School Leadership Review

Volume 10 | Issue 1 Article 5

2015

Professional Development for Educational Leaders in the Era of Performance Evaluation Reform

Kelly H. Summers
Northern Illinois University

Todd D. Reeves
Northern Illinois University

David A. Walker
Northern Illinois University

J. Schwartz
Northern Illinois University

Professional Development for Educational Leaders in the Era of Performance Evaluation Reform

Kelly H. Summers[i]
Northern Illinois University

Todd D. Reeves
Northern Illinois University

David A. Walker
Northern Illinois University

J. Schwartz
Northern Illinois University

The word "accountability" has become a mantra in public education. Arguably, this one word, and the movement it has produced, has shaped the direction of our field in the past decade more than any other (Harris, 2011). This movement has led to many positive changes including an examination of gaps in student achievement, the types of assessments used in schools, and the strength of the performance evaluation systems for principals and teachers. Many large urban school districts, as well as entire states, have revamped the way public school principals and teachers are evaluated. In fact many, including the State of Tennessee, Dallas Independent School District, Milwaukee Public Schools, Houston Independent School District, and the State of Illinois, have started or will start using some sort of student achievement metric as part of teacher and/or principal performance evaluations. The ideas surrounding using student growth seem simple enough: If student test scores improve, it means the teacher or principal is doing his or her job well and therefore should be rewarded. This seemingly simple idea is in fact quite complex. Many school administrators may not have the background or training to implement growth models as part of performance evaluations (Mitgang, 2012), which could lead to potentially unethical and incorrect implementation of newer forms of accountability such as growth modeling. Such problems have already arisen in a number of districts across the nation (Harris, 2011).

Training school leaders in the key areas of policy and state laws, basic statistical literacy, assessment types and purposes, and value-added models is particularly important because ideas shaping teacher performance evaluation are being considered at the national, state, and local levels. In fact, we argue that providing school leaders with this background is *necessary* in order to conduct ethical and effective performance evaluations. In this

[i] Dr. **Kelly H. Summers** can be reached at *ksummers@niu.edu*

article, we present several key areas for professional development aimed at providing educational leaders with this background. These key areas provide a framework (i.e., called Ethical and Effective Performance Evaluations or E^2PE) for designing workshops for educational leaders as they refine practices in conducting ethical and effective performance evaluations of both principals and teachers. Specifically, this article could be helpful to university personnel as they consider how to assist local school districts in implementing effective and ethical performance evaluations of principals and teachers. Often times federal and state laws governing performance evaluations change rapidly. Educational leadership preparation programs teach students what is current at the time they are in the program. The E^2PE framework provides an opportunity for university personnel to ensure educational leaders are up-to-date on the latest issues surrounding performance evaluation even if those leaders are not actively enrolled in a formal university certification or degree program. Using the E^2PE framework would allow university personnel to engage in out-reach practices with former students and local educational leaders on a regular basis in order to assist them on staying current with the most up-to-date practices for ethical and effective performance evaluations.

Practitioners will find this article helpful as it will give them ideas for the type of high quality professional development needed to support their effective and ethical performance evaluations. Figure 1 depicts our E^2PE framework for professional development in conducting effective and ethical performance evaluations. The importance of each foundational area is presented along with ideas about how to implement this framework as a professional development workshop. It should be noted that each foundational area could be presented in any order with the exception of statistical literacy and value-added modeling. Because knowledge of value-added modeling relies heavily on statistical literacy, statistical literacy should always be presented prior to value-added modeling. The order of the remaining components that form the foundation of the E^2PE framework could feasibly be presented in the order that works best in the local context.

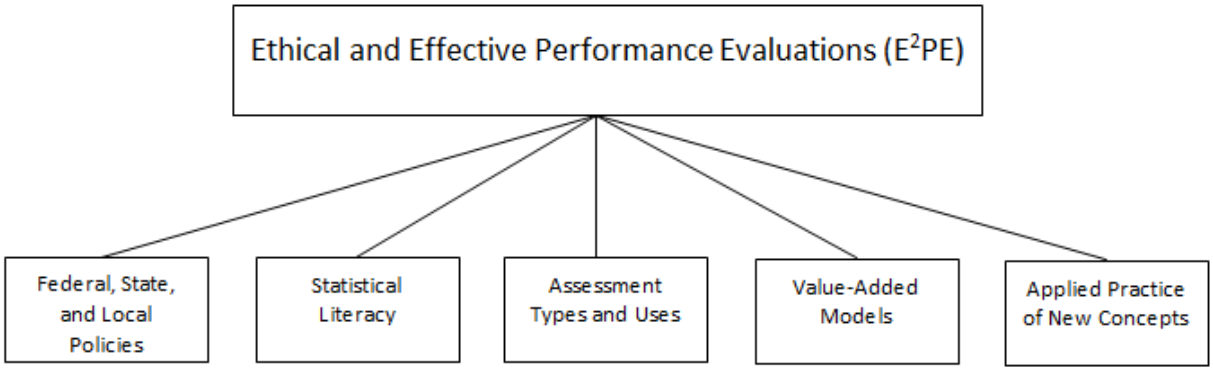

Figure 1. The E^2PE professional development framework.

The Policy Landscape

In order to understand the accountability movement in public education, educational leaders need to have an understanding of the policy landscape that has heavily influenced our current reality in education. Although there are many laws, policies, and court proceedings that have heavily influenced education, *A Nation at Risk*, No Child Left Behind, and Race to the Top are all directly linked to the accountability movement and should be reviewed in any professional development framework on effective and ethical performance evaluations for school administrators.

Many scholars trace the start of the accountability movement to *A Nation at Risk*, the federal report released in 1983, that highlighted the perceived inadequacies of public education (Berliner, 2011; Berliner & Biddle, 1995; Ornstein, 1988). Recommendations from *A Nation at Risk* included more rigorous standards and improved curricula coupled with frequent standardized assessment to ensure standards are being met (Amerin-Beardsley, 2014). Although scholars have since discredited most of the findings from *A Nation at Risk* (Berliner; Berliner & Biddle), the report prompted widespread fear about America's failing schools and set the wheels in motion for decades of trying to perfect accountability. Because this report had such far-reaching implications, it is important to provide a brief overview in the professional development framework.

The reauthorization of Elementary and Secondary Education Act (ESEA), commonly referred to as the No Child Left Behind Act (No Child Left Behind [NCLB], 2002), was the next big step toward increased accountability. NCLB called for 100% of American children to be proficient in math and reading by 2014; thus, introducing the idea of Adequate Yearly Progress for school districts. School districts as a whole were required to show yearly growth and several subgroups (e.g., based on race, income status, disability status, etc.) within those districts were required to show growth. For each year a school or district failed to show growth, sanctions increased accordingly, culminating in complete restructuring in year five of no growth. NCLB brought "formidable pressure to bear on states, school districts, and schools to meet the demands of the law" (Fowler, 2013, p. 320), but this pressure was deceptive as each state was allowed to develop its own set of learning standards and the corresponding tools to measure progress toward those standards. Public embarrassment occurred for districts and schools as a result of NCLB, but no real educational change occurred (Fowler). At present, the nation is waiting for Congress to re-authorize ESEA, but it has been stalled for a number of years.

Race to the Top (RTTT) is another policy that would be important to review as part of a professional development framework for conducting effective and ethical performance evaluations. In 2009 as part of the American Recovery and Reinvestment Act, $4.35 billion was earmarked for RTTT, a competitive grant for states seeking to implement innovative reforms in education. A total of 45 states and Washington, D.C. applied for the grant. Washington D.C. and 18 states were ultimately awarded the grant. Although RTTT is now five years old, several states are still in the process of phasing in the large-

scale changes that resulted from earning the grant. Many of the RTTT states adopted a growth model or value added model as their new system of accountability, which is what makes the development of a workshop on effective and ethical performance evaluations a timely endeavor. It is important for school leaders to understand the historical policy landscape that led us to where we are today, not just for RTTT states, but nationwide.

Overview of State Laws

School leaders should also be familiar with state laws concerning performance evaluation. Furthermore, if larger school districts have refined or developed their own performance evaluation systems, that information should be included in the workshop. Using Illinois as an example, in January of 2010 The Performance Evaluation Reform Act (PERA) was signed into law. Part of PERA requires all public school principals and teachers to be evaluated using student growth measures by the 2016 school year. Although the law is clear about the use of "growth measures" and the implementation date, individual school districts, almost 900 in Illinois, can determine their own method of growth assessment. Some growth models are simple and straight forward whereas others are quite technically sophisticated. Whether state law indicates exactly what each district is to do or gives discretion to each district, a discussion of state law specifics is necessary in a workshop on effective and ethical performance evaluations in order to ensure policy compliance.

In a professional development workshop, policy does not necessarily need to precede law in the manner we have presented here. However, both should be included as part of a professional development workshop on ethical and effective performance evaluations. These issues form the foundation for why we need to address our systems of performance evaluation in the first place. For this reason, both are included in our framework.

Statistical Literacy

The concept of "statistical literacy" has been discussed in the scholarly literature for over 20 years (Gal, 2002; Gal & Garfield, 1997; Wallman, 1993). Frameworks detailing the various levels that comprise statistical literacy have been offered by Watson (1997) and delMas (2002). Ideas of how statistical literacy might be defined have been outlined in work conducted by Gal (2000) and Watson. In spite of the importance placed on statistical literacy in the scholarly community, many teacher and principal training programs do not include specific coursework on statistical literacy in educational settings (Chick & Pierce, 2013). Because basic statistical literacy forms the foundation for understanding assessments and value-added models, statistical literacy should be a central theme in professional development on ethical and effective evaluations. For the workshop, the concept of developing and/or promoting "statistical literacy" for the participants could be focused around Rumsey's (2002, n.p.) components: 1). data awareness; 2). an understanding of certain basic statistical concepts and terminology; 3).

knowledge of the basics of collecting data and generating descriptive statistics; 4). basic interpretation skills; and 5). basic communication skills.

An important aspect of the workshop, and ultimately the framework for providing the workshop, is to assist school leaders in their recurrent exposure to and development of statistical literacy. The workshop's construct of "statistical literacy" is examined and operationalized via descriptive and inferential statistics as well as measurement concepts such as validity, reliability, and bias. A long-term intention is to support all educational leaders in communicating the importance of data awareness and applying various analyses in an attempt to determine, via the use of interpretation skills, if an idea such as "student growth" transpired and how to communicate said results to a broader base of internal and external constituents.

Assessment Purposes and Uses

Within the realm of the assessment of academic achievement (student mastery of content standards), school leaders need to understand that assessment occurs for a variety of purposes and the data are variously used (Pellegrino, Chudowsky, & Glaser, 2001). The professional development framework includes general and specific assessment purposes and uses (e.g., formative assessment, identifying strengths and weaknesses). In this section, we describe and exemplify the two general assessment purposes, formative and summative, and enumerate the ways formative and summative assessment data can be used.

Formative assessment is intended to support teaching and learning and generally occurs before or during some unit of instruction (Pellegrino et al., 2001). For example, diagnostic assessments, a subtype of formative assessment, might shed light on student strengths, weaknesses, errors, or misconceptions. These data can be used to select appropriate content, determine which students to provide extra support to, which instructional method to use, or how to group students for differentiation. Another sub-category of formative assessment is interim/benchmark assessment, which is intended to indicate whether students' are on-track to success on future assessments.

In contrast to formative assessment, summative assessment is intended to describe learning that occurred during some unit of instruction (Pellegrino et al., 2001). State mathematics and English/language arts tests in grade 3-8 mandated under No Child Left Behind are a well-known example. Other examples include any traditional test or performance-based assessment administered by a teacher for the purposes of grading. Outside of the classroom, summative assessment data are often used in part to make high-stakes decisions about students such as grade promotion/retention, graduation, and increasingly, to serve accountability uses such as teacher/school/principal evaluation. While most summative assessments focus on student status (level of student knowledge/skill at a fixed point in time such as the end of the school year); the focus of

assessment is increasingly on student growth (change in the level of student knowledge/skill over some time interval such as over the course of the school year).

Not all assessments defensibly support the making of all educational decisions and tests should be designed with the purpose and intended use in mind (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). School leaders need to have a deep understanding of the appropriate uses of formative and summative assessments so they understand the strengths and limitations of using these types of assessments as part of performance evaluations, as such these discussions should be a key component of any professional development framework on the topic of ethical and effective performance evaluations.

Value-Added Modeling

There are several categories of evidence that might be used within performance evaluation systems to support decisions about educators. These include indicators based on student-achievement (including value-added models, student growth percentiles, and unadjusted gains), teaching practice artifacts (e.g., lesson plans and teacher-developed assessments), observations, and student surveys. Amidst debates over the weight apportioned to different types of evidence (Baker et al., 2010), the focus of this section is one type of student achievement-based indicator of teacher effectiveness, namely value-added models. The use of value-added modeling (VAM) involves the application of statistical methods to student test score data with the aim of isolating the impact of individual teachers on students and thus identifying effective and ineffective teachers. Many states and districts have started to use VAMs as part of the evaluation process. Because they can be somewhat complex to understand, a discussion of VAMs should be included in any professional development for ethical and effective evaluations.

VAM methods compare students' observed achievement test scores to those predicted on the basis of a statistical model. The statistical model allows the estimation of an expectation for each student's level of achievement, given their prior achievement and possibly other variables (e.g., socioeconomic status and race/ethnicity). Each student's actual level of achievement is compared to that predicted on the basis of the model, and the discrepancy between these values (the residual) is taken as evidence for the effect of the teachers. Should a teacher's students typically perform better than they were expected to ("value-added" by the teacher), the inference drawn is that the teacher is more effective than other teachers whose students were represented in the data.

VAM-based indicators are advantageous relative to current achievement indicators such as cohort-to-cohort status comparisons (which depend on both cohort and actual year-to-year changes), and unadjusted gain scores (which depend on where students start). In particular, VAM is an attempt to deal with fairness issues relative to the context in which teachers work. The non-random assignment of students to schools and teachers possibly introduces bias in the comparison of teachers based on simple student achievement

indicators (Easton, 2008; Jennings, 2010; McCaffrey & Lock, 2008). By considering students' prior academic or other characteristics, VAM can set up more realistic achievement expectations for a given teacher's students. Thus, a teacher who serves low-achieving students from disadvantaged backgrounds is not unfairly penalized under the assumption that her students will be expected to achieve at the same level as their higher-achieving and less disadvantaged counterparts.

However, school leaders need to understand that VAM does not--and perhaps cannot--statistically account for all of the factors that are unevenly distributed across schools and teachers (Corcoran, 2010; McCaffrey & Lock, 2008), including things occurring before (e.g., summer learning loss, nutrition) and during (e.g., tutoring, and absenteeism) the school year. VAM models often omit other potentially important classroom-level factors as well (e.g., social composition, degree of behavioral problems) (Corcoran; Jennings, 2010). To the extent that these factors drive achievement and are distributed unevenly across classrooms, they can offer alternative explanations for VAM results one hopes to ascribe to teacher effectiveness. School leaders should understand that expectations for students are only as good as the appropriateness and comprehensiveness of the variables used to estimate them.

Building upon the background in statistical literacy and understanding of assessment types and purposes, school leaders should be presented with information about the consistency of VAMs. For example, there is some evidence that VAM estimates for individual teachers are unreliable (inconsistent) from year-to-year (Linn, 2008). Issues such as small numbers of students in a class, test exemption, mobility, and absenteeism can all contribute to inconsistency (Corcoran, 2010). At the same time, different VAM models that utilize different tests, and include different variables in the statistical model, often yield different results (Corcoran). Moreover, VAM-based estimates hinge on the quality of the test scores analyzed, in terms of reliability and validity (Linn). For example, if the test scores input to VAM fail to represent all important facets of the content standards (American Educational Research Association et al., 2014), inferences based on VAM-based estimates are restricted to only assessed content. These problems can translate to errors in the identification (misidentification) of effective and ineffective teachers. Another important VAM limitation is that it does not provide diagnostic information about teaching that can be used to support teacher development (e.g., evidence that a teacher is weak in a particular area of practice such as assessment). Given these and other issues, and a lack of extant evidence that VAM demonstrably results in improved teaching and learning, the technical and consequential aspects of VAM models are an active area of educational research (Baker et al., 2010; Harris, 2011).

Applied Practice

Once they become more familiar with educational policies, the principles of statistical literacy, assessment types, and value-added models, professional development participants should be given time and assistance during the workshop in applying to

practice ideas such as descriptive statistics, graphing data, inferential statistical models to monitor student growth, and conducting correlations to examine ideas of reliability with summative, interim, and formative assessments. District, school, or classroom level data from workshop participants' own contexts could be used. Advanced statistical programs, while helpful, are not necessary as Excel could be used in all of the practical experiences and exercises (see Figure 2). If participants are not proficient in use of Excel, the applied practice portion of the workshop could engage them in a primer on the use of Excel both in terms of data entry and conducting basic statistical analyses with their local data (see Figure 3). Group follow-up questions pertaining to results, interpretations, and how to communicate findings and areas for development with teachers in conjunction with local evaluation standards should be encouraged. Specifically, presenters could facilitate questions focused on the interpretation and communication of results with colleagues regarding descriptive statistics as well as score reliability via correlational analysis with interim and formative assessments.

	Instructor 1	Instructor 2
	3.1	2.5
	2.7	3.9
	3.2	4
	2.9	2.4
	2.8	2.7
	3	2.8
	3.3	4
	2.8	2.6
	3	3.1
	3.2	3
Mean	3.00	3.10
Median	3.00	2.90
Mode	3.20	4.00
Variance	0.04	0.40
SD	0.20	0.63
Minimum	2.7	2.4
Maximum	3.3	4
Skewness	0	0.65
Kurtosis	-1.23	-1.39
Q_1	2.83	2.63
Q_2	3.00	3.05
Q_3	3.15	3.10
P_{10}	2.79	2.49
P_{20}	2.80	2.68
P_{30}	2.97	2.77

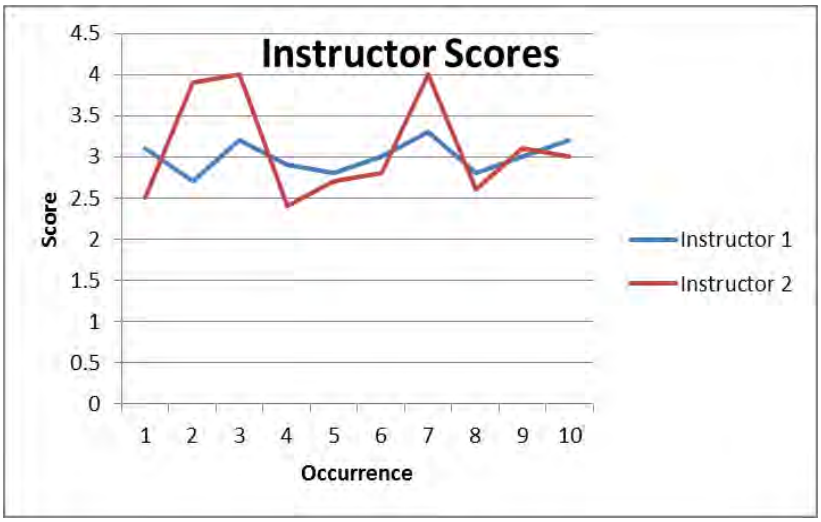

Figure 2. Workshop example of descriptive statistics and graphing use in Excel.

Student	Study	GPA				
1	42	3.3				
2	23	2.9				
			Correlation			
3	31	3.2	**(*r*)**	***t* value**	***p*-value**	
						Note: Two-tailed test with 6
4	35	3.2	**0.884**	**4.524**	**0.004**	df (n-2)
5	16	1.9				
6	26	2.4				
7	39	3.7				
8	19	2.5				

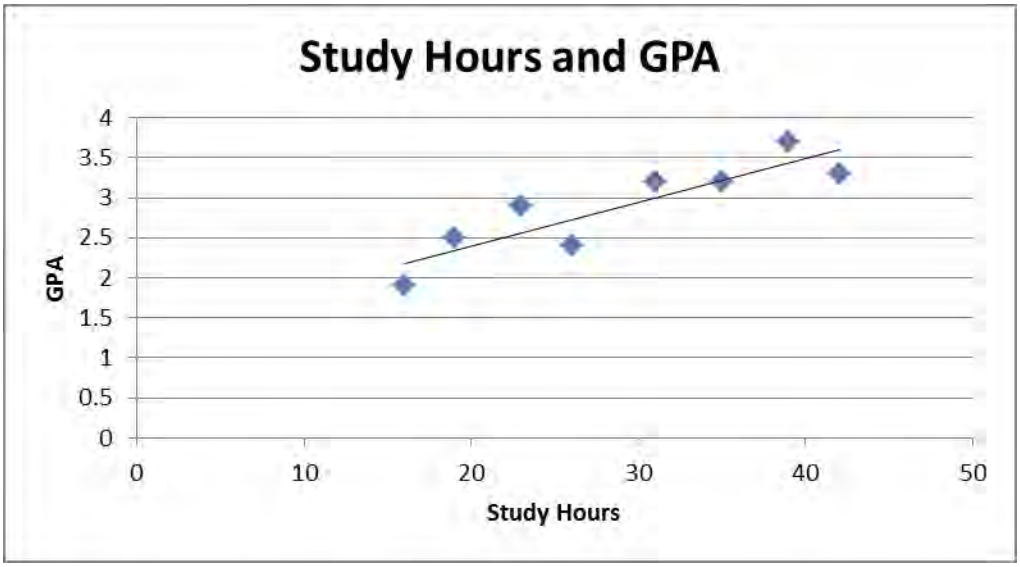

Figure 3. Workshop example of correlation use in Excel.

Concluding Remarks

The model presented, and its components (i.e., national policy and state laws, basic statistical literacy, assessment types and purposes, value-added models, and applied practice), provides a framework for developing workshops we believe will contribute to promoting effective and ethical performance evaluations of teachers. Such a workshop will provide school leaders with the knowledge and tools necessary to refine practices in conducting ethical and effective performance evaluations. Also, a workshop of this type would assist university personnel when working with local school districts in implementing effective and ethical performance evaluations, and assist practitioners in developing ideas for the types of high quality professional development that would support their effective and ethical performance evaluations.

Secondly, the model emphasizes applied, guided practice, an equally important element of any workshop that develops and promotes effective and ethical performance evaluations of teachers. Within our explanation of each component of the model, we provide ideas for developing a workshop, or series of workshops, that will assist participants in applying and practicing the ideas contained in each component such as discussing the historical policy landscape, graphing data, and conducting correlations to examine ideas of score reliability with assessments. We also strongly encourage the applied practice portion of the workshop utilize district, school, or classroom level data relevant to the workshop participants.

Although the specifics of any workshop will vary, the model and components we presented will provide school leaders, university personnel, and practitioners with a framework that will guide them in implementing effective and ethical performance evaluations of teachers and assist in their work towards "accountability" in public education.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability.* New York: Routledge.

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.* Retrieved from http://www.epi.org/page/-/pdf/bp278.pdf

Berliner, D. C. (2011). Rational responses to high-stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education. 41*(3), 287-302.

Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley.

Chick, H., & Pierce (2013). The statistical literacy needed to interpret school assessment data. *Mathematics Teacher Education and Development*, *15* (2). Retrieved from http://files.eric.ed.gov/fulltext/EJ1018712.pdf

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*. Retrieved from www.annenberginstitute.org/WeDo/NYC_ Conversations.php

delMas, R. C. (2002). Statistical literacy, reasoning, and learning: A commentary. *StatisticsEducation*, *10*(3). Retrieved from http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html

Easton, J. Q. (2008). *Goals and aims of value-added modeling: A Chicago perspective.* Retrieved from http://www7.nationalacademies.org/bota/VAM%20Goals%20and%20Aims%20-%20Easton.pdf

Fowler, F. C. (2013). *Policy studies for educational leaders: An introduction* (4th ed.). Boston: Pearson.

Gal, I. (Ed.). (2000). *Adult numeracy development: Theory, research, practice*. Cresskill, NJ:Hampton Press.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1-25.

Gal, I., & Garfield, J. B. (Eds.) (1997). *The assessment challenge in statistics education*. Amsterdam: IOS Press and the International Statistical Institute.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard Education Press.

Jennings, J. (2010). The policy and politics of rewriting the nation's main education law. *Phi Delta Kappan*, *92*(4), 44-49.

Linn, R. L. (2008). *Measurement issues associated with value-added methods*. Retrieved from www7.nationalacademies.org/bota/VAM_Robert_Linn_Paper.pdf

McCaffrey, D. F., & Lockwood, J. R. (2008). Value-added models: Analytic issues. Retrieved from http://www7.nationalacademies.org/bota/VAM_McCaffrey_Lockwood_Paper.pdf

Mitgang, L. (2012). *The making of the principal: Five lessons in leadership training*. Retrieved from http://www.wallacefoundation.org/knowledge-center/school-leadership/effective-principal-leadership/Pages/The-Making-of-the-Principal-Five-Lessons-in-Leadership-Training.aspx

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

Ornstein, A. C. (1988). The evolving accountability movement. *Peabody Journal of Education, 65*(3), 12-20. Retrieved from http://www.jstor.org/stable/pdfplus/1492816.pdf?acceptTC=true&jpdConfirm=true

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. J*ournal ofStatistics Education*, *10*(3). Retrieved from http://www.amstat.org/publications/jse/v10n3/rumsey2.html

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88*, 1-8.

Watson, J. (1997). Assessing statistical thinking using the media. In I. Gal & J. Garfield (Eds.).*The Assessment Challenge in Statistics Education* (pp. 107-121). Amsterdam: IOS Press and International Statistical Institute.