Educ Asse Eval Acc
DOI 10.1007/s11092-012-9146-6

Working with error and uncertainty to increase measurement validity

Audrey Amrein-Beardsley · Joshua H. Barnett

Received: 14 January 2012 / Accepted: 19 March 2012

Abstract Over the previous two decades, the era of accountability has amplified efforts to measure educational effectiveness more than Edward Thorndike, the father of educational measurement, likely would have imagined. Expressly, the measurement structure for evaluating educational effectiveness continues to rely increasingly on one sole indicator – performance on state-endorsed, large-scaled tests – and the use of these test scores in isolation of other indicators that also capture what it means to be effective. This manuscript describes unresolved questions in educational measurement and provides recommendations to increase measurement validity along both quantitative and qualitative dimensions to move towards a more holistic and appropriate system of measuring educational effectiveness. As Thorndike would have put it, we shall be aided, not hindered, by these tests.

Keywords Validity/reliability · Assessment · Accountability, research methodology · Policy analysis

1 Introduction

Measurement is often operationally defined as the numeric representation for how much or how many widgets there are. Measurements come in multiple forms of scale and worth: weight and mass, length and width, electricity, resolution, and time. And all are interpreted in reference to some base unit of magnitude: kilos and scruples, feet and inches, watts, pixels, and seconds. Measurements can be interpreted in absolute terms, whereby base units are used to reveal fixed values. And they can be interpreted in relativistic terms, whereby value is measured in comparison to another observation of the same discernible reality.

A. Amrein-Beardsley (✉) · J. H. Barnett
Arizona State University, Phoenix, AZ, USA
e-mail: audrey.beardsley@asu.edu

Published online: 05 April 2012

But even though measurement is scientifically based and rooted, measurements function much more like language – they are essentially arbitrary designations that have no inherent value. Rather their values are constructed given there is mutual agreement about how to use and interpret the numbers derived. For example, to measure in terms of miles or kilometers is not a naturally occurring event. Rather it is a mutually negotiated technique that facilitates conversations from which actions and decisions can be made. Because all measurements are ultimately approximations, their accuracy depending on how far they deviate from their true values, much effort must be exerted to ensure measurements are as exact and valid as possible, especially when measurement outcomes are to be used in consequential ways.

Unfortunately, or fortunately depending on where one ontologically stands, for most social science researchers the accuracy of measurements is confounded by the interactions of people involved. As Stone (1997) noted, "measuring social phenomena differs from measuring physical objects because people, unlike rocks, respond to being measured. Try measuring people's heights and watch how they stretch their spines upward…Measurement provokes people to 'play the role' and to present themselves as they want to be seen…(p. 177–178)." Werner Heisenberg took this one step further, applying even this logic within the "harder sciences" (Berliner 2002) of physics and quantum mechanics. Via his Uncertainty Principle, he evidenced that even when measuring physical properties, an uncertain measurement is likely to be taken even when atomic and subatomic particles are examined. Measuring one consistently distorts the measurement of the other and vice versa, and neither can be concurrently measured with certainty (Stanford Encyclopedia of Philosophy 2006).

Yet, in the ever-growing age of accountability in education, measurement is being promulgated as the panacea to the perpetual problems with our current education system. If we could only measure school and teacher effectiveness better and at a more granular level, then we could effectively use data to "fix" the system gone wrong. Such reasoning is fundamentally flawed, however, because measuring, for example, a teacher's effectiveness or ability is infinitely more complex than measuring the weight or length of an object, not to mention the position or momentum of an electron.

On standardized tests, the instruments too often utilized to measure educational quality, teacher effectiveness, and the like, error comes in two forms: random and systematic. Random errors, often considered as measurement noise, are caused by factors that impact test scores inconsistently and by chance. A student who experiences high levels of test anxiety may artificially deflate the student's observed score, for example, or a student who is highly efficacious may post an observed score artificially higher than what the student's true value probably was. Systematic errors, often considered as measurement bias, are caused by factors that impact test scores in more methodical and organized ways. Students whose teachers tell them that their test scores do not matter might post lower than true scores because they care about their performance relatively less or students whose teachers allow them more time to complete tests than allotted, provide hints along the way, or prepare them for tests in unprofessional ways (e.g., teaching to the test) may post artificially inflated scores. This is more likely when high-stakes consequences are attached to tests and distort test score outcomes (Carey 2006; Haladyna et al. 1991; Haney 2000; Koretz et al. 2001; Nichols and Berliner 2007).

Regardless, when random and systematic errors are summed with students' test scores, the resultant units or numerical values are designed to help approximate

students' *true* understandings of tested constructs or concepts. And the belief held firm, particularly by educational policymakers, is that as such standardized tests serve as the strongest and most exact measures of what students know and are able to do. This belief is so strong that some argue that test scores can be used in isolation of other indicators to make accurate statements about school and teacher effectiveness (Au 2011; Baker et al. 2010; Capitol Hill Briefing 2011; Dillon 2010; Felch et al. 2010; Ho et al. 2009; Mellon 2010; Papay 2010; SAS 2011). Proponents contend that as thermometers measure temperature, test scores measure knowledge. Then, once aggregated, test scores can accurately measure school and teacher quality. To some, measuring educational outputs is as simple and straightforward as that.

Others, however, contend that measuring a student's ability and the influence of the school and teacher is much more complicated and demands the use of multiple measures (AERA, APA, and NCME 1999; Capitol Hill Briefing 2011; Hursh 2008; Nichols and Berliner 2007). Opponents note that for test score data to be used in consequential ways, more effort must be devoted to acknowledging the measurement errors inherent within these tests. As well, more effort must be dedicated to considering the consequences caused by false negatives – misclassifications of passing students as failing or effective teachers as ineffectual – and false positives – misclassifications of the inverse. Must effort must also be dedicated towards examining the unintended consequences of such testing systems as a whole. Yet while the debate about the utility of test scores – whether used in isolation or in collaboration – continues across academic circles, policymakers continue to be promoted for their stronger accountability plans and practitioners continue to be victimized as they are consistently forced to increase test scores or face high-stakes consequences.

As such, those involved within the academic debate need to do more to inform this discussion, and they need to engage and work to develop more holistic systems that account for the weaknesses within the educational measurement systems being used across the country. What we know is that objectifying and reducing schools and teachers down to one impartial, measurably imperfect quantity, prizes and punishments attached, is not advancing us far enough away from the classic positivistic stance too many continue to take to capture complex phenomenon (Popper 1968). Where we need to move is to an era of enlightenment, so to speak, in educational measurement by which we critically question tradition and custom, we rely more on reason to construct meaning and truth, we acknowledge the complexities of human interactions in schools, we further understand and educate others about the inferences that can and cannot be drawn from standardized test scores, and we seriously consider whether tests must continue be at the center of our universe of educational accountability.

2 What might a more holistic measurement system include

There is widespread recognition that measuring student achievement more holistically would certainly be more beneficial (Capitol Hill Briefing 2011; Hursh 2008; Jones 2004; Nichols and Berliner 2007) and in line with the standards of the profession (AERA, APA, and NCME 1999). However, such a system is often rejected because it would be more difficult and costly. But given the consequences written into the stronger accountability tenet of No Child Left Behind (NCLB),

consequences that have endured its reauthorization, by not acknowledging these issues we are doing a disservice. By not acknowledging measurement fallacies, measurers' fallibilities, and by not working with measurement errors and the uncertainty with which we interpret standardized test scores, we are contributing to the victimization of schools and teachers alike. The over-reliance on standardized tests as sole indicators of school and teacher effectiveness is causing a stranglehold, limiting our greater expectations for schools and distracting us from the things that matter more – student development and social improvement. The current situation is especially troublesome given we have nearly a decade of evidence about NCLB's negligible intended, but profound unintended effects (Hursh 2008; Nichols and Berliner 2007; PACE 2006; Ravitch 2007).

Instead of such a narrow approach, using a set of measurements or observations collectively and holistically would reduce error and uncertainty, and ultimately yield a more valid measurement of effectiveness. Information theory scientists, for example, use various methods and multiple instruments to capture phenomenon as best they can, acknowledging all the while that with each measurement comes error and ambiguity. But with each additional measurement comes an increased level of validity; that is, if the measurement adds to the accuracy of the estimate by corresponding with other measurements included. Correlating numbers with more qualitative indicators, that to some degree can be quantified based on some set of negotiated and constructed understandings, might in fact be used to more holistically, and more accurately capture these phenomenon.

Consider Merrow's (2001) comparison to a person's health, where he asks: "Do you know your blood pressure? Your cholesterol level? Many people may have a rough idea of those numbers, but odds are that few know the error range of those measurements…for both figures the error range can be 20 or 30 points. Surely no good doctor would prescribe medication for 'high blood pressure' based on one reading, or multiple readings for that matter. No heart surgeon would operate based simply on [one's] dangerously high cholesterol level. More and different tests and procedures would be called for, because the goal is improved health" (p. 653). We concur with this rationale and support Merrow's (2001) claim that "we [as a discipline] need to develop multiple measures, not provide multiple opportunities to be measured the same way" (p. 658).

In short, our most robust numerical data, student test scores, might increasingly be used *correlatively* (see also Koretz et al. 2001) given the extent to which test scores correspond or correlate with other quantitative and qualitative indicators. Then, once all of the data are examined and combined, with co-relationships noted, they could be used to more appropriately and validly capture the construct of effectiveness and quality. Yet, to accomplish this task, we need to expand the quantitative and qualitative measurements collected and combined to understand the complexity inherent within these constructs.

3 Quantitative considerations and additional indicators

First, the most straightforward and perhaps simplest change to the current measurement system would be to administer the state's large-scaled tests twice versus once

per year, first at the outset of the school year as a pretest and then again before the spring semester's close as a posttest. This action would in itself "add value" to the ways we are currently measuring school and teacher value-added, progress, or growth over time. That is, by measuring growth in achievement over one academic year with school and teacher effects controlled, versus measuring growth from one academic year (often spring) to the next (often spring) as students migrate from one school or classroom to the next, would eliminate two of the biggest problems with the value-added models currently being designed and hyper-utilized: (1) controlling for one school's/teacher's residual effects on another's the following academic year and (2) variable levels of summer loss (Bracey 2004; Kupermintz 2003; NWEA 2006; McCaffrey et al. 2003). Academic year fall to spring, rather than across year spring to spring testing would also certainly help to alleviate the dismal levels of reliability currently contaminating all existing value-added systems (Baeder 2010; Baker et al. 2010; Koedel and Betts 2007; Papay 2010).

Second, if all students, including general, English Language Learner (ELL), and mainstreamed special education students, could be randomly assigned to classrooms, the measurement validity and the confidence with which researchers, policymakers, and practitioners could make consequential decisions using value-added scores would be increased (Corcoran 2010; Harris 2009; Ishii and Rivkin 2009; Nelson 2011; Linn 2008; Rothstein 2009). While the only (outdated) research on this topic suggests that random assignment of students into classrooms does occur as part of some local school and district policies, and this occurs more often than we might think (Monk 1987), such assignment would help to control more deliberately for varying student populations and their impact on value-added scores (Ballou et al. 2004; Kupermintz 2003; McCaffrey et al. 2004; Tekwe et al. 2004).

Now while such a policy decision would indeed reduce measurement error and increase the certainty with which we could make valid inferences using test scores, it would also reduce the choices parents often have when selecting teachers for their child(ren). As well, such a decision would reduce the freedom school personnel often have when placing students into classrooms based on what they perceive might work best for individual students given faculty weaknesses and strengths (Monk 1987). On the flip side, random assignment would negate related malpractices, for example, when highly accomplished teachers get stacked with students who are more difficult to teach, new teachers get classes disproportionately loaded with students who have been retained in grade, more traditional teachers are assigned students whose parents prefer back-to-the basics curricula, and the like.

Third, states might design a series of end-of-course exams, more precisely matching state standards. This decision might be similar to, but expand upon, what Virginia is doing with its Standards of Learning (SOL) tests and what New York is doing with its Regents Exams. That is, it might require the development of more tests, although these tests would be more closely linked to the specific subject areas and content taught, increasing content-related evidence of validity. States might use these, or even a similar set of common district benchmark tests that could be administered more than once per year, to facilitate a repeated measures or even archival time-series measurement design. This action would certainly help teachers use conceptually better and more intermittent assessment data, if released in a timely fashion, in more formative and instructionally relevant ways. And this action would certainly help all better

assess the "value" that a school or teacher "added" to student learning over the course of one class, quarter, semester, or academic year. As long as the metrics would be the same across schools, within or across districts or the state, much more valid assertions could be made about student growth over time.

Fourth, states might require a series of performance assessments and randomly sample culminating papers or projects to evaluate school effectiveness. States might do the same using portfolio assessments in which evidence of student growth might also be required. Advanced Placement and International Baccalaureate exams at the high school level might also be considered for inclusion in newer, more measurably holistic accountability systems. And if our Nation's gold standard test – the National Assessment of Educational Progress (NAEP) – is to be required of all schools as the national assessment linked to the new national standards, these scores might also be used to confirm whether gains on state's criterion-referenced tests are indeed valid.

This effort has been the focus of prior research studies on high-stakes tests, yet researchers have found that student gains on state tests have not substantially matched or correlated with state progress on the NAEP. This mismatch is highly problematic in the sense that both tests are testing the same constructs of knowledge, yet they are not showing concurrent or corresponding gains. This has illustrated that most likely students are not learning generalizable sets of knowledgeable, or knowledge that might transfer from one test for which they have been prepared to another, lower-stakes test (Carey 2006; Haney 2000; Koretz et al. 2001; PACE 2006). Using the NAEP alongside states' standardized test scores in order to validate whether gains on state tests are indeed true is one step in the right direction, as long as attaching consequences to the NAEP does not corrupt this assessment as well.

Fifth, states might also develop more authentic state assessments whereby random samples of students in all schools would participate via a sampling design similar to the one used with the NAEP. Not all students would be required to take the exams to make accurate statements about the effectiveness of the schools in which they are housed. In fact, not all students would be required to answer each of the items included on these tests given, like it is within the NAEP's sophisticated item sampling design (Johnson 1992). Although this action would not likely yield student-level data, it would permit the inclusion of better, more authentic, and higher-level assessment items. And the scoring of the assessments would be worth the costs, given fewer assessments and items would need to be evaluated and scored.

But beyond these five expanded and extended quantitative indicators, the educational measurement community would also be well served by including additional qualitative indicators to determining a students' understanding.

4 Qualitative considerations and additional indicators

First, with regard to evaluating teacher effectiveness more holistically, states might integrate a standardized and validated supervisor evaluation system. For example, the Teacher Advancement Placement system (TAP^{TM}) includes an observation protocol that, while still in its infancy in terms of the research conducted on it, might be a

viable alternative for states seriously working towards developing a more holistic system (Capitol Hill Briefing 2011). While some internal research exists evidencing that TAP improves schools by raising teacher quality (see TAP 2011), a recent report released by Mathematica (2010) found no measurable effects on teacher retention or student test scores in Chicago, two years post implementation. However, study critics argue real effects should not have been expected so soon (Sawchuck 2010). Regardless, because the system has been partially validated, and in itself it incorporates multiple measures, multiple measurers, and more measurements over time, it seems to be a viable alternative to supplement value-added analyses. In addition, this alternative is likely better than using teacher evaluation instruments, generated locally, that for the most part have not undergone validation studies. Another option might be for the National Board for Professional Teaching Standards (NPBTS) to develop a similar set of assessments framed by their Five Core Propositions, which would add another alternative for validation.

Second, as for evaluating administrator effectiveness more holistically, states might do well to incorporate the multi-dimensional assessment where teachers, principals, superintendents, and perhaps even parents or community members are included. For example, the Vanderbilt Assessment of Leadership in Education (VAL-Ed) (Murphy et al. 2007) provides a 360-degree evaluation of school administrators across six key core components and six key processes, the reports of which provide an easy to digest table indicating an administrators areas of strength and weakness along 36 key dimensions (the six processes by six components). As such, administrators are able to identify areas to seek out assistance and improvement far beyond examining a monolithic value-added score.

Third, and further broadening the definition of school and teacher effectiveness, states might also expand the use of school report cards, as also required within NCLB. States might continue to report schools' test scores, in aggregate and disaggregate forms, alongside indicators that help contextualize and better interpret extenuating and mediating circumstances (e.g., proportions of racial minority, ELL, and special education students) (Brennan et al. 2001). In addition, states might expand such information to include other, more holistically descriptive indicators. These might include, as informed by research and linked to increases in student achievement: how many teachers have master's degrees, not necessarily in general education but in the content areas they teach (Grissmer et al. 2000; Hanushek et al. 1999), how many years on average teachers in residence have been teaching (Grissmer et al. 2000; Nye et al. 2004), and how many teachers have earned external awards (e.g., National Board Certification) validating them as accomplished teachers (Cavaluzzo 2004; Goldhaber and Anthony 2004).

Fourth, class sizes and teacher-to-student ratios might also be used as measurable indicators of school effectiveness, as might numerical indicators of school safety (e.g. school violence), school resources, and technology use and innovation (e.g. computer to student ratios, opportunities to work with or integrate technologies, innovative technology programs). Other academic and unique opportunities offered to students within- and after-school might also be reported, although more difficult to measure numerically but qualitatively important as described. That which is measured is that what will count, and because the causes of low performance likely reside outside of the school (Berliner 2006; Coleman et al. 1966; Grissmer et al. 2000), it would be

useful to better understand how, as a part of defining a school's effectiveness, school personnel are attempting to help students better succeed given the circumstances they bring with them to school. Parent and student surveys might also be conducted, and their results used to inform reform and change.

While each of these qualitative indicators present their own challenges for data collection, again, the measurement community would be well served to continue exploring additional pathways all the while recognizing the complexity of capturing the school and teacher quality constructs, and all the while acknowledging the limited abilities of single assessments to measure the same constructs alone.

5 Conclusions

Transitioning to a more holistic view, based on research rather than intuition and the reification of school effectiveness using test scores (Gould 1996), would help educators, policymakers, and parents alike be more equipped to evaluate the effectiveness of schools, and teacher quality in those schools. As stated nearly a half century ago in the definitive handbook on educational evaluation, the more numerous and independent the ways by which worth can be demonstrated, the more plausible and likely the resultant definition, a definition far beyond one based solely on a school's or teachers' composite test or value-added scores (Campbell and Stanley 1963). In addition, expanding the definition of effectiveness by including more data elements is at the center of a more holistic measurement system, one that is necessary to help us focus more on an expanded and research-based definition of what counts.

As Selma Wassermann (2001) noted, our current trend is to "…put our faith in numbers because they remove us from the stressful world of uncertainty into what we believe is the definitive world of truth. Numbers provide us with a sense of security [and accuracy] in an uncertain world" (p. 31). However, "even among experts with the most advanced high–tech tools, measurement is subject to both human and technological error" (p. 35).

We contend that it is time to recognize the ramifications of our measurement-addled frenzy towards accountability and the issues associated with our singular strategy for determining worth. Only then will we be able to measure in essence the value of a teacher or school beyond their influence on a test score, and facilitate a conversation about having measures worth working toward. As Edward Thorndike (1921), the founder of educational measurement, wrote nearly a century ago, "It will be said that learning should be for learning's sake, that too much attention is given already in this country to marks, prizes, degrees and the like, that students work too much for marks rather than for real achievement…Students will work for marks and degrees if we have them. We can have none, or we can have such as are worth working for. Either alternative is reasonable, but the second seems preferable" (p. 378). To fully understand if the marks students are working towards are worth it, the measurement community must recognize the utility and limitations of the current measures in place and the need to move towards a more holistic system of measurement that provides educators with practical, formative, and improved feedback, in real time all of the time.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *AERA position statement on high-stakes testing in pre-K12 education*. Retrieved from http://www.aera.net/?id=378.

Au, W. (2011). Neither fair nor accurate: Research-based reasons why high-stakes tests should not be used to evaluate teachers. *Rethinking Schools*. Retrieved from http://www.rethinkingschools.org/archive/25_02/25_02_au.shtml.

Baeder, J. (2010). Gates' measures of effective teaching study: More value-added madness. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/on_performance/2010/12/gates_measures_of_effective_teaching_study_more_value-added_madness.html.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. The Economic Policy Institute. Retrieved from http://epi.3cdn.net/b9667271ee6c154195_t9m6iij8k.pdf.

Ballou, D., Sanders, W. L., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–66. doi:10.3102/10769986029001037.

Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researcher, 31*(8), 18–20. doi:10.3102/0013189X031008018.

Berliner, D. C. (2006). Our impoverished view of educational research. *Teachers College Record, 108*(6), 949–995. doi:10.1111/j.1467-9620.2006.00682.x.

Bracey, G. W. (2004). Serious questions about the Tennessee Value-Added Assessment System. *Phi Delta Kappan, 85*(9), 716–717.

Brennan, R. T., Wenz-Gross, M., & Siperstein, G. N. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review, 71*(2), 173–216.

Campbell, D.T. & Stanley, J.C. (1963). *Experimental and quasi–experimental designs for research on teaching*. In N. L. Gage (Ed.) Handbook of research on teaching. American Educational Research Association.

Capitol Hill Briefing. (2011). *Getting teacher evaluation right: A challenge for policy makers*. Washington DC: Dirksen Senate Office Building (research in brief). Retrieved from http://www.aera.net/Default.aspx?id=12856.

Carey, K. (2006). *Evidence suggests otherwise. Hot air: How states inflate their educational progress under NCLB*. Washington: Education Sector.

Cavaluzzo, L. (2004). *Do teachers with National Board Certification improve student outcomes?* Alexandria: The CNA Corporation.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity*. Washington: U.S. Department of Health, Education and Welfare.

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice*. Providence, RI: Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/products/Corcoran.php.

Dillon, S. (2010). Formula to grade teachers' skill gains acceptance, and critics: How good is one teacher compared with another? *New York Times*. Retrieved from http://www.nytimes.com/2010/09/01/education/01teacher.html?_r=1&emc=eta1.

Felch, J., Song, J., & Smith, D. (2010). Who's teaching L.A.'s Kids? *Los Angeles Times*. Retrieved from: http://www.latimes.com/news/local/la-me-teachers-value-20100815,0,258862,full.story.

Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Seattle, WA: Center on Reinventing Public Education. Retrieved from http://www.crpe.org/cs/crpe/view/csr_pubs/70

Gould, S. J. (1996). *The mismeasure of man*. New York: Norton.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What NAEP test scores tell us*. Santa Monica: RAND Corporation.

Haladyna, T. H., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*, 2–7. doi:10.2307/1176395.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis Policy Archives, 8*(41). Retrieved from http://epaa.asu.edu/epaa/v8n41/.

Hanushek, E.A., Kain, J.K, & Rivkin, S.G. (1999). *Do higher salaries buy better teachers?* National Bureau of Economic Research: Cambridge, MA. Retrieved from http://www.nber.org/papers/w7082.pdf.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An evaluation of the statistical properties and policy alternatives. *Education Finance and Policy, 4*, 319–350. doi:10.1162/edfp.2009.4.4.319.

Ho, A. D., Lewis, D. M., & Farris, J. L. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*, 15–26.

Hursh, D. W. (2008). *High-stakes testing and the decline of teaching and learning: The real crisis in education*. Lanham: Rowman & Littlefield.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy, 4*, 520–536. doi:10.1162/edfp.2009.4.4.520.

Johnson, E. G. (1992). The design of the national assessment of educational progress. *Journal of Educational Measurement, 29*(2), 95–110. doi:10.1111/j.1745-3984.1992.tb00369.x.

Jones, K. (2004). A balanced school accountability model: An alternative to high-stakes testing. *Phi Delta Kappan, 85*(8), 584–590.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. Working Paper No. 2007–03. Nashville: National Center on Performance Initiatives.

Koretz, D. M., McCaffrey, D.F., & Hamilton, L.S. (2001). *Toward a framework for validating gains under high-stakes conditions.* Center for the Study of Evaluation. (CSE Technical Report #551). Retrieved from www.cse.ucla.edu/products/Reports/TR551.pdf.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation & Policy Analysis, 25*(3), 287–298. doi:10.3102/01623737025003287.

Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies, 40*, 699–711. doi:10.1080/00220270802105729.

Mathematica Policy Research Inc. (2010). *Early results show no impact of Teacher Advancement Program in Chicago: No measurable effect on teacher retention, student test scores in second year of rollout*. Retrieved from http://www.mathematica-mpr.com/Newsroom/Releases/2010/TAP_5_10.asp.

McCaffrey, E. F., Koretz, D., Lockwood, J. R., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica: Rand.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101. doi:10.3102/10769986029001067.

Mellon, E. (2010). HISD moves ahead on dismissal policy: In the past, teachers were rarely let go over poor performance, data show. *The Houston Chronicle*. Retrieved from http://www.chron.com/disp/story.mpl/metropolitan/6816752.html.

Merrow, J. (2001). Undermining standards. *Phi Delta Kappan, 82*(9), 653–659. doi:10.2307/1164374.

Monk, D. H. (1987). Assigning elementary pupils to their teachers. *Elementary School Journal, 88*(2), 166–187.

Murphy, J., Elliott, S. N., Goldring, E., & Porter, A. C. (2007). Leadership for learning: A research-based model and taxonomy of behaviors. *School Leadership and Management, 27*(2), 179–201.

Nelson, F. H. (2011). *A guide for developing growth models for teacher development and evaluation*. Paper presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge: Harvard Education Press.

Northwest Evaluation Association (NWEA). (2006). *Achievement gaps: An examination of differences in student achievement and growth.* Lake Oswego, OR. ED498429

Nye, B., Konstantopoulos, S., & Heges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(4), 237–257. doi:10.3102/01623737026003237.

Papay, J. P. (2010). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*. doi:10.3102/0002831210362589.

Policy Analysis for California Education (PACE) (2006). *Is the no child left behind act working? The reliability of how states track achievement.* University of California, Berkeley. Retrieved from http://pace.berkeley.edu/2006/06/01/is-the-no-child-left-behind-act-working-the-reliability-of-how-states-track-achievement/.

Popper, K. W. (1968). *Conjectures and refutations: The growth of scientific knowledge*. New York: Harper & Row, Publishers, Inc.

Ravitch, D. (2007). Get congress out of the classroom. *New York Times*. Retrieved from http://www.nytimes.com/2007/10/03/opinion/03ravitch.html.

Rothstein, J. (2009). *Student sorting and bias in value-added estimation: Selection on observables and unobservables*. Cambridge, MA: The National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14607.

SAS. (2011). SAS® EVAAS® for K-12: Assess and predict student performance with precision and reliability. Retrieved from http://www.sas.com/govedu/edu/k12/evaas/index.html.

Sawchuck, S. (2010). Performance-pay model shows no achievement edge. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2010/06/01/33tap.h29.html?tkn=SLLFZe8XVYfHJJSsSgGYCI87ZvETbCbN%2FXmT&cmp=clp-edweek.

Stanford Encyclopedia of Philosophy. (2006). *The uncertainty principle*. Retrieved from http://plato.stanford.edu/entries/qt-uncertainty/.

Stone, D. (1997). *Policy paradox: The art of political decision making*. New York: Norton.

TAP. (2011). *The system for teacher and student advancement*. Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved March 12, 2011 from http://www.tapsystem.org/policyresearch/policyresearch.taf.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11–35. doi:10.3102/10769986029001011.

Thorndike, E. (1921). Measurement in education. *Teachers College Record, 22*(5), 371–379. doi:10.1037/11013-001.

Wassermann, S. (2001). Quantum theory, the uncertainty principle, and the alchemy of standardized testing. *Phi Delta Kappan, 83*(1), 28–40.