Validity and Test Use:
An International Dialogue on Educational Assessment, Accountability and Equity

Chapter 1
Global Forces and Educational Assessment – A Foreword on Why We Need an International Dialogue on Validity and Test Use

Madhabi Chatterji
Teachers College, Columbia University

Abstract

While educational testing is a ubiquitous and international phenomenon today, the validity (or the meaningfulness and defensibility) of test scores and assessment reports is frequently compromised in test use settings by conditions and factors that are still largely undocumented and not well understood. Some of these factors might concern the test itself (inadequate domain frameworks, poorly constructed items or scales), but others lie completely outside the test, having to do with ways in which particular instruments or information generated by the assessments are employed in applied environments. This volume examines the overlapping topics of validity, validation, and use of tests in educational practice, policy, and evaluative contexts around the globe. Reflecting perspectives of a cadre of international policy experts as well as measurement and evaluation scholars, it raises new questions and suggests innovative directions, solutions, and frameworks for improving validity at the nexus of assessment design and use. The Foreword introduces the section themes, authors and individual chapters in this volume, using illustrative case studies to highlight validity issues prevalent in three test use contexts treated in the book’s sections.
Global Forces and Educational Assessments – A Foreword on Why We Need an
International Dialogue on Validity and Test Use

Purpose of the Volume

While educational testing is a ubiquitous and international phenomenon today (Kellahan & Greaney, 2001; Ryan & Feller, 2009; Tamassia & Adams, 2009), the validity of test scores and assessment reports is frequently compromised in various test use settings by conditions and factors that are still largely undocumented or poorly understood. Some of these factors might concern the test itself (e.g., inadequate domain frameworks, poorly constructed items or scales), but others lie completely outside the test, having to do with ways in which particular instruments or information generated by the assessments are employed in applied environments. There could be serious consequences to validity oversights, especially in cases where the test data are fed into high stakes decisions that directly affect the lives of test-takers, teachers, school leaders, school systems, and indirectly, society at large.

This volume examines the overlapping topics of validity, validation, and use of educational tests in practice, policy, and evaluative contexts around the globe. Few would question the utility of educational tests today, given the exponential rise in the levels and kinds of test use internationally. Fueled by an active media, reports of educational testing efforts like the Programme for International Student Assessment (PISA) are now a part of everyday public conversations in both the developed and developing world (see Dillon, 2010; Layton & Brown, 2012). As educational assessment and accountability programs mushroom and multiply globally, this volume seeks answers to two broad questions: To what extent are these various applications
of tests and assessment information valid? Also: What can be done to optimize validity internationally, and with it, fair and equitable assessment and evaluation practices in education?

The book compiles three invited papers based on keynote speeches along with four reaction papers to each. The papers were originally presented at an invitational assessment conference held in March, 2012, co-sponsored by the Educational Testing Service, the National Science Foundation, and Teachers College, Columbia University through its Assessment and Evaluation Research Initiative (AERI). It is my privilege to introduce the chapters in this book and the collected works as a whole.

Orientation and Some Definitions

To begin, a few words on the orientation of the volume and some clarifying definitions. As the title states, the book represents an international dialogue. As such, it offers a range of perspectives by a cadre of experts knowledgeable about the measurement and evaluation sciences, as well as accountability trends and policies affecting educational systems in their regions and around the world.

The reader will find that the present book is not about validity theory. Instead, it is concerned with evaluating and enhancing the meaningfulness and utility of information generated by visible tests and assessment programs in applied contexts around the world. The premise is that, as assessment instruments are moved from relatively constrained test design and validation settings to more diverse and uncontrolled test use environments, validity can be and is routinely challenged. To tackle these types of challenges, the distinguished contributors rely on aspects of validity theory most pertinent to practical aspects of test design, validation and use in
“real” world settings. Readers will encounter three, fairly well-defined test use contexts that correspond with the three sections of the volume (detailed next).

Given our applied focus, in this volume, the term *validity* refers to the meaningfulness and defensibility of “*inferences*” and “*actions*” based on test scores, given the constructs and domains intentionally tapped by particular tests and assessment instruments (after Messick, 1989, p. 13, emphasis in original). This formulation of validity has been endorsed by Kane (2006) in the fourth edition of *Educational Measurement* (Brennan, 2006), as well as in the most recent *Standards for Educational and Psychological Testing* (American Education Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME, 1999)). Our deliberate leaning towards one (and still dominant) view of validity in modern measurement theory is not intended to discount newer theoretical formulations or debates on alternate conceptions of validity or its philosophical foundations (see for examples, Borsboom, Mellenbergh & van Heerden, 2004; Cizek, 2012; Lissitz & Samuelson, 2007; Lissitz, 2009; Sireci, 1998).

The reader will also find references to both the terms “evaluation” and “assessment”. Here, we use “educational assessment” broadly to refer to a range of instruments employed to tap individuals’ cognitive or non-cognitive traits, skills or abilities in educational contexts, such as, multiple choice tests, essay examinations, performance assessments, surveys or other rating scales (after Pellegrino, Chudowsky, & Glaser, 2001, p. 20). The term “evaluation”, on the other hand, denotes either a formative or summative use of data generated by educational assessments and other sources for monitoring the effectiveness or improving the quality of an entity, such as, a school or school system, program, curriculum, or an institution. Evaluation could also refer to
individual performance appraisals of students, teachers, and educational personnel/leaders (Rossi, Freeman, & Lipsey, 2004; Yarborough et al., 2010).

**Organization and Broad Themes**

Three widespread forms of international test use involve the testing of individuals for selective admissions, placement, certification or merit rewards (Section 1); system-wide testing of students for evaluating and holding teachers, schools and education programs accountable (Section 2); and large scale, international assessment programs to evaluate the progress of nations (Section 3). Acknowledging that there could be various other applications of tests and assessments in education—such as, in teaching and learning, research, clinical diagnoses or guidance applications—we begin the validity conversations around these three selected themes.

Accordingly, the book’s sections are labeled as follows:

*Section 1--Validity, Fairness and Testing of Individuals in High Stakes Decision-making Contexts;*

*Section 2--Validity Issues in Models of Teacher Evaluation and School Accountability;*

and

*Section 3--Validity Issues in International Large-Scale Assessment (ILSA) Programs.*

Organized in another sense, the dialogue in the volume reflects viewpoints of broad groups of stakeholders of educational assessment programs. As “stakeholder” perspectives, the chapters could be seen as representing particular value positions and interests. Treated from this angle, the reader will find interspersed in the book, perspectives of:

(a) those concerned most with the science and technicalities of developing, validating or conducting research on educational tests and testing programs (see for examples, the
contributions of Kane, Feuer, von Davier, Ercikan & Olivieri, Backoff, and Wagemaker in this volume); (b) those interested most in building sound evaluation systems that utilize student tests and other kinds of information as ingredients in organizational decision-making (see the chapters by Visscher, Gitomer, Georges, and Yon in this volume); and (c) those who rely upon and look towards data from educational tests and other sources in applied decision-making or policy settings, and are interested in the futures of their societies, school organizations, and individuals in small or large ways (see the chapters by Plisko, Pallas, Laurie, and Wandall in this volume).

This last perspective (c), includes that of test users at large, with test-takers, educational practitioners, policy-makers, parents, the public and media in the mix. In addition, the volume looks forward to the future of educational assessment as a process and an enterprise (see Edmund Gordon’s chapter in this volume).

**Clarifying the Problem: How Instances of (In)validity Arise**

To set the stage for the contributions that follow and provide a rationale for inviting an international dialogue, I now turn to a few illustrative cases of invalidity—or commonly-found validity oversights— in the three international test use settings mentioned earlier. My goal in this section is to clarify the book’s central concern for the reader, and make connections to specific sections and chapters as I do so.

**SECTION 1. Issues of Validity and Fairness in the Testing of Individuals.** The literature on international uses of tests of English for high stakes decision-making in college admissions contexts, provides examples of some recurring validity challenges in real world
settings. Decisions in this test use context are made at the individual test-taker level, with validity closely tied to perceptions of equity, fairness and bias (see Kane, this volume).

English as a language now has many native and non-native varieties that differ from the Standard English that is used in the U.S. and the U.K today, two regions where large scale English language tests are typically developed (Lowenberg, 2000). One source of systematic error in such assessments arises from fixed scoring keys used for test items that reflect norms of native English speakers to define what is correct with respect to style, form, syntax, or grammar. For example, Lowenberg (2000) provides the following excerpt of a now retired item from a large scale test, where the examinee is asked to identify the error in a sentence provided in American English.

“To obtain a full refund on your purchase, you must return back the merchandise within ten days” (pp. 50-51).

Here, the incorrect phrase is “return back”. Based on the assessment specifications and scoring key, a proficient examinee would be expected to correct it to “return”. The “back” would be considered redundant in American English. However, the author points out, “return back” would be a permissible and correct usage in several non-native English-speaking nations where the particular language test is routinely used, such as, southeast Asian countries like Indonesia, Malaysia, Phillipines, and Singapore. Cultural factors could likewise influence the ease with which ethnically diverse test-taking groups respond to multiple choice test formats predominant in some high stakes testing instruments.

In considering the above instance, we must note that a single test item does not make for a bad or invalid test. However, the above example illustrates that levels of validity and
meaningfulness of scores for particular sub-groups of test-takers, are influenced by the properties of items on a test, both in terms of their content and format. The same items may be entirely defensible for use with one group of test-takers, yet yield less valid or invalid information with others. In any testing program, therefore, new evaluations of validity and item quality become necessary as test use expands to include new or different test-taking populations. Whether the proportion of differentially-performing items is large enough to affect validity of resulting scores in significant ways, becomes a pertinent empirical question in such studies.

In addition to the inter-related issues of content-based validity, item properties and population shifts, shifts in purposes of such tests from college admissions assessments (that are kept strictly secure from prospective test-takers), to more low stakes teaching and learning tools, could also affect the validity of results in test use contexts. For instance, in seemingly innocuous teaching-learning environments, instructors could expose particular groups of test-takers to the test’s content with samples of actual test items. If such practices occur routinely before the instruments are employed in large scale test administrations for high stakes, college admissions decisions, they would inadvertently favor some groups of test-takers over others. Validity is thereby challenged at the high stakes decision-making point. If some groups are denied access to college based on the test’s results, fairness and equity issues also arise from the differential levels of item exposure before testing (see Spolsky, 1995; Chalhoub-Deville & Deville, 2006; Pearlman, 2003).

Section 1 of this volume discusses precisely these types validity issues related to high stakes, individual testing applications, including chapters by Michael Kane (U.S.), Alina von Davier (U.S.), and Kadriye Ercikan (Canada), Sebastien Georges (France), and Edmund Gordon (U.S.). In his lead paper, Kane (this volume) recommends adoption of an argument-based
approach to validation to pre-emptively address issues of validity, fairness and bias in individual assessment applications. Pointing out that decision-makers and measurement specialists tend to “talk past one another”, Kane recommends that relevant stakeholder perspectives (that of the measurement, test-taking, and decision-making communities) be considered together to improve outcomes with respect to validity.

Ercikan and Oliveri (this volume) and von Davier (this volume) respond to Kane’s proposition for argument-based validation by offering new psychometric and technical solutions for addressing issues of demographic diversity in tested populations. Georges (this volume), views the dilemma from both an educational decision-maker’s and psychometrician’s angles, providing France’s testing programs as the backdrop. Gordon (this volume), in contrast to the rest, calls for a shift away from the current reliance on instruments that are “standardized” and involve large numbers of individual test-takers or group-testing efforts. Questioning the utility of historical traditions of test development and psychometrics on which the worldwide testing industry still draws, Gordon pushes us to support student learning through new paradigms of assessment.

SECTION 2. Validity Issues in Models of Teacher Evaluation and School Accountability. Section 2 is concerned with limitations, misinterpretations or misuse of test data from group testing efforts. For an illustrative case, I quote from the description of a Canadian system given in a recent publication: “Each May, students in Grade 3 and 6 in Ontario, Canada are required to sit for large scale assessments in reading, writing, and mathematics. The following October…(for) schools and school districts, the percentage of their students who performed at or above the provincial standard is reported. Although the assessments have very low stakes for students—the results affect neither students’ marks nor their promotion to the next
grade—they have higher stakes for schools, which are ranked in some newspapers by their results, and the Ontario Ministry of Education (provides) additional resources to schools with declining or static results” (Hollingshead and Childs, 2011, p. 36, parentheses added for clarity).

Here, we find a low stakes testing situation for individual test-takers, contrasted with a high stakes and consequential decision-making for schools that relies on the percent of students meeting proficiency standards on the very same educational test. In their study, Hollingshead and Childs (2011) evaluate the influence of small sample size—the numbers of students taking the test--on inferences made and actions taken regarding school quality from aggregated data at the school or upper levels of the education system. The statistic used is the percentage of students lying above the cut score (PAC). The authors evaluated the errors associated with the size of the school (total numbers of students tested) across different years. To verify findings, they also estimated errors at cut-points for randomly sampled data sets of different sizes drawn from the complete 2009 student data set. Their findings showed high standard errors associated with the PACs for schools with small groups of students in a grade, making school-specific inferences error-prone and suspect. The authors concluded, “To ignore group size when making comparisons is to increase the risk of concluding that…smaller schools are improving or declining, when the changes may simply represent random variations” (p. 43). They urged education officials to take heed.

While technically, the estimation of random errors at PACs is a reliability issue (Haertel, 2006), the particular actions taken by governmental officials, and the inferences made regarding school performance from the student assessment data aggregated at the school level, makes this a classic case of stakeholders overlooking or undermining an important validity issue (AERA, APA, & NCME, 1999). Officials are not as interested in making inferences about students’
achievement in mathematics, reading or writing from the test data, with individual students (for whom the tests were originally designed) serving as the units of analysis. Although student test scores are being used, the construct of interest to education officials is the quality of a school; and the relevant unit of analysis for them is the individual school. Descriptive data from student assessments aggregated as PACs were used to make direct causal inferences about the quality of a school and of schooling processes. Without experimental and other controls for background differences of students, such inferences cannot be defended. The PAC aggregates simply describe the achievement status of students at a given point in time. What would be the responsibilities of various stakeholders—test developers, psychometric researchers, evaluation system designers and test users/policy-makers—in such cases?

Section 2 of this volume presents the descriptive and evaluative analyses of Adrie Visscher (The Netherlands), Jakob Wandall (Denmark), Aaron Pallas (U.S., and specifically, New York City), Drew Gitomer (U.S.), and Haniza Yon (Malaysia) on validity issues surrounding the use of student test data and other assessments for evaluating schools and teachers. As the second case illustrates, issues of validity are often complicated by the aggregation of data at classroom and school levels, and use of the same data in complex statistical models and research designs which aim to make cause-and-effect conclusions about whether or not a teacher, school or program is effective.

Visscher’s chapter (this volume) appraises the implementation of The Netherlands’ Achievement-oriented Work (AOW) policy—a formative evaluation model that utilizes student achievement test data, classroom observations, and other performance criteria for Dutch schools and teachers. He appraises the quality and validity of the processes in place with reference to the Joint Committee’s Program Evaluation Standards (Yarborough et al., 2010).
Wandall (this volume) compares societal contexts, values and cultures of educational systems with a special focus on school-based assessment and evaluation practices and data in use in Denmark and Netherlands, contrasted with the U.S. He comments on validity from these European vantage points. Pallas (this volume) and Gitomer (this volume) respond to Visscher by critically considering validity factors and theories of action guiding more high stakes and summatively-oriented school and teacher evaluation systems emphasized by current U.S. policies. Finally, Yon (this volume) describes Malaysia’s emerging evaluation system compared to The Netherlands, raising her own questions and making recommendations for enhancing validity.

SECTION 3. Validity Issues in International Large Scale Assessment (ILSA)

As mentioned at the start, national evaluation programs today are heavily influenced by outcomes and national rankings reported by ILSA programs, the theme of Section 3. Administered under the auspices of the International Association for the Evaluation of International Achievement (IEA) since the 1960s, elementary and middle grade students (mostly grade 8 samples) from a large number of nations have been tested at regular intervals. These tests are best known today under the title of the Trends in International Mathematics and Science Study (TIMSS). A number of other subject areas have also been tested by IEA, including reading literacy (PIRLS) and civic education (CIVED, and ICCS). A more recently instituted international student achievement survey is the PISA, operated by the Organisation for Economic Cooperation and Development (OECD), which assesses 15 year old students in reading literacy, mathematics literacy, science literacy, and general problem solving.

Countries are ranked on the aggregated results of these assessments. In this case, weighted marginal mean estimates are employed as aggregates in reports issued by IEA and
PISA, respectively. Governmental agencies then release statements to the public and media like the following: “On the (first) eighth-grade TIMSS assessment, the U.S. scored somewhat above the international average in science but somewhat below average in mathematics” (excerpted from Koretz, 2008, p.104).

Similar reports of U.S. results on the PISA in 2003 generated great concern among politicians, policy-makers, and the like, because average scores of U.S. students in mathematics were well below those of developed nations like Germany and Australia, but closer to nations such as Latvia and Spain (see Feuer, this volume). The common inference was that low mean scores indicate that the education system in the U.S. is failing to prepare students in mathematics. And that these findings suggest a decline of the nation’s status as a world economic power.

Yet, many assessment and psychometric specialists informed about such programs (see Brennan, 2001; Goldstein, 2004; Koretz, 2008; Mazzeo, Lazer, & Zieky, 2006) caution users about some of the popular interpretations of group score reports and validity concerns thereof. Whether or not by intention, direct causal attributions--i.e., from average scores to conclusions about poor versus good schooling in a given nation-- are encouraged by published reports that rank nations on assessments like the TIMSS and PISA. As in the second case, cause-effect conclusions about the quality of the education system are technically indefensible here, as reports provide descriptive or correlational results.

Further, the ranks given to nations are also technically suspect. Properly done norm-referenced interpretations (ranks) assume that there is a carefully selected and representative comparison group of examinees or schools (AERA, APA, & NCME, 1999). Norming in a formal sense is not conducted with either the TIMSS and PISA assessment development programs. The
fact that different countries participate in different cycles of these studies raises an obvious
problem for cross-time comparisons. Although there are formal design recommendations and
sampling protocols, participation of schools and classrooms tends to be locally determined based
on priorities, politics and resources—especially in developing nations (the Shanghai PISA
sample is an example of this). Further, the PISA assessments are designed as criterion-referenced
assessments, with cut-scores to delineate different levels of student performance.

A plethora of other conditions could threaten the meaningfulness of inferences and
actions tied to the “group scores” or mean estimates of large scale international assessments,
bulleted next (see Mazzeo et al., 2006 for a more technical discussion on “group scores”). Some
of these problems are more easily solved than others (e.g., first two bullets next could be
addressed with modified reporting and training/orientation of users). For example,

- The content coverage and sampling of topics in a subject area assessment and the
degree of match with what children of that age are typically taught in a given
nation can vary greatly from region to region, affecting content-based inferences
and what scores indicate (AERA, APA & NCME, 1999).

- When the original content and test specifications suggest a multi-dimensional
domain that taps more than one type of student skill and ability, the use of more
efficient scaling methods based on Item Response Theory (IRT) models could
generate unidimensional metrics that force the inclusion of only items that
statistically “fit” that dimension. This would tend to undermine what the final test
actually measures, changing content-based meanings in scores and reports.

- The degree to which samples from different nations and regions are based on the
prescribed sampling design, yielding comparable and representative groups,
affects validity of norm-referenced interpretations and descriptive averages (mean estimates)

- The size of the standard errors associated with marginal “group score” estimates (averages reported for groups) affects reliability of national ranks based averages of scores and therefore, affects the overall validity of the proposed interpretation.

High sampling errors will be found for nations or regions where small numbers of students participate in the testing. Also, the rank of a nation may change depending on nations included in the comparison group.

Section 3 offers thoughtful, contextualized, and up-to-date appraisals of Michael J. Feuer (U.S.), Hans Wagemaker (IEA), Eduardo Backoff (Mexico), Valena Plisko (U.S.) and Robert Laurie (Canada). Feuer (this volume) begins by debunking popular myths surrounding the relationship between student test performance and economic productivity of different nations based on ILSA results, underscoring the utility of ILSA programs in other ways within the scope of national development programs. Referring to Messick’s (1989; 1995) notion on consequential validity, Feuer offers a framework for more contextualized validation of assessment instruments of the ILSA programs, addressing policy environments more directly.

Responding more from a research perspective, Wagemaker’s and Backoff’s chapters (this volume) demonstrate with empirical data, specific kinds of validity challenges that researchers could confront when performing ILSA studies with large and diverse cross-national or national samples. Providing a public policy perspective, Plisko and Laurie (this volume) respond by referencing mind sets of policy-makers and the public in the U.S. and Canada, respectively, who prioritize, seek out and/or fund ILSA programs. Laurie’s chapter also applies Feuer’s validation framework to the Canadian context based on the PISA 2000 results. Together, Section 3 authors
provide recommendations on making ILSA program results and reports more relevant in a
globalized world.

**New Directions**

Measurement theorists have long understood that validity depends not only on how well
tests and assessments are designed and validated, but also on how defensibly the resulting
information is put to use in applied and policy settings (Cronbach, 1971; Messick, 1989; Kane,
2006). Test data uses today are more complex and multi-level than before. Interpretations and
decisions made with test scores involve statistical procedures yielding results that are several
steps removed from the original data source and units of analysis. Educational evaluation and
accountability policies entail the use of more than one data source, with tests and test scores
weighted according to societal values and politics.

Further, many international testing and evaluation efforts today operate under
uncompromising and tight timelines with high stakes for particular stakeholders. Such forces add
stress to the assessment design and validation efforts, as well as to the preparation of test users at
large for appropriate and meaningful information use. When the consequences of assessment
results have high stakes for individuals, programs or institutions (Proctor & Silverman, 2011),
validity issues must become a part of the public discourse. In the Afterword, I summarize new
insights and future directions emerging from the validity conversations in the volume. I hope
readers—including stakeholders representing all sides of international assessment programs--
find this search for constructive solutions to be useful, stimulating and informative.¹

¹ A companion series dealing with validity concerns in U.S.-centered test applications may be found in a
forthcoming issue of the *Teachers College Record* (see Baker, in press; Casey, in press; Chatterji, in press; Henig, in
press; Steiner, in press; Shepard, in press; and Welner, in press).
References


*Psychological Review, 111* (4), 1061-1071.


Casey, L. (in press). The will to quantify. *Teachers College Record.*


Henig, J. (in press). The politics of testing when measures “Go Public”. *Teachers College Record*.


Welner, K. G. (in press). Consequential validity and the transformation of tests from measurement tools to policy tools. *Teachers College Record*. 