The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0968-4883.htm

SECTION 1. ASSESSMENT POLICY: UNDERSTANDING VALIDITY ISSUES AROUND THE WORLD

(GUEST EDITORS: PROFESSOR MADHABI CHATTERJI AND PROFESSOR KEVIN G. WELNER)

Understanding validity and fairness issues in high-stakes individual testing situations

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

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

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

Abstract

Purpose – This policy brief discusses validity and fairness issues that could arise when test-based information is used for making "high stakes" decisions at an individual level, such as, for the certification of teachers or other professionals, or when admitting students into higher education programs and colleges, or for making immigration-related decisions for prospective immigrants. To assist test developers, affiliated researchers and test users enhance levels of validity and fairness with these particular types of test score interpretations and uses, this policy brief summarizes an "argument-based approach" to validation given by Kane.

Design/methodology/approach – This policy brief is based on a synthesis of conference proceedings and review of selected pieces of extant literature. To that synthesis, the authors add practitioner-friendly examples with their own analysis of key issues. They conclude by offering recommendations for test developers and test users.

Findings – The authors conclude that validity is a complex and evolving construct, especially when considering issues of fairness in individual testing contexts. Kane's argument-based approach offers an accessible framework through which test makers can accumulate evidence to evaluate inferences and arguments related to decisions to be made with test scores. Perspectives of test makers, researchers, test takers and decision-makers must all be incorporated into constructing coherent "validity arguments" to guide the test development and validation processes.

Originality/value – Standardized test use for individual-level decisions is gradually spreading to various regions of the world, but understandings of validity are still uneven among key stakeholders of such testing programs. By translating complex information on test validation, validity and fairness issues with all concerned stakeholders in mind, this policy brief attempts to address the communication gaps noted to exist among these groups by Kane.

Keywords Fairness, Validity, High stakes testing, Individual testing, Assessment policy

Paper type Research paper

Quality Assurance in Education
Vol. 22 No. 1, 2014
pp. 5-18
Emerald Group Publishing Limited
0968-4883
DOI 10.1108/QAE-12-2013-0054

Introduction

All of us would like tests and assessment reports to be valid and fair, especially when these instruments serve as a gateway for admitting individuals into specialized or competitive programs[1]. But do the terms – validity and fairness – mean the same thing to all stakeholders concerned with these test use contexts? In this policy brief, we examine how the validity of tests, scores and assessment reports is related to issues of fairness and equity viewed from three different stakeholder perspectives (after Kane, 2013):

1. Test makers – or those who design and conduct research on standardized tests and testing programs, generating information that is fed into "high-stakes" decisions that could impact the future of individual examinees.
2. Test takers – or those who opt to take such a test with a specific goal in mind.
3. Decision makers – or those who rely on the test results and reports of such testing programs to meet their decision-making goals.

More specifically, our purpose is to look closely at tests designed and used at the individual level, either in educational, professional or social contexts. In all these applications, the test results are used for taking some kind of "high-stakes" action – or decisions that could have lasting and sometimes, adverse consequences for test-takers either by delaying opportunities or closing doors to future opportunities in school, work, or life. Some examples of such high stakes actions include:

- identifying qualified individuals for merit recognitions or awards;
- promoting students to the next educational level;
- selecting and admitting persons into colleges and higher education programs, for job opportunities, or for immigration into countries of their choice; and
- in certifying or licensing professionals for particular occupations like teaching, counseling, or medicine.

To help enhance levels of validity and fairness in these particular test use contexts, this policy brief summarizes an "argument-based approach" to validation forwarded by Kane (2013).

Who and what this eBrief speaks to

The targeted audience for this eBrief is broad. In addition to test makers and affiliated researchers, we include both test takers and decision makers as important stakeholders in the test use contexts identified. All three groups would very likely care a great deal about the tests, the quality of the information they yield, and the possible repercussions of misuse or misinterpretion of test results, for themselves or their organizations.

Test makers could benefit from the information in this eBrief by gaining insights into how decision-makers and test-takers typically think, for example:

1. What is the logic undergirding the actions that decision-makers take with results of such tests?
2. What are the validity, fairness and equity issues from the perspective of the test-taker or decision-maker in these contexts?

(3) What kinds of validity evidence will help ensure that a test's scores and score-based reports have the properties to support the actions intended by the typical test user and/or decision-maker?

Prospective decision makers and test takers could similarly benefit by gaining a window into the thinking and recommendations of established measurement and policy scholars on the following types of issues:

(1) What steps should standardized testing programs take to ensure that a test yields information of sufficiently high quality to support the intended actions with test results?

(2) What technical procedures are available to help make the test-based information more fair and equitable for diverse and global test takers?

(3) What are some philosophical, political and societal tensions that currently exist surrounding this form of testing, views on what is important to measure in education, and how to use the results to improve social conditions?

Method

This policy brief is based on a synthesis of conference proceedings and selected pieces of extant literature. It begins by summarizing perspectives of an invited panel of measurement and policy experts on the topic. To that, we add practitioner-friendly examples with their own analysis of key issues. We conclude by offering our own thoughts and recommendations.

The content of this eBrief is derived from the keynote presentation and chapter by Michael T. Kane (Kane, 2013), reactions by four discussants (Gordon, 2013; Georges, 2013; Ercikan and Oliveri, 2013; von Davier, 2013), and audience comments at a March, 2012 conference held at Teachers College, Columbia University, titled: "Educational assessment, accountability and equity: conversations on validity around the world." Following a summary of the main ideas from the panelists, we provide our own thoughts on the validity, fairness and test-use issues in these contexts.

A summary of main themes

Michael Kane's main ideas

Kane (2013) recommends that test makers follow an argument-based approach to investigate validity and fairness issues for particular tests and test results. He describes the proposed interpretation or use of the test scores as a chain of inferences leading from a test taker's score to an interpretation of that score and a decision to be made about that person. All of the inferences and assumptions inherent in the interpretation or use have to be plausible for the interpretation or use to be considered valid. The inferences can be thought of as "the spans of a bridge" that must be crossed to get from the score to the score interpretation and use. "If one span of the bridge is out, the bridge is out" (Kane, 2012). For a test to be judged as valid for its purposes, there must be solid evidence that supports the inferences inherent in decisions based on the test's results, particularly when the consequences of poorly made decisions can be serious.

Using this argument-based approach (described briefly below) could help test makers identify and evaluate all the proposed interpretations and uses of a test, thereby making it more likely that inferences made from a test's scores and reports are valid and are fair to all test takers. To apply this approach, Kane suggests that test makers and measurement researchers first lay out the claims that test takers or decision makers wish to make from the test's results. For example:

(1) Would the results simply reflect a person's competence in one subject area? Or, would they also predict how the person will perform in other areas in future?

(2) Would the results convey that a person could perform equally well in a real world setting?

(3) To be considered minimally qualified for admission, promotion or certification, what performance level must the person achieve on the test?

(4) Are the scores reliable enough to support the proposed uses and decisions? How much error will decision-makers and test takers be prepared to tolerate at the score that serves as the cut-point for decisions?

For test designers and researchers, an "interpretive argument" lays out the chain of inferences and assumptions that are proposed, starting from the test taker's observed performance on a test, leading to the claims and decisions made by test users based on the test scores. By evaluating the plausibility of the inferences and assumptions inherent in the proposed interpretations and uses of the scores through logical and empirical "validation studies", test makers can gauge the extent to which the test will provide information that is valid and useful for the actions to be taken.

The evidence from validation studies also tells us to what degree the test scores and results will likely be fair and dependable (reliable) for the proposed uses. Kane reminds us that it is not just the test that needs to be validated, but the proposed uses and interpretations based on the scores that must be validated by researchers before a test can be put to use.

Kane cites Toulmin's (1958) framework as a guide for applying his approach to validation. The backing for the individual inferences defines the evidence that measurement researchers would need to collect to support the claims about what test scores mean. Validation could include results from a variety of studies, such as content validation studies, generalizability studies, reliability studies, or criterion-related validity studies. Each type of study would yield a particular type of evidence, which must be holistically evaluated to decide whether the test is able to support the intended actions.

Finally, Kane reminds us that there could be exceptions to the claims about what a test score means in particular test-use contexts. For example, if the test taker has a disability, he/she might need an accommodation. If the student is denied the accommodation, the intended meaning of the score and test-based report might be compromised. Test takers and decision makers must be aware of such limits to meanings of test-score reports and should interpret the results accordingly.

Kane presents the argument-based approach to validation as a single general model for validation, applicable to a variety of proposed test-score interpretations and uses in different populations and assessment contexts. Although the approach can be used for validating scores from any test or assessment tool in a variety of decision-making contexts, the primary focus in his article (Kane, 2013) is on high-stakes testing of individuals.

Kane's three perspectives; fairness in test-based decisions

Kane defines "fairness" as a part of validity. Fairness means that scores from a test should convey consistent meanings for all groups or sub-groups that take a test. For standardized test makers, fairness investigations are typically also a part of the validation research program.

Kane specifies three perspectives on testing that should be considered by everyone involved so as to obtain higher levels of validity and fairness in practical settings (see also, Dorans, 2012):

(1) A “measurement perspective,” reflecting the typical mindset of test developers and measurement researchers in standardized testing programs, describing how this group tends to approach the tasks of test development and validation.

(2) A “contest perspective,” reflecting the typical mindset of test takers, who view passing a high stakes test as being similar to winning a contest to help them reach their goals.

(3) A “pragmatic perspective” or the mindset of typical decision makers, who count on a test to be cost-efficient, objective, fair and dependable for taking necessary actions from an institutional perspective, without adverse or untoward repercussions.

When it comes to fairness, Kane reminds us that each perspective supports equitable treatment of test takers and consistency in score meanings for all, but in different ways and to varying degrees. First, the measurement perspective is primarily concerned with the meaning, accuracy and the precision of test scores. For example, test developers and measurement researchers typically would want to know to what extent a test taker’s score on a math achievement test represents the “true” measure of his/her mathematical ability. Also, they would be interested in estimating the extent to which the test scores will remain consistent over time, and how much error is associated with test scores?

The contest perspective is more concerned with getting the highest score possible on the exam. For example, a test taker will try to get the highest score possible on a math achievement test, especially if it will help determine his/her admission into a competitive high school. The key for this stakeholder is that the test is fair in providing every test taker an equal chance to pass. A test taker wants to “win” at taking the test and also wants to assume that it is a level playing field for everyone.

Finally, the pragmatic perspective of a decision maker is concerned with the same goals as the test makers, but for very different reasons. Meaning, accuracy and precision are also necessary here in order for the decisions themselves to benefit the organization. An organizational decision-maker seeks the most qualified and able candidates for meeting the larger goals and mission of the organization. At the same time, these decision makers are also concerned about public perceptions of unfairness or discrimination against some groups, and potential legal repercussions of actions based on test results. From this viewpoint, a test must be able to support high-stakes decisions that are fair.

Even when tests yield technically defensible results (i.e. scores are accurate and precise), significant differences in score patterns of particular subgroups have led to serious policy discussions on “achievement gaps” in the US. With regard to test-score gaps by race and wealth on many standardized tests, for example, there are certainly accusations that the tests may not be fair. But it is also well accepted that the test score gaps reflect real differences in what is being measured (e.g. math skills or language arts skills), with the evidence replicated on different measures and levels of education.

That evidence is now being interpreted as indicative of underlying and persistent opportunity gaps among different groups of test takers. In other words, the gaps revealed by standardized tests are broadly perceived as serious educational and social problems, but the main issue in policy debates is how best to close the observed gaps (Carter and Welner, 2013).

Both the test maker and the decision maker might consider such group differences to be problematic. All three of these perspectives on testing are "legitimate", according to Kane (2013), and a high-stakes testing program must satisfy each one in order to be effective.

Reactions to Kane's position: main ideas

Edmund W. Gordon, Sébastien Georges, Kadriye Ercikan and Alina von Davier offered reactions to Michael Kane's thinking on these matters.

Gordon points out that Kane's comprehensive analysis of validity, fairness, and testing issues echo the tensions operating within the Gordon Commission on the Future of Testing (2013). The commission is tasked with understanding current issues in student assessment, with a view towards informing future teaching and learning practices. In light of rapid changes in education and society, Gordon (2013) expresses concern that if current trends in standardized assessment programs continue, they will likely become obsolete. The focus should change to learning in an increasingly complex world. While we can still honor notions of standardization and universalization, we need to recognize that tests may be less universally applicable. Given changes in the way in which we think of and generate knowledge about how students learn, he emphasizes the crucial need to consider ideas of validity, reliability and fairness through different lenses. He particularly stressed the need to design assessments suited to differences in student contexts and backgrounds.

Georges (2013), von Davier (2013), and Ercikan and Oliveri (2013) also affirm the importance of both understanding and interpreting issues of validity, particularly in light of how, why, and for whom assessment decisions are made.

For instance, Georges points out that different assessment actors often have different test information needs and very different points-of-view about what kind of a test is valid. In France, for example, the public tends to view essays as more valid than multiple choice tests, regardless of whether these tools are standardized and well-designed from a technical standpoint. A test and testing program should gather validity evidence to make sure all the different ways in which the same test (and scores) will likely be used have been validated. There should be suitable kinds of evidence to support test use requirements of all stakeholders. Some tests, however, may be more or less valid for some purposes, and test users and decision makers must learn to accept the limitations of tests and what testing programs can realistically deliver (Georges, 2013; see also Chatterji, 2013a, b).

von Davier (2013) highlights different ways in which psychometric studies can be helpful in improving validity and fairness, particularly through the processes of standardization and test-form equating with diverse populations. If test design is implemented in a specific way, we can effectively preempt and address the validity argument during evaluations of a test. The sampling of items and people tested should be well-matched to get maximum validity. In a recent study Duong and von Davier (2012) examined the use of a test on a more heterogeneous population than originally intended. They concluded that the secondary use of the test was not appropriate, but if a different type of test design had been considered and used (such as a computer adaptive test with a multi-stage design) and coupled with an altered sampling strategy for equating, a more comprehensive accounting of examinee background differences may have been possible in the score reports. These actions would, in von Davier's view, improve the validity and fairness of the test and test-score reports for heterogeneous test-taking populations.

Ercikan and Oliveri (2013) are also concerned with the technical aspects of fairness, as investigated using differential item functioning (DIF) in heterogeneous sub-populations. DIF occurs when examinees from various groups with similar ability levels perform differently on particular test items. These items, if detected before a test is used, can be deleted or revised to make the test more fair and the resulting information more valid for the proposed uses.

Issues of fairness and potential bias emerge when DIF detection methods fail. When this happens, items underestimate the attributes for one group (when compared to another) or measure attributes that lie outside the domain for the assessment for some students. Given added sources of diversity in global populations that take standardized tests today, Ercikan and Oliveri offer a new method for detecting items that show DIF. They caution that we must be very specific about our claims regarding validity and fairness with diverse populations. DIF detection, if inaccurate, can lead to the development of biased tests and threaten the consistency of score meanings across subgroups.

Stakeholder views

Audience concerns

Audience members at the conference expressed some concerns worth noting here.

A representative from a youth life skills and leadership program in India had concerns about the validity of their student-selection process based on portfolio-based assessments, with decisions made by individual schools. Each school is likely to want their favorite student to receive a scholarship, prompting the question, "What rules do we use to minimize bias and maximize efficacy in selecting 11th and 12th grade students for our leadership program?" Kane suggested a social moderation alternative where all involved schools agree on a common set of standards they use to evaluate students. Ercikan recommended also evaluating students who are not selected into the program and identifying potential patterns or trends that describe their assessed abilities. Such careful evaluations can provide formative feedback and facilitate long-term fairness in making decisions. Similarly, von Davier proposed a focus on collecting "predictive validity evidence" where students would be evaluated the year after being admitted to determine if the selection process is identifying students who would actually succeed. To what extent did the scores, used for awarding scholarships, correlate with their future levels of student success? She cautions, however, that portfolio-based assessments tend to be less reliable that other forms of testing.

Another question came from an educational consultant and former university professor in measurement, who asked the following: "Do you think validation should remain within the purview of the educational measurement profession, or do you envision [that] concerns of stakeholders that fall within contest and pragmatic perspectives could actually work their way into interpretive and validity arguments? If so, can you give us an example of what that might look like?"

Kane responded that many changes have already been made to incorporate varied perspectives into validity arguments in testing programs. He believes it is definitely possible for more to be done, but he acknowledged that it is a difficult task, especially in our world's changing demographic landscape. Gordon asserted that we need to use a kaleidoscope model when making high-stakes decisions where we examine the many "hues" of a test taker.

A third audience question explored the struggle between assessing students on established standards and assessing them on content they have not had the opportunity to learn. Gordon responded that the absence of opportunity to learn is a moral issue: "We are investing much too much in assessing the outcomes of inadequate opportunities to learn, and we ought to be investing more effort in improving opportunities to learn in school." He further contended that we should shift from viewing school and life as a duality, and find ways to view both contexts together. Therefore, measurement of things learned in life and school could be the new goal.

Conclusions

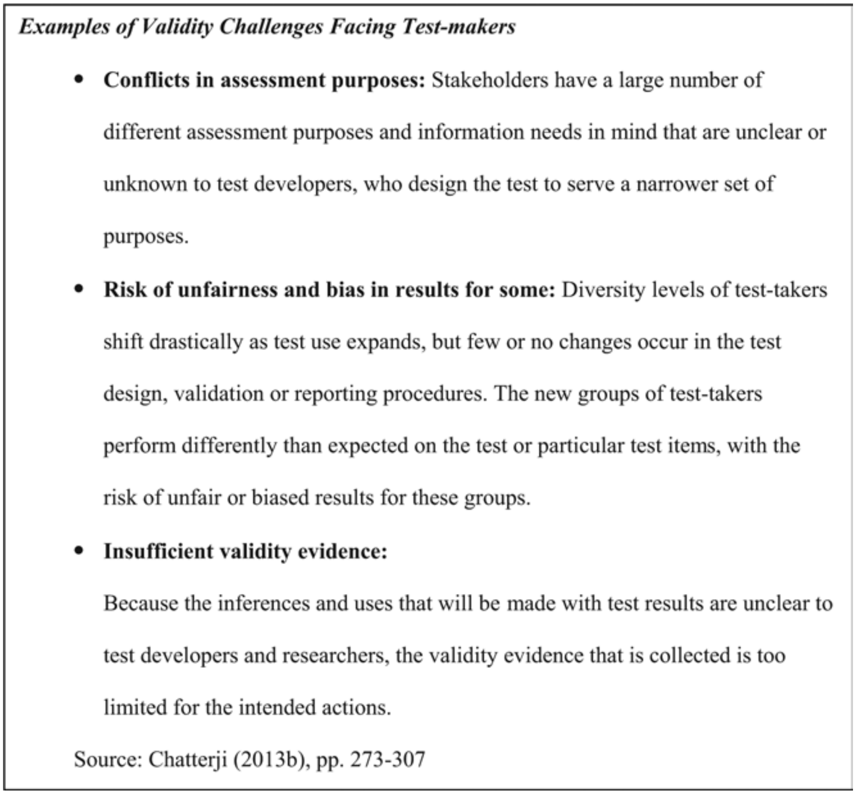
Our thoughts

Validity is a complex and evolving construct, especially when considering issues of fairness. Based on a content analysis of validity issues discussed in the full-length chapters, Chatterji (2013a, b) identified some of the current challenges facing test makers and test users (the latter group including test takers and decision makers). The challenges are summarized in Figures 1-2, with our recommendations for test developers and test users in Figures 3-4.

Stakeholders may have multiple and different assessment purposes and information needs that remain unknown to test developers, who end up designing the test to serve a narrower set of purposes. In other instances, the diversity levels of test-takers shift drastically as test use expands, but few or no changes occur in the test design, validation or reporting procedures. The new groups of test-takers perform differently than expected on the test or particular test items, with the risk of unfair or biased results for these groups. All such oversights lead to collection of validity evidence that is too limited for the intended actions by decision-makers (see Figure 1). The argument based approach offers a viable solution for addressing these issues systematically.

Aligning perspectives of different stakeholder groups may be a challenge when they "talk past one another" (Kane, 2013). Technical limitations of a test, the testing program or test reports are often overlooked, misunderstood or undermined by decision makers and test users who are guided by other agendas and needs. They want to believe that tests are error-free, although results ought to always be interpreted with the understanding that there will always be some uncertainty due to measurement error (see Figure 2).

Kane's argument-based approach offers an accessible framework through which test makers can accumulate evidence to evaluate inferences and arguments. But, as one audience member at the conference suggested, the approach falls primarily within the "measurement perspective." If Kane's approach is to be broadly applicable, the "contest" and "pragmatic" perspectives must also be incorporated into constructing coherent "validity arguments." Ideally, this should occur during the test development process, well before tests are disseminated and used widely. We hope our



Examples of Validity Challenges Facing Test-makers

- **Conflicts in assessment purposes:** Stakeholders have a large number of different assessment purposes and information needs in mind that are unclear or unknown to test developers, who design the test to serve a narrower set of purposes.
- **Risk of unfairness and bias in results for some:** Diversity levels of test-takers shift drastically as test use expands, but few or no changes occur in the test design, validation or reporting procedures. The new groups of test-takers perform differently than expected on the test or particular test items, with the risk of unfair or biased results for these groups.
- **Insufficient validity evidence:**
  Because the inferences and uses that will be made with test results are unclear to test developers and researchers, the validity evidence that is collected is too limited for the intended actions.

Source: Chatterji (2013b), pp. 273-307

Figure 1.
Example of validity challenges facing test-makers

recommendations for various stakeholder groups will facilitate communications among them (see Figures 3-4).

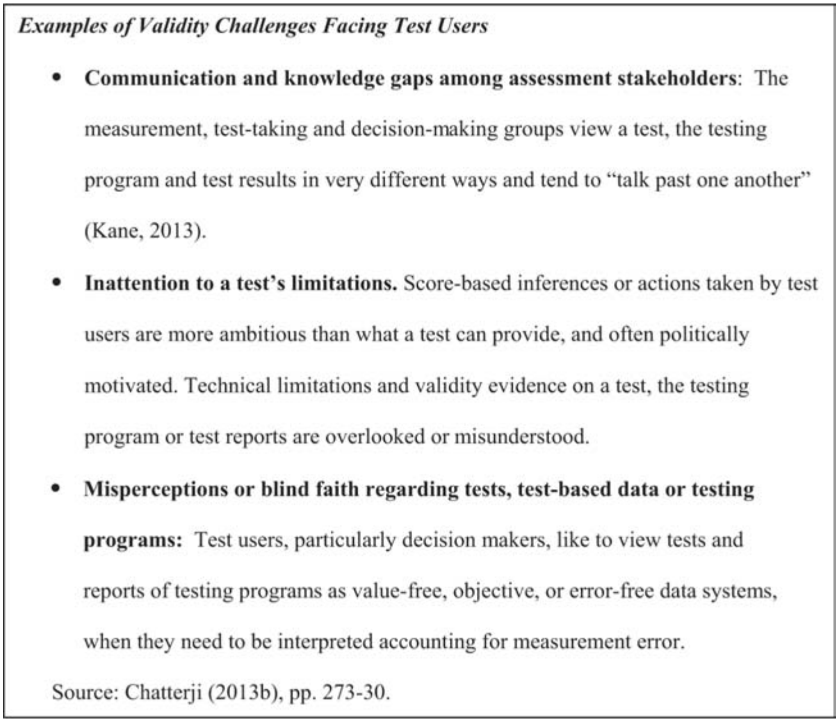
In sum, we endorse the argument based approach to validation. Because such an approach to validation may seem somewhat technical for the typical teacher or educational practitioner, we provide the following example illustrating how levels of fairness and validity could be maximized as understandings build among concerned stakeholders.

An applied example

> Malik is an 8th grader. He is a student in Ms Johnson's Algebra I class. He needs to master the objectives (knowledge and skills) in that course.

In this scenario, the test designer and user is a teacher making "high stakes" grade promotion decisions for students in middle school. She begins with inferences about student achievement based on a classroom assessment. That interpretation is then extrapolated to a broader inference about learning in a domain tapped by a state test, which leads to a decision about promotion. Initially, let's consider just the classroom assessment.

Applying Kane's approach, we can propose the following inferential chain:

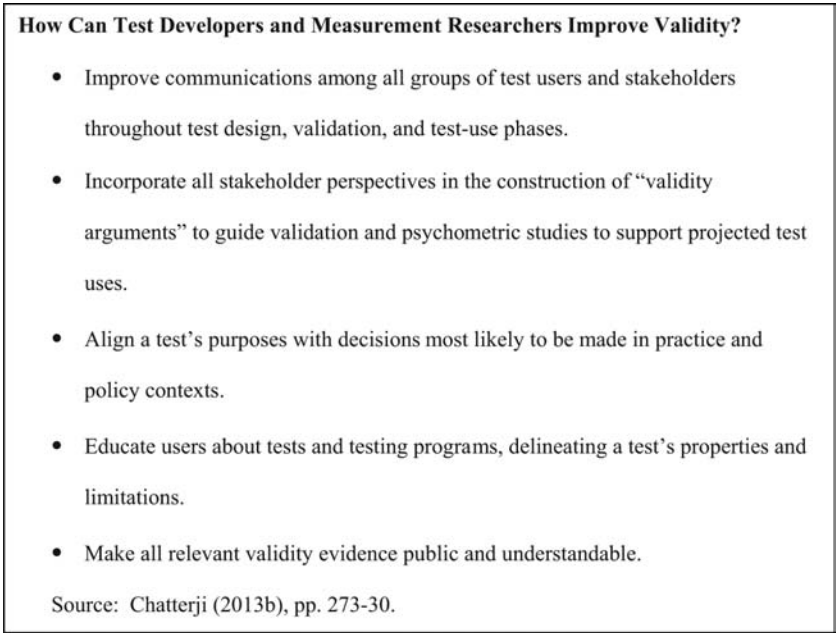


Examples of Validity Challenges Facing Test Users

- **Communication and knowledge gaps among assessment stakeholders**: The measurement, test-taking and decision-making groups view a test, the testing program and test results in very different ways and tend to "talk past one another" (Kane, 2013).
- **Inattention to a test's limitations.** Score-based inferences or actions taken by test users are more ambitious than what a test can provide, and often politically motivated. Technical limitations and validity evidence on a test, the testing program or test reports are overlooked or misunderstood.
- **Misperceptions or blind faith regarding tests, test-based data or testing programs:** Test users, particularly decision makers, like to view tests and reports of testing programs as value-free, objective, or error-free data systems, when they need to be interpreted accounting for measurement error.

Source: Chatterji (2013b), pp. 273-30.

Figure 2.
Example of validity challenges facing test users

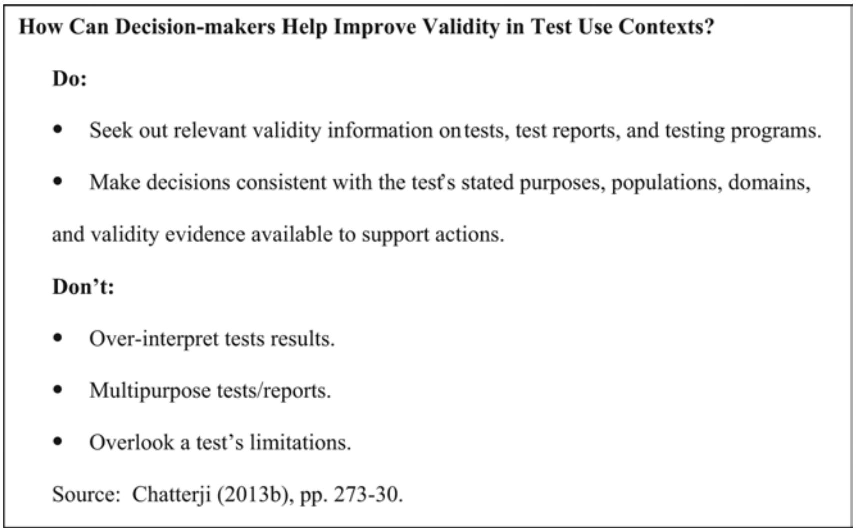


How Can Test Developers and Measurement Researchers Improve Validity?

- Improve communications among all groups of test users and stakeholders throughout test design, validation, and test-use phases.
- Incorporate all stakeholder perspectives in the construction of "validity arguments" to guide validation and psychometric studies to support projected test uses.
- Align a test's purposes with decisions most likely to be made in practice and policy contexts.
- Educate users about tests and testing programs, delineating a test's properties and limitations.
- Make all relevant validity evidence public and understandable.

Source: Chatterji (2013b), pp. 273-30.

Figure 3.
How can test developers improve validity?



How Can Decision-makers Help Improve Validity in Test Use Contexts?

Do:

- Seek out relevant validity information on tests, test reports, and testing programs.
- Make decisions consistent with the test's stated purposes, populations, domains, and validity evidence available to support actions.

Don't:

- Over-interpret tests results.
- Multipurpose tests/reports.
- Overlook a test's limitations.

Source: Chatterji (2013b), pp. 273-30.

Figure 4. How can decision makers help improve validity in test use contexts?

> Malik's observed performance on Ms Johnson's math test → Malik's score on Ms Johnson's math test → Score indicates level of mastery in algebra objectives covered in Ms Johnson's course.

Examining this inferential chain from Ms Johnson's perspective (the test maker), the focus is mainly on the performance at hand, and there are no general interpretations beyond the domain tested in the algebra class. This interpretation of the score makes validity claims only with regard to the content tested. To make a claim that her test's scores are valid, Ms Johnson needs to establish a good match between her course objectives, what she taught in class, and the test questions themselves (assuming these are well-written and clear). Assuming she has a consistent scoring procedure, this may be all the evidence she needs to claim that the scores indicate students' levels of mastery of the content.

However, suppose we also want to propose that:

> Malik's performance on Ms Johnson's Algebra test → Malik's math score on a statewide math achievement test → Malik's math competence and abilities on state's standards → Eligibility for promotion to next grade.

Now, in this more elaborate chain of score interpretations, several added pieces of evidence may be necessary in order to support these inferences that go beyond the classroom test to a broader test domain and to the high-stakes action (e.g. greater scoring accuracy and reliability evidence, sufficient correlations with the state test scores, generalizability studies, and so on). For test makers, thinking clearly about the content of the test should be a critical, initial consideration (Lissitz and Samuelsen, 2007).

The student, Malik's focus would be on "winning the contest" (Kane, 2013) or passing the test. Given this test taker perspective and Kane's focus on fairness, validity arguments should take into account contextual factors related to Malik and the testing situation (e.g. if Malik had a disability that had not been adequately accommodated, or if he has a language or cultural difference that could alter score meanings, and so on). While Malik may not care about the rest of the validation process – the scientific effort to get to his "true" score on the state test's domain – this should be a relevant concern for standardized test makers (Kane, 2013).

Examining the inferential chain from the test taker's perspective, the biggest concern would be that the exam reliably measures Malik's math competence and abilities, is adequately scored, and has clear rules that are known ahead of time to all test takers like himself. The outcome should be one that can be empirically verified as fair for all (Dorans, 2012). Now let's turn to the decision maker.

Decision makers (e.g. a school-based committee making promotion decisions) are often concerned with the same goals as test makers: do Malik's test scores on the classroom assessment and the statewide math achievement test represent a "true" measure of his mathematical abilities? Does his score warrant promotion to the next grade level? They are concerned about these aspects, however, for a very different reason than the test maker. From their pragmatic perspective, the statewide math test must be consistent and objective in order for high-stakes promotion decisions to be made fairly for all students. So, Malik's score must be accurate and reliable enough so that decision makers can make both sound and fair decisions regarding his eligibility to advance to the next grade level and succeed. Further, if compared to others in the same class, no student should feel that performance was gauged for some by different or variable standards. The Kane (2013) approach focuses on these legitimate policy concerns, setting aside the various political elements that may also factor in to the decision-making in this area (e.g. the perception of taking a tough stance on accountability), as well as the educational soundness of promotion versus grade retention policies (see Moreno, 2012).

To help decision makers act responsibly, the backing (or the evidence) that supports the test scores must be easily interpretable for decision-making audiences. If the backing in this example includes an empirical study relating the statewide math achievement test scores in middle school to course grades in high school, then there would be adequate support for the claim that the capacities on the class test and state assessment overlap sufficiently to predict future performance. This "predictive validity evidence" is one kind of evidence to support a fair policy decision. The challenge is for test makers to create easily interpretable materials that explain the chain of inferences within the validity argument for decision makers to use.

To what extent can a student's individual context be taken into account in policy-making and in standardized testing practice? There is, no doubt, a lack of clarity and consensus regarding how we should address the differences among the measurement, contest, and pragmatic perspectives. What is clear, however, is that mutual understandings of these different perspectives on validity, testing, and fairness are critical in any efforts we make to ensure accurate interpretations and uses of "high stakes" test scores and reports.

Another thing is certain: as conversations about validity, equity, and fairness continue to expand amongst test makers, test takers, and decision makers, things are likely to improve. The argument-based approach gives us a conceptual framework with which to think about the tests and their intended purposes – helping us move one step closer to using test-based information better so as to justify more fair and valid decisions in applied and policy contexts (Chatterji, 2013a, b).

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. This series of policy briefs, Understanding Validity Issues around the World, is a joint product of AERI and NEPC. eBriefs, or electronic versions of the items in the series are also available at the AERI and NEPC websites.

References

Carter, P.L. and Welner, K.G. (Eds) (2013), *Closing the Opportunity Gap: What America Must Do to Give All Children an Even Chance*, Oxford University Press, New York, NY.

Chatterji, M. (2013a), "Bad tests or bad test use: a case of SAT® use to examine why we need stakeholder conversations on validity", *Teachers College Record*, Vol. 115 No. 9, pp. 1-10.

Chatterji, M. (2013b), "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.

Dorans, N.J. (2012), "The contestant perspective on taking tests: emanations from the statue within", *Educational Measurement: Issues and Practice*, Vol. 31 No. 4, pp. 20-37.

Duong, M.Q. and von Davier, A.A. (2012), "Observed-score equating with a heterogeneous target population", *International Journal of Testing*, Vol. 12 No. 3, pp. 224-251.

Ercikan, K. and Oliveri, M.E. (2013), "Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 69-85.

Georges, S. (2013), "Fairness and validity from the viewpoints of different assessment actors: a French reaction to Michael Kane's contribution", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 87-91.

Gordon, E.W. (2013), "The Gordon Commission's perspectives on the future of assessment", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 93-99.

(The) Gordon Commission on the Future of Assessment in Education (2013), "To assess, to teach, to learn: A vision for the future of assessment", Technical Report, Executive Summary, Princeton, NJ.

Kane, M. (2012), "Validity, fairness, and testing", paper presented at the inaugural invitational conference, co-organized by AERI and ETS, titled "Educational Assessment, Accountability, and Equity: Conversations on Validity around the World" held at Teachers College, Columbia University on March 28-29.

Kane, M. (2013), "Validity and fairness in the testing of individuals", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 17-53.

Lissitz, R.W. and Samuelsen, K. (2007), "A suggested change in terminology and emphasis regarding validity and education", *Educational Researcher*, Vol. 36, pp. 437-448.

Moreno, A. (2012), *Does Retention (Repeating a Grade) Help Struggling Learners?*, Marisco Institute for Early Learning and Literacy, Denver, CO, available at: www.du.edu/marsicoinstitute/policy/Does_Retention_Help_Struggling_Learners_No.pdf

Toulmin, S. (1958), *The Uses of Argument*, Cambridge University Press, Cambridge.

von Davier, A. (2013), "Standardized assessments and implications for improving test and sampling design: Applying Kane's principles for validation", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 55-67.

About the author

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. Jade Caines is the corresponding author and can be contacted at: jade.caines@unh.com

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.

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.