy makers, action takers, and public users do?

legitimately provide (Figure 2). To do so, she references the example of the 2010 Shanghai Phenomenon where the US media extrapolated the results to all of China – an indefensible and invalid inference that was pointed out by the scholar from China (see stakeholder views in this eBrief).

Media claims surrounding the most recently released 2012 PISA results are similar, fueling public concerns that as the students of the US are ranked below those of Shanghai, the US education system is declining in its effectiveness, and that the nation could lose its status as a global economic force in comparison to China (Figure 2). Such claims are erroneous because, first, Shanghai's sample of students are not comparable to those of China as a whole. Second, economic policy researchers such as Feuer (2013) have shown that linkages are tenuous between student performance on international large scale achievement tests and macro-economic indicators of productivity of nations. While such claims are attention grabbers – they have little evidentiary support. PISA results are based on descriptive surveys and do not permit direct causal inferences between student test scores and overall quality of schools and schooling processes.

Acknowledgements

The first volume of eBriefs "Understanding Validity Issues Around the World" is produced via a partnership between the Assessment and Evaluation Research

Initiative (AERI) at Teachers College, Columbia University and the National Education Policy Center (NEPC) at the University of Colorado, Boulder. The inaugural AERI conference that generated the first series of eBriefs was co-sponsored by Educational Testing Service, Teachers College, and the National Science Foundation. Websites: www.tc.edu/aeri and nepc.colorado.edu

Note

1. Because this attempt to distill a great deal of information will necessarily lose some nuance and detail, readers are encouraged to access the original articles listed in the References section. AERI-NEPC eBriefs, or electronic versions of the items in the series are also available at the AERI and NEPC websites.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999), *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, DC.

Chatterji, M. (2013), "Insights, emerging taxonomies, and theories of action toward improving validity", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 273-307.

Feuer, M. (2013), "Validity issues in international large-scale assessments: 'truth' and 'consequences'", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 197-215.

Messick, S. (1989), "Validity", in Linn, R. (Ed.), *Educational Measurement*, Macmillan, New York, NY, pp. 13-103.

Gitomer, D. (2013), "International parallels in response to accountability requirements: validity considerations", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 173-183.

Pallas, A. (2013), "Reflections on rationality, evidence and issues of validity: evaluating schools and teachers", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 163-171.

Visscher, A. (2013), "Evaluation-centered school improvement: potential, prerequisites, and validity considerations", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 101-135.

Wandall, J. (2013), "Education, testing, and validity: a Nordic comparative perspective", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 137-161.

Yarborough, D.B., Shulha, L.M., Hopson, R.K. and Caruthers, F.A. (2010), *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users*, 3rd ed., Sage, Thousand Oaks, CA.

Yon, H. (2013), "Addressing validity and other challenges in evaluation-centered school improvement models: possibilities in the Netherlands and Malaysia", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 185-194.

About the authors

Beatrice L. Bridglall is a Fulbright Specialist in Higher Education with the Council for International Exchange of Scholars (CIES) and currently teaches at Montclair State University in Montclair, New Jersey. Her most recent book is: Teaching and Learning in Higher Education: Studies of Three Student Development Programs (2013). She received her doctorate in education from Teachers College, Columbia University in 2004. Beatrice L. Bridglall is the corresponding author and can be contacted at: Bridglall@exchange.tc.columbia.edu

Jade Caines is an Assistant Professor at the University of New Hampshire, and completed her doctoral training at Emory University in 2011. Her primary research interests relate to educational measurement. She studies validity and fairness issues as it relates to standard setting, instrument development, and stakeholder participation.

Madhabi Chatterji is Associate Professor of Measurement, Evaluation, and Education and the founding Director of the Assessment and Evaluation Research Initiative (AERI) at Teachers College, Columbia University. Dr Chatterji's publications focus on the topics of instrument design, validation, and validity; evidence standards and the "evidence debate" in education and the health sciences; standards-based educational reforms; educational equity; and diagnostic classroom assessment.