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



# Understanding validity issues surrounding test-based accountability measures in the US

Nancy Koh
*Assessment and Accreditation at Lynch School of Education, Boston College, Chestnut Hill, Massachusetts, USA*

Vikash Reddy
*Education Policy and Social Analysis, Teachers College, Columbia University, New York, New York, USA, and*

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

## Abstract

**Purpose** – This AERI-NEPC eBrief, the fourth in a series titled "Understanding validity issues around the world", looks closely at issues surrounding the validity of test-based actions in educational accountability and school improvement contexts. The specific discussions here focus testing issues in the US. However, the general principles underlying appropriate and inappropriate test use in school reform and high stakes public accountability settings are applicable in both domestic and international settings. This paper aims to present the issues.

**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 recurring validity issues arise with tests used in school reform and public accountability contexts, because the external tests tend to be employed as policy instruments to drive reforms in schools, with unrealistic timelines and inadequate resources. To reconcile the validity issues with respect to educational assessment and forge a coherent understanding of validity among multiple public users with different agendas, the authors offer several recommendations, such as: adopt an integrated approach to develop content and standards of proficiency that represent a range of cognitive processes; support studies to examine validity of assessments and the effects of decisions taken with assessment data before results are fed into high stakes accountability-related actions that affect teachers, leaders or schools; align standards, curricula, instruction, assessment, and professional development efforts in schools to maximize success; increase capacity-building efforts to help teachers, administrators, policy makers, and other groups of test users learn more about assessments, particularly, about appropriate interpretation and use of assessment data and reports.

**Originality/value** – Baker points out that in response to growing demands of reformers and policy-makers for more frequent and rigorous testing programs in US public education, results from a single test tend to get used to meet a variety of public education needs today (e.g. school accountability, school improvement, teacher evaluation, and measurement of student performance). While this may simply be a way to make things more cost-efficient and reduce the extent of student testing in schools, a consequence is inappropriate test use that threatens validity in practice settings. This policy brief confronts this recurring validity challenge and offers recommendations to address the issues.

**Keywords** Validity, High stakes testing, Educational accountability, Standards-based reforms

**Paper type** Research paper



Quality Assurance in Education
Vol. 22 No. 1, 2014
pp. 42-52
Emerald Group Publishing Limited
0968-4883
DOI 10.1108/QAE-12-2013-0051

## Introduction

Today, results of students' standardized achievement tests are used for a variety of purposes in America's public schools[1], some of which include the following:

- depicting the status of student learning in different grade levels;
- making promotion decisions as individual students move from one level to the next;
- placing students in special education programs;
- counseling for educational or career guidance; or
- teacher, school or program evaluation and accountability.

Of these, the type of test use that is the most talked about, and almost incessantly debated in public forums and media outlets concerns the use of tests and test-based information in "high stakes" school reform and public accountability contexts.

Conversations on test-based accountability issues become heated and repeatedly so. Depending on regional jurisdictions and policies in effect, students' test scores can cause superintendents and school principals to be hired (or fired); teachers to receive merit recognitions (or not); schools to be closed (or completely reorganized and rebuilt); parents to receive vouchers with choices to keep their children in neighborhood public schools (or take them away) (Amrein and Berliner, 2002; Darling-Hammond, 2004; Jones *et al.*, 2003).

Given what we know about how standardized achievement tests are designed, the types of information they can reasonably yield, and the purposes they best serve, which of these actions reflect valid and appropriate uses of tests and test-based information in school reform and public accountability settings? And, what does the word "validity" mean when viewed from the perspectives of different education stakeholders – such as, the measurement researcher and/or test-maker, the high-level education policy-maker, the teacher union leader, or the politically- and policy-minded thinker? In an ideal context, how should "validity" play out in the world of educational testing practice?

This eBrief, the fourth in a series of AERI-NEPC policy briefs focusing on validity, looks closely at issues surrounding the validity of test-based actions in educational accountability and school improvement contexts. The specific discussions here focus testing issues in the US. However, the general principles underlying appropriate and inappropriate test use in accountability and school improvement settings are applicable in both domestic and international settings.

### *Who and what this eBrief speaks to*

Several distinct audiences might benefit from reading this eBrief. Applied measurement and evaluation specialists or researchers with similar interests could find answers to questions like these:

- How should we go about designing and operating better test-based evaluation systems for schools, school-based programs, and teachers that can meaningfully support decision-making needs of school system stakeholders?
- How can our testing and evaluation system designs be more responsive to values of policy-makers, who fund reforms and emphasize public and tax-payer accountability?

- What could we do better to improve validity during test and evaluation information use in the "real world" of decision-making – particularly, in high stakes school accountability contexts?

Decision-makers, educational leaders, teachers/educators, media, and public stakeholders at large will find information on issues like:

- If test-based information (with or without other kinds of data) is used in high stakes school or teacher evaluation systems for accountability purposes, what are the appropriate uses?
- What are the limitations of such reports?

*Method*

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. Specifically, the content is derived from the keynote presentation and lead article by Eva Baker (2013), reactions by an invited panel of discussants (Casey, 2013; Henig, 2013; Steiner, 2013; Welner, 2013), a subsequent comment offered by Lorrie Shepard (2013), and audience discussions that followed at a March, 2012 conference, titled: "Educational assessment, accountability and equity: conversations on validity around the world". To that synthesis, the authors add their own analysis of key issues. They conclude by offering recommendations for test developers and test users.

We start by summarizing the main points of Eva Baker's keynote address on the concept of validity as it has evolved, showing how it operates in academic, policy, and real world contexts of public schooling today. A summary of the panelists' reactions follow, reflecting different "stakeholder" voices from public education. We end by highlighting audience queries and concerns, with our own thoughts and recommendations on the future of validity in the context of "high stakes" reforms and test-based accountability policies.

## A summary of main themes

*Eva Baker's main ideas*

To start, Baker (2013) notes that historically, the term "validity" has been used differently and has evolved over time in the educational measurement field. The term has different connotations outside the measurement arena.

Validity scholars, test developers and psychometricians have had their own definitions, which continue to change as understandings expand. Teachers, educational administrators, parents, test makers, politicians, policy-makers and other public users, embrace other definitions. These differences can be attributed to a wide range of formal and distinctive understandings of the term that exist today in dictionaries, academic and public spheres.

For example, since 1999, assessment experts, test developers and associated researchers have come to accept a "unified" definition of "validity" with regard to educational test development practice. Assuming tests are developed to produce information for a particular purpose, this definition holds that arguments for claiming validity should be well grounded in fact (evidence), reason, or both and evaluated as a whole with that purpose in mind. Put another way, validity refers to the usefulness of

information that a test provides for decisions that need to be made (AERA, APA, and NCME, 1999; Baker, 2013; Kane, 2013; Messick, 1989; Shepard, 2013).

Regardless, in a given test application context, a policymaker may think about validity much more narrowly and quite differently. An administrator of a special education program concerned about avoiding discriminatory actions against particular pupils based on test results may embrace a legal definition of the term – that is, a definition based on what is a legally acceptable action for different social groups based on their performance on a standardized test. Their guiding question might look something like this: would it be legally acceptable and non-discriminatory to place a particular child in a special education program based on the test scores?

What could be the repercussions of such widely disparate interpretations of the terms, "valid" or "validity"? Baker (2013) suggests that perspectives that conflict in this manner may cause confusions, and complicate matters when it comes to assuring validity in practical settings. To foster appropriate uses of standardized test information in the "real world" of schools where test information is put to use on a daily basis, it would be better to have some agreement on what the term validity means.

Baker (2013) acknowledges that validity scholars themselves have changed their views on the meaning of validity over time. Definitions have changed from thinking of discrete types of validity towards the more unified view referenced earlier.

For example, in the past, it was held that a valid assessment was one that had observable properties intended to measure something (e.g. valid tests of elementary algebra would require students to solve algebraic expressions, and hence look like an algebra test). Baker (2013) identifies this property as "surFACE validity", and as insufficient by itself. In other words, this definition suggests that a test is deemed valid if it "looks valid" on the surface, but it may fail to address how well the test actually measures the desired attribute or construct, and gives limited attention to various other factors that could affect validity defined in a unified sense. For example, we would not know if the algebra test scores would correlate positively with other math tests, as would be reasonable to expect. Or, if the test scores yielded consistent and reliable scores for the same students. All of these kinds of evidence should be put together and evaluated as a whole to decide the extent to which a test is validly tapping some domain.

In the last century, significant developments in measurement and validity theory signaled shifting conceptions of validity in educational measurement. Distinguished scholars like Cronbach and Meehl (1955), Bloom *et al.* (1956), Glaser (1963), Cronbach (1988) and Messick (1989) went on to explain that multiple sources of evidence would be needed to support validity arguments claimed by test-makers.

Messick (1989) also directly addressed the notion of test use as being related to validity. Outside measuring something well, the validity of a test must be evaluated with some purpose in mind. A test that gives valid information for one purpose (e.g. evaluating student achievement in the classroom) may not be equally valid for another purpose (e.g. judging the effectiveness of teaching). There are consequences of appropriate versus inappropriate test interpretation and use. Extrapolating Messick's (1989) ideas on consequences and valid test use, we can conclude that when test-based data are used in invalid or inappropriate ways for making judgments and decisions in the practical world, consequences for those evaluated – test-takers, teachers, schools – could be adverse.

*Assessment purposes and the consequences of test data use in judgments concerning validity*

Baker (2013) acknowledges that discussions on the consequences of test interpretation and use continue to be a contentious topic today. In the academic field of measurement, there is a belief that every testing purpose can be supported with corresponding analytic tools to generate the needed evidence on validity. This, she suggests, may be an expectation that cannot always be met – or is very hard to meet – in test development contexts. She emphasizes that no single test can be made valid for every possible purpose that test users, policy-makers or practitioners desire.

In response to growing demands of policy-makers for more frequent and rigorous testing programs (e.g. No Child Left Behind, 2002; GovTrack.us, 2010), results from a single test tend to get used to meet a variety of public education needs today (e.g. school accountability, school improvement, teacher evaluation, and measurement of student performance). While this may simply be a way to make things more cost-efficient and to reduce the extent of student testing in schools, a consequence of such pressure is inappropriate test use that threatens validity in practice settings.

When test results are used inappropriately regardless of the test's intended functions and purposes, often without regard to a test's inherent or associated limitations of the processed information, we find unintended and negative consequences on students, teachers, schools, and the education system as a whole.

*Baker's views on the 1999 Standards for Educational and Psychological Testing*

Baker (2013) discusses the many challenges and concerns for upholding validity principles as she reflects on her experience as the co-chair of the national committee that assembled the 1999 Standards for Educational and Psychological Testing – the most recent edition available as this eBrief goes to print. Published periodically and jointly by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, the standards provide guidelines addressing best practices with respect to test development and use, along with professional and technical issues in testing for a range of assessment purposes and applications. The standards reflect wide consensus on principles of sound measurement and trends affecting validity in practice settings.

In the 1999 standards, validity was a central theme (Baker, 2013). It was viewed as "a unitary construct" as applicable to particular assessments and results these yielded. As mentioned earlier, the kind of evidence one needs to claim validity are application-specific. It depends on the purpose of the test, the inherent properties of the test and test items, and how well the evidence on the test supports the inferences to be deduced from test results. The 1999 standards also clarify issues on fairness in the use of assessments for policy decisions.

The 1999 standards expected test users to exercise good professional judgment when adopting standardized tests or testing programs for their ends. Users of tests and testing programs apparently sought more prescriptive guidelines. The less prescriptive the approach, the more criticism the 1999 standards committee faced. Baker (2013) describes openly the challenging debates that eventually led to the 1999 standards, which are now undergoing a partial revision.

*Baker's views on standards-based assessment*

Standards-based assessment reforms in the US are ongoing efforts to delineate and gauge what students know and are able to do as they progress through school. Federal, state, and local education policies have engaged in comprehensive reform efforts to improve schools and schooling processes since the 1990s (Borman *et al.*, 2003; McLaughlin and Shepard, 1995; Porter *et al.*, 2011; Swanson and Stevenson, 2002). The Common Core State Standards (National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010) movement is the most recent curriculum reform effort in US public education.

Baker (2013) delineates a set of validity criteria developed by herself and colleagues (Baker *et al.*, 1993; Linn *et al.*, 1991) to guide standard-based assessment design in education that continues to have relevance today. They are: consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost. Baker (2013) endorses standardizing assessments as a means for ensuring equity, or giving every test taker the same conditions to take tests and assessments in education.

However, Baker (2013) also points out that the quality of standards-based assessments employed in schools today are often poor approximations of what is valued by educators/reformers and should be taught ideally. She discusses the many practical and technical challenges to validating standards-based assessments, concluding that the aspirations of the 1999 standards are yet to be fully realized.

In teaching and learning contexts, a continuing validity challenge is in aligning the curriculum standards developed by school reformers, with instruction and assessment in the classroom. Schools and teachers who are held accountable with high stakes sanctions, tend to teach to the external tests to raise student scores, rather than teaching to the standards. The narrowing of standards and content to what is (or can be included) on standards-based assessments weakens the curriculum that is eventually delivered. All this undermines the very point of education reforms, which were intended to support teaching and learning goals and improve overall schooling in the first place.

In sum, barriers to overall education reforms compromise validity claims associated with standards-based assessments. Some barriers to validating and bettering standards-based assessments include:

- insufficient time and a lack of resources (i.e. money) to collect evidence of validity and map changes in student learning;
- inadequate ways to judge the quality of teaching and learning on a larger scale; and
- constraints regarding testing time and resources that limit the number of assessment tasks needed to capture complex kinds of student knowledge and learning.

*Baker on the future of validity*

In examining how validity could play out in the future, however, Baker (2013) ends with optimism. She challenges us to move towards novel forms of testing and assessment in education supported by advances in technology, while being responsive to changes in globalization, demography, societal expectations, and individual

preferences. She imagines a world where new forms of testing are compatible with students using the internet and social media sites in and outside of the classroom. To manage the myriad of possible changes in the future, Baker (2013) recommends that educational assessments focus on students' learning cognitive, social, and intrapersonal skills needed for success in a dynamic global society.

*Reactions from panelists: main ideas*

Leo Casey, Jeffrey Henig, David Steiner and Kevin Welner offer reactions to Eva Baker's thinking on these matters. Lorrie Shepard (2013) provided an afterword to the published proceedings subsequently; her views are also incorporated. Nick Lemann's comments – author of a history of how the use of the SAT® evolved in America and its repercussions on society, and who also participated in the panel – are featured with audience discussions.

In general, the panelists dwelt on the importance of considering the intended purposes of the assessment tools that are employed in policy-making decisions. The consequences of repurposing student assessments for teacher evaluations when those assessments were designed to measure something else, was a central issue of concern.

Representing the voice of teachers and teacher unions, Casey (2013) denounced the recent publication of Teacher Data Reports by the New York City Department of Education, ranking teachers' performances based on their students' test scores, publicly. The city used "value added" measurements based on students' test scores. He attributed the use of standardized test results for such high stakes teacher evaluations to the prevailing "market model of education reform" and accountability, which ignores validity and reliability issues surrounding the tests and reports.

As a political scientist, Henig (2013) observed that while validity issues tend to be discussed in highly technical terms in the educational measurement world, when tests and measures enter broader public spheres, the political realities call for reliance on broader kinds of expertise to ensure appropriate score interpretation and use.

Responding to the recent outcry in New York City against the use of value added measures to evaluate teacher performance, Steiner (2013) draws on his recent experience as the New York state education commissioner to react to Baker and other panelists. He points out the urgent dilemma faced by education policymakers nationwide: the real absence of valid and appropriate tools to identify and reward the most effective teachers, and to make sure that students are not taught by ineffective teachers. He comments on the observed preoccupation with "objective measurements" in US schools, attributing it to a malaise of societal distrust in general. He concludes by endorsing the need for accountability as a means to uphold democratic education principles.

Welner (2013) delves deeper into the idea of consequential validity when student tests are used as accountability measures in school and teacher evaluation contexts. He raises the question: When an assessment is used as a policy tool instead of a measurement tool, can it really accomplish what it was originally intended to accomplish? Regarding test score usage, Welner cautioned against investing too much importance on such results when teacher differences account for less than 20 percent of student variance on those very scores. If the test becomes the sole indicator of teacher performance, he says, teachers' sole objective will be to raise test scores, regardless of how well the tests capture student learning.

Shepard (2013) acknowledges that tests are used as tools to push educational reform policies. However, validation of tests and test-based results in such settings should be guided by the underlying claims and assumptions of the education reform efforts. She views the use of value-added methods as a rather ambitious application of a "useful statistical tool" at present, calling for more validity studies to warrant their use with test-based data.

## Stakeholder views

*Audience views and questions*

Audience members at the conference acknowledged the imperfections of the current tests and testing programs, but they expressed somewhat different concerns. A former Deputy Chancellor in New York City asked why, given the high stakes and grave consequences of the testing outcomes for children's futures, public school children and parents should not have a right to accountability-related information like the Teacher Data Reports in the city? Casey responded by questioning the purpose of the accountability system, contrasting a fear-based system of accountability and a trust-based alternative.

In response to the question about testing accommodations with the SAT® for particular social groups who typically perform poorly on the test, Nick Lemann invoked the theme of the test's original purpose. Such tests were not intended to create winners and losers, he said, especially when the magnitude of the win or loss is so great for the students' futures. If the test's original purpose was upheld in practice, the question of accommodations would not be nearly as consequential. Baker acknowledged the need to employ assessments that are cognitively appropriate, and discussed the need for less secrecy in test-based reports and testing practices.

## Conclusions

*Our thoughts*

How can we reconcile validity issues with respect to educational assessment and forge a coherent understanding of validity among multiple public users with different agendas? The Common Core State Standards (National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010) movement, a national education initiative to align diverse state standards with instruction and assessment, is one step in the right direction. Our comments on validity pertain to this policy direction in the US.

In states implementing the reforms supported by large federal grants, the Common Core State Standards and accompanying assessments have sparked a fierce national outcry against testing in 2013. Educators, parents and local officials reasonably fear that, yet again, tests are serving as blunt policy instruments to drive top-down reforms with inadequate time and resources for designing deeper curriculum and assessments to match, with little or no professional development of teachers and school leaders, and in neglect of critical supports that schools need to succeed.

The common core tests have been criticized as too long, superficial or overly narrow, and out of alignment with curriculum and standards. Yet were the same tests implemented after curriculum standards were refined, teachers and schools readied, parents and students oriented, tests validated to measure what students

actually learned better, and results freed from external rewards and sanctions, the backlash might well disappear. With ill-prepared schools, what do the test results really mean?

Sound educational testing and assessment are integral to good teaching and learning in classrooms and necessary for evaluating school performance. Improving validity would begin with understanding that achievement tests yield meaningful information on student learning (or the quality of schooling) only when test-based information is used appropriately. How, when, and where a test's results are applied, and the defensibility of inferences drawn or actions taken, affect the levels of validity that we can claim from test scores and test-based reports

*Recommendations*

History suggests that school reform and improvement efforts typically wane over time and well-grounded solutions can be elusive. Despite this deterrent, we offer the following recommendations to help increase validity with evidence and support for appropriate decision-making with regard to the common core efforts:

(1) Adopt an integrated approach to develop content and standards of proficiency that represent a range of cognitive processes.

(2) Support studies on validity of assessments and the effects of decisions taken with assessment data before results are fed into high stakes accountability-related actions that affect teachers, leaders or schools.

(3) Align standards, curricula, instruction, assessment, and professional development efforts in schools to maximize success.

(4) Increase capacity-building efforts to help teachers, administrators, policy makers, and other groups of test users to learn more about assessments, particularly, about the appropriate interpretation and use of assessment data and reports.

(5) To improve understandings of how validity concepts and principles play out in practice, expand school-based research studies on teaching and learning in conjunction with assessment.

(6) Create longitudinal databases and defensible systems to monitor student growth.

(7) Develop research-based guidelines for assessment design and item-writing for teachers and non-technical stakeholders so that they can become engaged in building better assessment systems.

**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. Please see the references section for details on a Special Issue of the *Teachers College Record* (Vol. 115, No. 9) and individual articles.

## 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.

Amrein, A. and Berliner, D. (2002), "High-stakes testing & student learning", *Education Policy Analysis Archives*, Vol. 10, p. 18.

Baker, E.L. (2013), "The chimera of validity", *Teachers College Record*, Vol. 115 No. 9, available at: www.tcrecord.org (accessed 8 September 2013).

Baker, E.L., O'Neil, H.F. Jr and Linn, R.L. (1993), "Policy and validity prospects for performance-based assessment", *American Psychologist*, Vol. 48 No. 12, pp. 1210-1218.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R. (1956), *Taxonomy of Education Objectives: The Classification of Education Goals. Handbook I: Cognitive Domain*, David McKay, New York, NY.

Borman, G.D., Hewes, G.M., Overman, L.T. and Brown, S. (2003), "Comprehensive school reform and achievement: a meta-analysis", *Review of Educational Research*, Vol. 73 No. 2, pp. 125-230.

Casey, L. (2013), "The will to quantify: the 'bottom line' in the market model of education reform", *Teachers College Record*, Vol. 115 No. 9, available at: www.tcrecord.org (accessed 8 September 2013).

Cronbach, L.J. (1988), "Five perspectives on validity argument", in Wainer, H. and Braun, H.I. (Eds), *Test Validity*, Erlbaum, Hillsdale, NJ, pp. 3-17.

Cronbach, L.J. and Meehl, P.E. (1955), "Construct validity in psychological tests", *Psychological Bulletin*, Vol. 52, pp. 281-302.

Darling-Hammond, L. (2004), "Standards, accountability, and school reform", *Teachers College Record*, Vol. 106 No. 6, pp. 1047-1085.

Glaser, R. (1963), "Instructional technology and the measurement of learning outcomes: some questions", *American Psychologist*, Vol. 18, pp. 519-521.

GovTrack.us (2010), "H.R. 6244–111th Congress: Race to the Top Act of 2010", available at: www.govtrack.us/congress/bills/111/hr6244 (accessed 27 September 2013).

Henig, R.J. (2013), "The politics of testing when measures 'go public'", *Teachers College Record*, Vol. 115 No. 9, available at: www.tcrecord.org (accessed 8 September 2013).

Jones, G., Jones, B. and Hargrove, T. (2003), *The Unintended Consequences of High Stakes Testing*, Rowman & Littlefield, Lanham, MD.

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.

Linn, R.L., Baker, E.L. and Dunbar, S.B. (1991), "Complex, performance-based assessment: expectations and validation criteria", *Educational Researcher*, Vol. 20 No. 8, pp. 15-21.

McLaughlin, M. and Shepard, L.A. (1995), *Improving education through standards-based reform: A report by the National Academy of Education Panel on Standards-Based Education Reform*, National Academy of Education, Stanford, CA.

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

National Governors Association Center for Best Practices and Council of Chief State School Officers (2010), *Common Core State Standards*, National Governors Association Center for Best Practices and Council of Chief State School Officers, Washington, DC.

No Child Left Behind (NCLB) Act of 2001 (2002), Pub. L.(107-110, § 115, Stat. 1425 (2002).

Porter, A., McMaken, J., Hwang, J. and Yang, R. (2011), "Common core standards: the new US intended curriculum", *Educational Researcher*, Vol. 40 No. 3, pp. 103-116.

Shepard, L.A. (2013), "Validity for what purpose? An afterword", *Teachers College Record*, Vol. 115 No. 9, pp. 1-12.

Steiner, D.M. (2013), "Trusting our judgment: measurement and accountability for educational outcomes", *Teachers College Record*, Vol. 115 No. 9, available at: www.tcrecord.org (accessed 8 September 2013).

Swanson, C.B. and Stevenson, D.L. (2002), "Standards-based reform in practice: evidence on state policy and classroom instruction from the NAEP state assessments", *Educational Evaluation and Policy Analysis*, Vol. 24 No. 1, pp. 1-27.

Welner, K.G. (2013), "Consequential validity and the transformation of tests from measurement tools to policy tools", *Teachers College Record*, Vol. 115 No. 9, available at: www.tcrecord.org (accessed 8 September 2013).

## About the authors

Nancy Koh is the Director of Assessment and Accreditation at the Lynch School of Education at Boston College. Her research interests include diagnostic classroom assessment, evaluation methodology, and education policy reform. She received her doctorate from Columbia University in 2012. Nancy Koh is the corresponding author and can be contacted at: nkoh888@gmail.com

Vikash Reddy is currently pursuing his PhD in Education Policy at Teachers College, Columbia University. After graduating from Dartmouth College in 2005 with an AB in Government, Vikash taught third grade in Brooklyn as a Teach for America corps member and alumnus. In addition to his dissertation research at Teachers College, Vikash is a Senior Research Assistant at the Community College Research Center, and a teaching assistant to former Mayor David Dinkins at Columbia's School of International and Public Affairs.

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.