

Defining the Effective Teacher: Current Arguments in Education

Tim Markley

White Mountains Regional School District
New Hampshire

Abstract

The high-pressure educational world of today demands accountability of its practitioners. Legislators are reluctant to increase educational funding without exacting a corresponding increase in accountability. This increase in accountability means that educational leaders must be able to assess and identify quality teachers. To assist in this, educators need to review the current state of evaluation practices. This article provides a review of the history of teacher evaluation and a discussion of emerging trends in teacher evaluation.

Introduction

The evaluation of teachers is not a new issue. Teacher evaluation existed in the days of the one-room school. The initial purpose was to determine job continuation and pay increases. Evaluation tended to happen at the local level with standards that were based upon local education objectives. The industrial revolution brought about some changes in the evaluation process as schools became larger and unions started to exert their influence. Unions started to set specific evaluative criteria for teachers and rules for dismissal and advancement. These criteria tended to be minimal and were still dominated by local boards of education. During the 1950s, more men entered the teaching profession. There also emerged an increase in professional activity and union membership. Sputnik and The Cold War focused additional attention on education by raising fears that Soviet students were better educated than American students. The cold war brought about the desire to find better teachers in order to compete with the Soviets. This led to even more men entering the teaching profession and unions increasing their influence. Clark (1993, p. 7) said, "Their influence and role in evaluation of teachers offered the profession the respect long overdue." America prospered, and students went to college in larger numbers than ever before (Clark, 1993).

The *A Nation at Risk Report* (National Commission on Excellence in Education, 1983) changed the educational landscape by telling the country that education was again in trouble, and students were not learning and lacked even basic skills. Clark (1993) described what happened in the intervening decades:

Education had evolved into a system based on the premise that teacher-proof curriculums, test-based instructional management, and student competence testing alone would improve learning. These policies assumed the adherence to a predetermined teaching format would result in the desired level of learning. Teachers were viewed as laborers implementing a prescribed program in a manner determined by policy makers further up the educational hierarchy rather than as professionals with a repertoire of techniques and the ability to decide for themselves how techniques should be applied

(p. 7)

One of the prime outgrowths from the *A Nation at Risk* (National Commission on Excellence in Education, 1983) report was the effective schools movement. Effective schools emphasized minimum requirements to be a teacher and the ability to implement specific correlates, such as punctuality and provision of a safe learning environment. Teacher evaluations gained a new importance as a call for effective teachers spread across the United States. Systems, such as North Carolina's TPAI (North Carolina Department of Public Instruction, 2000), had their origins in the effective schools movement of the 1980s (Papanastasiou, 1999). Clark (1993) wrote that three questions emerged from this movement; "(1) what is an effective teacher? (2) how can they best be evaluated? and (3) what can we do with this evaluation?" (p. 11) These questions still drive the debate about teacher evaluation today.

Defining the Effective Teacher

Research offered a plethora of definitions of an effective teacher. Clark (1993, p. 10) wrote that, "Obviously, the definition involves someone who can increase student knowledge, but it goes beyond this in defining an effective teacher." Vogt (1984) related effective teaching to the ability to provide instruction to different students of different abilities while incorporating instructional objectives and assessing the effective learning mode of the students. Collins (1990), while working with the Teacher Assessment Project established five criteria for an effective teacher: (a) is committed to students and learning, (b) knows the subject matter, (c) is responsible for managing students, (d) can think systematically about their own practice, and (e) is a member of the learning community (Clark, p. 11).

Swank, Taylor, Brady, and Frieberg (1989) created a model of effectiveness that was based upon teacher actions. For them, *effective* meant increasing academic questions and decreasing lecture and ineffective practices, such as negative feedback and low-level questions. The authors believed that these factors become easily identifiable in the assessment of performance. Million (1987) based effectiveness on the lesson design and method of delivery. If teachers met a preset list of criteria during their evaluation, they were deemed effective. Papanastasiou (1999) stated "that no single teacher attribute or characteristic is adequate to define an effective teacher" (p. 6).

Wenglinsky (2000) believed that the classroom practices are important to learning. In his research, Wenglinsky (2000) found that what happens in the classroom is critical and that how a teacher teaches is important. Practices that promote higher order thinking and active participation are most successful. The problem is to translate this knowledge into an acceptable evaluation procedure. Clark (1993) pointed out that "One area that was avoided by most authors was the idea of using student achievement as a measure of effectiveness" (p. 12).

Researchers appear to have taken student achievement for granted; they have believed that effective teaching techniques would automatically yield positive student achievement. Only recently has research seriously begun to look at achievement data. As Clark (1993) pointed out, the problem is determining how best to measure student achievement. The research of Sanders (Sanders, 1996, 1999; Sanders, Wright, & Horn, 1997) and others at the University of Tennessee and of Wenglinsky (2000) offered a possible solution to this question. Their work demonstrated

that teacher effectiveness can be measured and may be critical to student success. Both Sanders' (1999) and Wenglinsky's (2000) work asserted that teacher effectiveness is the single biggest contributor to student success. Teacher effectiveness outweighs all other factors, such as class size, socioeconomic status, and gender.

Sanders and his associates used data from the Tennessee Value-Added Assessment System (TVAAS) database to run multivariate analyses of students who took the Tennessee Comprehensive Assessment Program test. The results of their longitudinal study showed that teacher effectiveness is both "additive and cumulative with little evidence of compensatory effects" (Sanders & Rivers, 1996, p. 1). Sanders, Wright and Horn (1997), who followed up the original work of Sanders (1978), found that successive years with effective teachers created an "extreme educational advantage" (p. 3). Conversely, successive years of an ineffective teacher placed students at an extreme disadvantage due to the cumulative effects of poor instruction. Minority students suffered the most. African American students were twice as likely to be assigned to ineffective teachers (Sanders & Rivers, 1996).

Wenglinsky (2000) built on the work of Sanders and Rivers(1996)and others by trying to identify practices that improve student outcomes. Data from the eighth-grade science report of the National Assessment of Educational Progress (NAEP) provided the basis for this study. Wenglinsky (2000) acknowledged that this snapshot limited his study, thus providing avenues for additional research. The research showed that teacher input, professional development, and classroom practices all influence student achievement. The most significant of the three areas was classroom practices, especially those geared toward high-order thinking (Wenglinsky). Darling-Hammond (2000) studied data from the 1993-1994 Schools and Staffing Surveys and the NAEP data to gauge teacher effectiveness. The results indicated that states, such as North Carolina, that invested heavily in improvements to teacher quality and student accountability showed the greatest gains on NAEP assessments.

Not all researchers were convinced that teachers provide measurable input into student gains. English researcher Goldstein (2001) asserted, "In secondary schools, it is very difficult to ascribe the progress of any one pupil in a given subject to the teacher of that subject" (p. 4). Instead, other factors influence the student, such as other teachers, student background, and school setting. Citing Gray (1979) and Saunders (1978) as well as his own work, British researcher Long (2000) concluded that there is no established connection between teaching and learning. Long stated that "Findings from a number of different areas therefore consistently indicate that there is little variation between teachers in terms of their impact on pupil's progress" (p. 7).

Since the publication of the *A Nation at Risk* (National Commission on Excellence in Education, 1983) report, the definition of teacher effectiveness has been slowly evolving. Initially, *effectiveness* was defined by meeting a set of vague criteria associated with the effective schools movement. This evolved into the multiple strategies of methods and instruction. While this was an improvement over the earlier evaluation methods in that the role of teacher was recognized, it still left large gray areas. New research (Sanders 2002, Strauss and Vogt in press)advocated the increased use of student data, especially gain data that measure student growth from the beginning of the year to the end of the year. The definition of teacher

effectiveness most likely lies in the middle. Teachers must have adequate knowledge of the techniques and methods that are related to their profession and must understand that student learning must increase over the course of the year.

Evaluation Practices

No matter how one defines effectiveness, there is an understanding that teaching “involves a complex set of knowledge, abilities, and personal attributes in dynamic interplay” (Davey, 1991, p. 121). Davey explained that evaluating teachers is different from evaluating laborers or assembly line workers in that there is no end product to assess. Because there is no simple way to evaluate teachers, multiple methods have evolved. The most common method is classroom evaluation. One study found that 99.8% of public schools use principals’ classroom observations as the primary source of data for teacher evaluation (Sullivan, 2001). Other methods include teacher portfolios, student evaluations, value-added assessment, and peer evaluations.

The most common method of evaluation involves observation and feedback. The North Carolina Department of Public Instruction (2000) provided a typical example of this. Under this system, teachers were observed and rated on their lesson designs and teaching techniques.

There is the wide-ranging debate over what constitutes good practice. One recent report cited the lack of agreement by teacher colleges and other professionals on what constitutes good practices (Cochran-Smith, 2000). Wilson and Wood (1996) pointed out a number of flaws with this system. Observations do not take teaching differences into account; instead, observers tend to look for the same practices from different subject teachers. Principal observations force teachers to limit their performance to established evaluation criteria. The National Center for Educational Statistics (1999) noted several criticisms of observations: (a) limited competence of the principal; (b) teacher resistance and apathy; (c) role conflict for the principal; and (d) lack of expertise in specialized areas, especially at the secondary level. In North Carolina, a new teacher must be observed four times, and tenured teachers require only one observation. Sullivan (2001) pointed out that this method is not linked to student performance.

Research indicated that observation is important to teacher evaluation because teachers must demonstrate that they can perform certain preestablished competencies, such as lesson presentation and classroom management (Clark, 1993). One caution with observation evaluations was “the implication that, if it looks good, then it is good. Unfortunately, it is just not that simple” (Clark, p. 18). Both Sullivan (2001) and Clark urged a multifaceted approach that incorporated observation used with some other evaluative tool.

Other evaluation methods incorporate observations and feedback from nontraditional observers, such as students, peers, and parents. Each of these has proven less than effective in gauging effectiveness. Andrews (1995) found that “Poor instruction is not identified through the use of student ratings” (p. 5). Sullivan (2001) substantiated this finding and concluded that there is no evidence of increased student achievement with the use of student evaluations. Parent and peer evaluations showed similar results (Sullivan). Andrews found that peer evaluations tended to be even more inflated than principal observations. Papanastasiou (1999) believed that peer

reviews have merit, but research on the subject is very limited.

One method of evaluation was becoming more prevalent at the time of this writing: data-driven evaluation that is based upon student achievement. Sullivan (2001) stated, "As state and school districts begin to gain an appreciation of just how inefficient current evaluation systems are and investigate how to create data driven systems to raise performance, value-added systems of accountability are likely to become more prevalent" (p. 7). There are, however, several criticisms of student achievement tests. The first concern is that it is hard to assess teacher input into student learning. The second criticism is that standardized tests examine outcomes of students who start at different levels (Mehrens, 1998). Kohn (2000) feared that any standardized testing program would narrow the scope of teacher evaluation. Sanders and others who helped develop the TVAAS system in Tennessee tried to address these criticisms. Sanders and Horn (1995) pointed out that past efforts in the use of test data were not feasible because of the cost and lack of computing power. This meant that there was no way to differentiate educational influences from external factors. The advent of powerful computers and sophisticated software changed this equation.

The research demonstrated that observations comprise the bulk of the evaluation process, whether the principal or others conduct them. New research (Sanders and Rivers 1996, & Mendro, Gomez, & Anderson 1998) shows that data-driven evaluations are gaining acceptance and reliability. The next logical step is to combine observations with data analysis. In a search for the best evaluation model, a combination of observation and data-driven evaluation deserves a serious examination. Observation is necessary to ensure that the teacher provides instruction using accepted pedagogy and that there is an understanding of the teaching process. A data component provides a quantifiable method of determining teacher input and student learning. Combining the two would maintain the salient features of the old system and incorporate the new wave of available data.

Measuring Effectiveness

Although there exists extensive literature about how to evaluate a teacher and what method is best, there has been very little published that relates directly to quantifying teacher effectiveness. Sullivan (2001) pointed out that 99% of evaluations are the result of teacher observation. Most of these observations are then formalized into rating systems. Vogt (1984) suggested using a system with four levels of performance: (a) exceeds district expectations, (b) meets districts expectations, (c) needs improvement, and (d) is unsatisfactory. Variations of this system were still in use at the time of this writing. North Carolina's TPAI (North Carolina Department of Public Instruction, 2000) used a similar ranking system.

Frase and Streshly (1994) reported a serious flaw with this ranking system. Their research showed that teacher evaluations tended to be inflated due to tenure laws and union regulations. Their study of six eastern districts found that ratings tended to skew toward the high side of the rating scale. In most districts, none of the teachers rated "below standard." Frase and Streshly were alarmed at these results because the school districts they studied had demonstrated very poor instructional practices.

Research by Wilson and Wood (1996) raised concerns with evaluation instruments. They concluded that these instruments are not sensitive to innovative teachers or differences in teaching across content areas. Scrivens (1987) questioned the ability of administrators to judge teachers who teach outside their area of expertise. Can a history-teacher-turned-principal evaluate an effective French lesson?

Perhaps because of the problems that have been highlighted in the research, other methods of quantifying teacher effectiveness were emerging at the time of this writing. The Teacher Assessment Project (Sullivan 2001) program urged a mixed approach that included observations and portfolios. The portfolios contained evidence of knowledge and skills that were used during the school year. Student work, various teacher assessments, and lesson plans could be included in a typical portfolio. The intent was to provide documentary evidence of the teachers' competence and effectiveness (Clark, 1993). Savage (1982) advocated a portfolio system that included the artifacts of teaching. Savage believed that portfolios worked well in conjunction with observations. Portfolios, though, are not perfect and may not even be a true alternative. Alexandrov (1989) found that portfolios were not an effective tool for measuring teacher performance. Alexandrov found that there was little direct improvement in classroom instruction. Sullivan (2001) pointed out that portfolios are highly subjective and may not reflect the true ability of a teacher.

Sullivan's (2001) research into evaluation methods concluded that nearly all methods are subjective in nature and lack any connection to student achievement. The only measure that is both objective and related to student achievement is the use of test data to determine teacher effectiveness. As Sullivan pointed out, administrators must "embrace the idea that the ultimate measure of a teacher is whether his or her students are learning" (p. 19).

The origin of the idea that school factors, such as teachers, may or may not have an effect on learning can be traced to the Coleman (1966) report, which said that schools have little, if any, impact on the education of the child. Hanushek (1986) reviewed 147 empirical studies that examined the relationship between school factors and student achievement. His conclusion, as paraphrased by Lee (2001), was that "Most empirical studies have a limited set of tools mainly based on single equation regression analysis to estimate the relationship between school characteristics and student achievement" (p. 2). The fundamental problem with using test data for measuring effectiveness is that the results inevitably lead to skewed or biased evaluations. In the 1970s, Cronbach (1976) began researching effects between classrooms and not just between schools. He stated that "The majority of studies of educational effects--whether classroom experiments or evaluations of programs or surveys--have collected and analyzed data in ways that conceal more than reveal" (p. 1). Unfortunately, this means that most accountability models in the United States use test data inappropriately. Stevens, Estrada, and Parks (2000) pointed out the problem with this. They said,

If data are analyzed at the student level, the school variables are repeated exactly for each student in a school, giving a false impression of their variability. If data are analyzed at the school level (as is done in almost all state accountability systems), then all student variables within the school must be averaged, thereby losing important information about student differences. Neither analytic approach is correct, and each will result in biased

interpretations of the true relationships among the variables of interest (p. 12).

Raudenbush and Bryk (1986) concurred with the idea that educational policy is a multilevel problem but that researchers have used traditional multilevel approaches that produce erroneous results.

Only recently have researchers recognized these limitations in data analysis and begun looking for other approaches to the problem. Williams (1999) pointed out that, 20 years ago, the question was what level was the appropriate one for analysis: the student, the classroom, or the school. This separation of levels is wrong, and only recently have researchers understood this and begun to examine schools as hierarchies that are interrelated. This has led to the development of a new tool in educational research, called *multilevel models*, *mixed effects models*, *covariance components models*, or *mixed linear models*. The name used for the purposes of this study is Bryk and Raudenbush's (1992) *HLM*.

The origins of HLM can be traced to the work of C. R. Henderson, a pioneer in statistical modeling. Henderson's work provided the basis for much of the research of W. Sanders, who was one of the first to see the implications of using mixed models in education (McLean, Sanders & Stroup, 1991). Raudenbush and Bryk (1986), two more pioneers in the field, believed that education is a field with multiple levels and that data can be analyzed at different levels. They believed that "The new approach provides a flexible statistical tool for studying how variations in policies and practices influence the educational process" (p. 3).

Sanders' background was in genetics and agriculture, but he was given the task of creating one of the first true HLM models in Tennessee. Sanders (Sanders & Horn, 1994) developed complex statistical analyses to examine student scores on the Tennessee Comprehensive Assessment Program test to measure student achievement gains over time. The data are used to assess how effectively a teacher or school increases what students know. This is what Sanders called "value-added" (Sanders & Horn, 1998). What makes TVAAS different from other systems is that there are 3 or more years of longitudinal student data available for analysis. Sanders claimed that it is possible to assess individual teacher effectiveness using a form of HLM statistical analysis (Sanders & Horn, 1994, 1995, 1998). Sanders, Rivers and others stated that teacher effectiveness is paramount to student success and that the effects, both positive and negative, are cumulative. One feature of the TVASS system is the use of gain scores instead of raw scores in measuring teacher effect (Sanders & Rivers, 1996).

TVAAS is not without its critics. Several reviews of the program have raised some concerns. Baker, Xu, and Detch (1995) raised concerns about the variation in scores from year to year, problems with identifying teacher factors, and data collection issues. Sanders (Sanders et al., 1997) stated that this system answers many of these concerns and is a valid indicator of teacher effectiveness. Bock, Wolfe, and Fisher (1996) also found some areas of concern with the model. These included problems with missing data, the number of questions on the new science and social studies test, test changes over time, and some test administration issues. Bock, Wolfe, and Fisher (1996) were concerned about the use of national norming numbers in the TVAAS system. Although there were concerns, the researchers did conclude:

We agree that the central concept of the assessment system is the only present, fair, objective, and dependable methods of evaluating teacher effectiveness based on scores and the measurement of achievement gain shown by students during a period a teacher is responsible for their instruction in the subject matter measured by the test (Bock, Wolfe, & Fisher, 1996, p. 69).

Bock, Wolfe, and Fisher (1996) concluded that the underlying statistical model was reasonable. A follow-up assessment by Fisher (1996) raised other questions. Fisher was primarily concerned with contractual problems and data collection problems. He noted that the complexity of the model hinders its effectiveness as an assessment tool.

Goldstein (1997), a leading British HLM researcher, highlighted another common complaint about the TVAAS model. This is the lack of an independent evaluation of the model. With the exception of the one study cited above, studies that provide details about the statistical model that was used have been lacking. Instead, there have been a number of articles (Sanders, 1999, Sanders & Horn, 1994, 1995, 1998, Sanders & Rivers, 1996) that explained the results and conclusions but left out the procedures. Goldstein believed, though, that HLM has a place in educational research. "The statistical models now available, together with the powerful and flexible software, enable researchers to explore the inherently complex structure of schooling in a manner that begins to match that complexity" (Goldstein, p. 18).

Additional work using HLM analysis was done by Webster and Mendro (Millman, 1997) when they helped to create the Dallas Value-Added Accountability System. The Dallas system is a two level HLM model that examines student and school achievement. The results provide the assessment of the effectiveness of each school and the progress of individual students. Although the Dallas model has received good reviews, not everyone is convinced. Researchers Thum and Bryk (Millman) said, "On balance, there is much to commend about this work. Nevertheless, our examination of the system provides reason for caution" (p. 108). Their reasons for caution include complexity and the inherent problems of using test data to measure student outcomes.

Bryk has become one of the leaders in HLM research. He teamed with Raudenbush (1992) to create a computer program for HLM analysis. The version that was current at the time of this writing was HLM5.04 (Raudenbush, Bryk, Fai Cheong, & Congdon, 2001). The advances in statistical theory and the advent of powerful, affordable computers and software enabled researchers to solve much more complex problems. Although the system is complex, that should not preclude TVAAS or similar systems from using statistical modeling as an evaluation tool. Even Wenglinsky (2000) used student results on standardized tests to measure teacher practices against student achievement.

Adequate measurements of teacher effectiveness have been lacking. What evaluation methods exist are highly subjective and bear little connection to student achievement. If one agrees with Sullivan (2001) that ensuring student learning is the prime purpose of evaluation, then the new statistical models of Sanders (1999) and others (Raudenbush & Bryk, 1986) offer the best hope for creating a truly useful evaluation measure.

Summary

The review clearly demonstrated that there are problems associated with teacher evaluation practices. The problems include evaluation inflation, highly subjective instruments, and lack of objective measures. The review also showed that teachers are important to student achievement and that effectiveness is quantifiable. The work of Sanders and Horn (1994, 1995), Wenglinsky (2000), and Darling-Hammond (2000) demonstrated that teacher practices correlate to student performance. This is a change from the beliefs of the past. Research also showed that there are a number of ways to assess teacher quality, but each has its limitations. Observations provide insight in the mechanics of teaching but are subjective and fraught with problems when used as a measure of effectiveness. Although student achievement tests are not without critics, the recent advances in computer technology and software sophistication offer a possible area for study. A Sanders-type value-added system is the next logical step if one believes that teachers are important to improved student outcomes. Value-added takes the aggregate data of Wenglinsky (2000) and provides the mechanism by which individual effect can be measured. The underlying principles of value-added, demonstrated in Dallas and Tennessee, are worthy of further investigation. The primary problems that concerned the reviewers were correctable procedural errors. There were few problems with the underlying statistical model or the idea that teacher quality is measurable. For combining evaluation models, new computer technology, and advanced statistical theory, HLM value-added assessment offers an excellent tool in the evaluation of teachers. Sanders (1999) cautioned that value-added methodology is not a stand-alone assessment but a new tool in the overall evaluation program.

The review also provided a means to define teacher effectiveness or teacher quality. Paraphrasing Clark (1993) and Sullivan (2001), an effective teacher is one who demonstrates knowledge of the curriculum, provides instruction in a variety of approaches to varied students, and measurably increases student achievement. The best means to measure this is with an approach that combines observation with data-driven assessment.

References

- Alexandrov, D. (1989). *Teacher evaluation in an era of reform*. Unpublished evaluative report.
- Anderman, P. (2002). Hierarchical Linear Modeling (HLM). In *University of Kentucky, Department of Psychology Web Site for Class Handouts*. Retrieved December 17, 2002, from http://www.coe.uky.edu/EDP/707/HierarchicalLinearModeling_HLM.ppt
- Andrews, H. (1995). *Teachers can be fired: The quest for quality*. Chicago: Catfeet Press.
- Baker, A., Xu, D., & Detch, D. (1995, April). *The measure of education: A review of the Tennessee Value-Added Assessment System*. Nashville, TN: Office of Educational Accountability.

- Bock, R., Wolfe, F., & Fisher, T. (1996, April). *A review and analysis of the Tennessee Value-Added Assessment System*. Nashville: Tennessee Comptroller of the Treasury, Office of Educational Accountability.
- Byrk, A., & Raudenbush, S. (1992). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Press.
- Byrk, A., & Raudenbush, S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Press.
- Clark, D. (1993, June). *Teacher evaluation: A review of the literature with implications for educators*. Unpublished Seminar Paper, California State University at Long Beach.
- Cline, F., & Fay, J. (1996). *Teacher handbook: Discipline with love and logic*. Golden, CO: Love and Logic Press.
- Cochran-Smith, M. (2000, April). *The outcomes question in teacher education*. AERA Vice Presidential Address for Division K (Teaching and Teacher Education). Under review for publication in *Teaching and Learning*. Retrieved March 28, 2001, from <http://www2.bc.edu/~cochrans/mcsvpaddress.html>
- Coleman, J. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.
- Collins, A. (1990, March). *Transforming the assessment of teachers: Notes on a theory of assessment for the 21st century*. Paper presented at the annual meeting of the National Catholic Education Association, Boston, MA.
- Cronbach, L. (1976). *Research on classroom and schools: Formation of questions, design, and analysis*. Occasional paper of the Stanford Evaluation Consortium.
- Darling-Hammond, L. (2000, January). Teacher quality and student achievement: A review of state policy evidence. In *Educational Policy Analysis Archives*, 8(1), Retrieved October 20, 2000, from <http://olam.ed.asu.edu/epaa/v8n1/>
- Davey, B. (1991). Evaluating teacher competence through the use of performance assessment task: An overview. *Journal of Personnel Evaluation in Education*, 5(1), 121-132.
- Fisher, T. (1996, January). *A review and analysis of the Tennessee Value Added Assessment System*. Tallahassee: Florida Department of Education.
- Frase, L., & Streshly, W. (1994). Lack of accuracy, feedback, and commitment in teacher evaluation. *Journal of Personnel Evaluation in Education*, 8(1), 47-57.

- Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, (4)1. 367-395.
- Goldstein, H. (2001). *Using pupil performance data for judging schools and teachers: Scope and limitations*. London: University of London.
- Gray, J. (1979). *Reading progress in English infant schools: Some problems emerging from a study of teacher effectiveness* (Research Report No. RR216). London: University of London
- Hanushek, E. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24(3), 1141-1177.
- Hox, J. (1995). *Applied multilevel modeling*. Amsterdam: TT-Publications.
- Hox, J. (1998). *Multilevel modeling: Why and where*. Occasional paper. Amsterdam: University of Amsterdam.
- Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.
- Lee, Y. (2001, November). *Measuring school effects on student achievement in hierarchical school system: A comparison of multiple techniques*. Paper prepared for a presentation at the School of Public Policy and Management of The Ohio State University, City, OH.
- Long, M. (2000, November). *Teacher effectiveness: Do teachers matter*. Retrieved October 21, 2002 from http://www.psych-ed.org/Topics/teacher_effectiveness.htm
- Maas, C. J., & Hox, J. (2001). *Robustness of multilevel parameter estimates against small sample sizes*. Utrecht, the Netherlands: Utrecht University.
- Maas, C. J. M., & Hox, J. (2001). *Sample size for multilevel modeling*. Utrecht, The Netherlands: Utrecht University.
- McLean, R., Sanders, W., & Stroup, W. (1991, February). A unified approach to mixed linear models. *The American Statistician*, 45(1), 54-64.
- Mehrens, W. (1998, July). Consequences of assessment: What is the evidence. *Educational Policy Analysis Archives*, 6(13), Retrieved March 28, 2001, from <http://epaa.asu.edu/epaa/v6n13.html>

- Mendro, R., Jordan, H., Gomez, E. & Bembry, K. (1998, June). *An application of multiple regression in determining longitudinal effectiveness*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA
- Meyers, R. (2000). Value-added indicators: A powerful tool for evaluating science and mathematics programs and polices. *National Institute for Science Education Brief*, 3(3).
- Million, S. (1987, May). *Demystifying teacher evaluation: The multiple-strategies model used as an assessment device*. Paper presented at the annual meeting of the National Council of States on in-Service Education, San Diego, CA.
- Millman, J. (Ed.). (1997). *Grading teachers, grading schools*. Thousand Oaks, CA: Corwin Press.
- National Center for Educational Statistics. (1999, January). *Teacher quality: A report on the preparation and quality of public school teachers*. Washington, DC: U.S. Department of Education Office of Educational Research and Improvement.
- National Commission on Excellence in Education. (1983, April). *A nation at risk: The imperative for educational reform*. Washington,DC. Government Printing Office.
- North Carolina Department of Public Instruction. (2000). *TPAI 2000 the evaluation of beginning teachers*. Raleigh, NC: Author.
- Papanastasiou, E. (1999). *Teacher evaluation*. Unpublished manuscript, Michigan State University, East Lansing.
- Raudenbush, S., & Bryk, A. (1986, January). A hierarchical model for studying school effects. *Sociology of Education*, 59(1), 1-17.
- Raudenbush, S., Bryk, A., Cheong, Y.F., & Congdon, R. (2001). HLM5: Hierarchical linear modeling (Version 5.04) [Computer software]. Lincolnwood, IL: Scientific Software International.
- Raudenbush, S., Rowan, B., & Cheong, Y.F., (1992). Contextual Effects on the self-percieved efficacy of high school teachers. *Sociology of Education* 65(2), 150-167. Retrieved January 16, 2002, from <http://www.jstor.org>
- Sanders, W., & Rivers, J. (1996). Cumulative and residual effects of teachers on future student academic achievement (Research progress report). In *University of Tennessee Value-Added Assessment Center, Knoxville, TN*. Retrieved March 28, 2001, from http://mdk12.org/practices/ensure/tva/tva_2.html
- Sanders, W., Wright, W., & Horn, S. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 4(1), 3-7.

- Sanders, W. (2002, March). Keynote address to the North Carolina Association of Supervision and Curriculum annual conference, Pinehurst, NC.
- Sanders, W., & Horn, P. (1995, March). Educational assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. *Education Policy Analysis Archives*, 3(6). Retrieved March 28, 2001, from <http://o lam.ed.asu/ epaa/v3n6.html>
- Sanders, W. (1999, Fall). Teachers! Teachers! Teachers! *Blueprint Magazine, Online edition*. Retrieved March 28, 2001, from <http://www.ndol.org/ blueprint/fall/1999/solutions4.html>
- Sanders, W., & Horn, P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8(1), 299-311.
- Sanders, W., & Horn, S. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3), 247-256.
- Saunders, J. (1978). Teacher effectiveness: Accepting the null hypothesis. *The Journal of Educational Thought*, 12(3), 184-189.
- Savage, J. (1982). Teacher evaluation without classroom observation. *National Association of Secondary School Principals Bulletins*.
- Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education*, (8)1, 9-23.
- Stevens, J., Estrada, S., & Parks, J. (2000, April). Measurement issues in the design of state accountability systems. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA
- Strauss, R., & Vogt, W. (in press). It's what you know, not how you learned to teach it: Evidence from a study of the effects of knowledge and pedagogy on student achievement.
- Sullivan, C. (2001). Rewarding excellence: Teacher evaluation and compensation. Alexandria, VA: National School Boards Association.
- Swank, P., Taylor, R., Brady, R. & Frieberg, T. (1989). Sensitivity of classroom observation systems: Measuring teacher effectiveness. *Journal of Experimental Education*, 57(2), 171-186.
- Vogt, W. (1984, Winter). Developing a teacher evaluation system. *Spectrum*, 2(1), 41-46.

Webster, W. J., & Mendro, R. L. (1997). Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools* (pp. 81-99). Thousand Oaks, CA: Corwin Press.

Wenglinsky, H. (2000, October). *How teaching matters: Bringing the classroom back into discussions of teacher quality*. Princeton, NJ: The Milken Family Foundation and Educational Testing Service.

Williams, J. D. (1999). Basic concepts in hierarchical linear modeling with applications for policy analysis. *Handbook of educational policy* (pp. 473-493). New York: Academic Press.

Wilson, B., & Wood, J. (1996). Teacher evaluation: A national dilemma. *Journal of Personnel Evaluation in Education*, 10(2), 75-82.