PART VI

Analyzing teaching

Introduction

There are many reasons why we want to know more about how teachers teach their subject matter. Researchers are interested in studying excellent teachers to better understand the elements of effective teaching and to use this information to better prepare preservice teachers. Likewise, novice or struggling teachers can benefit from feedback from teacher assessments to improve their skills and better manage their classrooms. Principals and supervisors use formative assessments of teaching as primary tools to assist teachers to learn new strategies or improve current practice, while teachers receive summative assessments associated with annual or periodic reviews to provide information for school or grade level placement or other personnel decisions.

Recently, in the United States, the analysis of teaching has taken on heightened intensity as teacher job security and salaries have become tied directly to classroom performance, particularly student learning. Neoliberal perspectives on high stakes testing have led to finger-pointing when students do not score well on these tests or do not improve from year to year based on external metrics of success. The natural next step is to assess the teaching learning environment and particularly teaching behavior as the source of the problem.

For the last decade, however, physical educators have been buffered from the more punitive forms of teacher evaluations due to the peripheral position and perceived low status of physical education within the school environment. With most of the high stakes testing and teacher evaluations targeted at literacy and mathematics, physical educators, though forced to teach increasingly larger classes, for shorter time periods, and with fewer resources, have not received the same level of scrutiny as have their classroom colleagues. Increasingly, though, the federal government has offered large state level multimillion dollar grants, such as “No Child Left Behind” and the most recent “Race to the Top,” as carrots to entice states and school districts to evaluate teachers in all subject areas as one condition to receive this funding (United States Department of Education [USDE], 2001, 2009). Many state Departments of Public Instruction have taken on the task of designing evaluation systems to assess teacher effectiveness in all subject areas, including physical education.

The chapters in this part look at a range of trends and issues associated with teaching analysis and teacher effectiveness. In the first chapter (25), McCullick, Gaudreault, and Ramos
open this part with a comprehensive review of research associated with teacher accountability in the United States. They begin by describing many of the concerns associated with the teacher accountability movement and highlight the contradictory research results that provide the foundation on which teacher effectiveness decisions are currently being made. They then distinguish between performance-based and student growth/value-added accountability models currently competing for acceptance in states and school districts across the United States. Performance-based models evaluate the teaching process, while student growth models evaluate teachers using student outcomes (typically based on student scores on high stakes achievement tests). The authors explain the theory and policies associated with these two approaches and detail a number of issues associated with use of these teacher evaluation models in the classroom and gymnasium.

It appears clear that physical education teachers are experiencing teacher accountability initiatives and are assuming a precarious place in the high stakes environment. These efforts to hold teachers directly responsible for student learning in their classes shows a blatant ignorance of or disregard for the plethora of factors that are known to impact students’ opportunities to learn. Physical education professionals at every school level will need to understand the consequences of value-added, student growth accountability models to have even a reasonable chance of steering clear of the pitfalls.

Performance-based models have been used for many years in education and physical education to provide formative feedback and assessments to teachers in physical education. Subramaniam and Wuest’s chapter (26) examines several methods for measuring teaching in physical education. After a brief history of research on the measurement of teaching in physical education, they describe several quantitative methods with particular emphasis on systematic observation systems. While systematic observation has a rich history in physical education beginning with the work of Cheffers (1977) and Metzler (1983), these systems are finding new and varied purposes when coupled with wireless technologies and mobile device applications. The authors conclude by glimpsing the future of ecologically valid instruments that allow researchers to measure student learning in authentic game play settings.

This section concludes with Cale’s (27) analysis and critique of the issues surrounding teaching about active lifestyles. She summarizes and critiques the research evidence on the role of physical education as a positive influence on young people’s physical activity. She then considers issues teachers face when teaching about active lifestyles. This includes critiquing current lifestyles pedagogy and some common practices, notably fitness testing. Although highly touted when first developed, fitness tests have not proven to be as motivating to students as once predicted. Unfortunately, they hold potential to become the “high stakes” tests in physical education similar to the standardized tests in reading and mathematics used to evaluate classroom teacher effectiveness. Cale explains that fitness tests/student test scores can become the “student growth” measure used to compare physical education teachers and evaluate their “effectiveness.” She provides a powerful critique of these tests and concludes with a futuristic assessment of issues and concerns for the future.

Catherine D. Ennis
University of North Carolina at Greensboro, USA

Stephen J. Silverman
Teachers College, Columbia University, USA
Part VI introduction

References


In many professions, relatively simple calculations and logically linked outcome measures determine one's effectiveness. Two such examples are sales and law. The evaluation of sales professionals is based on the dollar amount brought in on a yearly basis while prosecuting attorneys' effectiveness is usually dependent on the number of cases won and lawbreakers brought to justice. To our knowledge, few debate whether these measures are fair let alone valid and reliable. Many would accept the premise that if a real estate agent sold $1 million worth of real estate in a calendar year then he was an 'effective' sales professional. Similarly, a prosecuting attorney who won 50% or more of her cases would be regarded almost universally as an effective lawyer. Although there is room for debate, it is safe to say that, in both cases, a direct line can reasonably be drawn between the individual's work and the outcome.

When it comes to evaluating teacher effectiveness, however, the process and confounding factors make the practice considerably less straightforward. If it were easy to measure good teaching by readily determined metrics such as profit margins or winning records in the courtroom, there would be little need for this text to address the topic. As it is, the identification of teaching that results in student learning is an exercise that is marred with a multitude of variables and an inability to control for them. Teacher effectiveness literature indicates the importance of understanding that directly connecting teacher behavior and student achievement and "equating effectiveness with success in producing achievement gain" (Brophy & Good, 1986, p. 328) is a slippery slope at best.

Perhaps the culprit preventing effective teaching from easy measurement of student achievement is that "the difficulty in identifying the concept of effectiveness in teaching lies in the complexity of teaching" (Rink, 2014, p. 408). That is, a teacher's 'product' – student learning, is heavily influenced by a host of dynamics beyond the control of schools and teachers. Even the research base identifying what teacher behaviors are most indicative of teacher effectiveness has cautioned that a relationship between effective teaching behaviors and student learning is not as easily conceptualized because "in reality, instruction always occurs within particular contexts" (Brophy & Good, 1986, p. 369) and those contexts heavily influence student achievement.
This ‘context’ makes equating teacher behaviors with student achievement such a dubious task. The nature of schools and other overwhelming variables such as students’ previous school experiences, resources outside of school necessary to practice and learn, and opportunities within home life mediate the effects of teacher behaviors on student learning (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012). Any individual who has spent one day in a school could attest to this assertion. Incredulously, it is those writing and enacting policy who seemingly fail to recognize the complexity of this issue. In turn, it allows for a much ‘cleaner’ determination of teacher performance. Conversely, there are many concerns resulting in myriad issues associated with labeling teachers effective or ineffective since the bulk of research indicates that most popular modes of teacher evaluation are “fraught with inaccuracies and inconsistencies” (p. 8). As a result, these individuals stand as advocates for using student achievement gains to measure teacher effectiveness and, eventually, make personnel decisions.

For physical education (PE) teachers the evaluation of teacher effectiveness provides additional concerns. A majority of research and debate about teacher effectiveness guiding state and national policy centers on the classroom with little attention being paid to content areas that are taught in different contexts. PE scholars have given much attention to teacher effectiveness research but only recently has the debate on evaluating PE teachers’ effectiveness based on student performance been brought to the forefront (Roberts, 2014). As with the debate regarding teachers in ‘regular’ classrooms, the debate regarding PE teacher effectiveness has similar problems. Specifically, the amount of instructional time, class sizes, and available resources are areas in which teaching in the gymnasium is different from the classroom. At the same time, PE brings with it additional complexities of marginalization, lack of resources, isolation, and its historical connection to coaching sport, all of which make for a strong argument that in PE drawing a connection between teacher effectiveness and student outcomes becomes an even flimsier proposition than in any other subject area.

While hardly anyone would take a stance that teachers should not be held reasonably accountable for effective teaching, the debate seems to center on how teachers are held accountable, what measures determine it, and the impact on teachers’ professional lives when employment, advancement, and rewards are tied to these measures. Therefore, the purpose of this chapter is three-fold and will highlight: (a) confounding dynamics in the evaluation of teacher effectiveness, (b) widely used models of assessment, and (c) the impact of this movement on the teaching of PE and PE teachers. The chapter concludes with reflections on where teachers, physical education teacher education (PETE) researchers, and faculty should focus their efforts in the near future.

Confounding dynamics in the evaluation of teacher effectiveness

For a concise and well-presented history on the genesis and proliferation of the teacher accountability movement in the United States, we direct readers to Metzler (2014). In short, over the last 30-plus years despite evidence to the contrary, the position that teachers should be held accountable and that there are valid and reliable measures that can assuredly indicate which teachers contribute most to student learning (Milanowski, 2004) has continued to gain support. Shockingly, the movement has pressed on and even advanced in the face of evidence suggesting that making teacher effectiveness determinations based on student outcomes is precarious, at best. Given less attention, however, are the complications that make this issue an area of fervent debate. While far from a comprehensive list, those complications are: (a) uncontrollable factors, (b) contradictory research results, (c) ineffective staffing practices, and (d) resources.
Teacher accountability

Uncontrollable factors

Regardless of the subject matter, perhaps, the most influential hindrance to the validity of teacher effectiveness measures is that of the factors influencing student learning outside the teacher’s and school’s sphere of influence. Unlike the example of the real-estate sales professionals whose products and performance are easily linked, student learning and being prepared to learn is heavily dependent on factors beyond even the student’s control. Perhaps the most influential of these factors is the student’s socio-economic status (SES) and associated “sub-components” that have far-reaching impact on resources which contribute most to student achievement (Hattie, 2013). Additionally, data suggest that influencing student achievement are factors such as teacher turnover (Ronfeldt, Loeb, & Wyckoff, 2013) and parental aspirations and expectations (Hong & Ho, 2005).

At the same time, students bring with them a number of factors that affect their own achievement. Among others, children’s prior school achievement, dispositions and attitudes, and pre-school experiences appear to have the most powerful effect on student achievement (Hattie, 2013). Arguably, while schools and teachers can possibly foster these, the evidence suggests that they are best done so in pre-schools. Whether children are able to get those experiences, however, is not within the control of the school or teacher.

Contradictory research results

We give more attention to this subject in a subsequent section of this chapter, but the body of knowledge on teacher evaluation measures is inconsistent. There are scholars positing that “rigorous teacher evaluation system can be substantially related to student achievement” (Milanowski, 2004, p. 34), while others claim that, even those systems which are statistically strong, produce ratings that “differ substantially from class to class and from year to year, as well as from one statistical model to the next” (Darling-Hammond et al., 2012, p. 9). Further complicating the issue is that both sides put forth strong arguments that would leave even the most independent arbitrator unconvinced that either had the stronger case. In the end, not only have scholars put forth much effort to create and test evaluation models, but they have also worked just as hard at convincing others of their position. Without consensus, this leaves policy makers to their own devices when taking a side. When those policy makers are individuals with little knowledge of teaching, learning, and how schools operate, expediency takes precedent over appropriateness.

Ineffective staffing practices

Perhaps the single-most controversial outcome of the teacher evaluation debate concerns how evaluation results should influence personnel decisions. Teacher evaluations are usually linked to retention, promotion, or salary increases. Economists have recognized that one purpose of employee evaluation is to motivate employees to improve their work (Taylor & Tyler, 2011). However, they also recognize that one problem with this is that “evaluation induced improvements may … be transient if the incentives only operate on behavior that contributes to an evaluation score. By contrast, the effects may be more likely to persist if the evaluation spurs employees’ investment in human capital” (p. 1).

Teacher evaluation, at best, can be a vital component of teacher development. However, as Murphy, Hallinger, and Heck (2013) noted, the definition of the term evaluation connotes a superior-subordinate relationship fraught with the tensions that accompany it—especially when the results of the evaluation are used to make decisions about the terms of employment” (p. 349).
The idea that measuring teacher value based on student outcomes is seemingly too complex to allow the same approaches used in business and other professions to be applied to teaching.

**Resources**

Arguing against the high cost of evaluating teaching is low-hanging fruit for anyone opposed to teacher evaluation and is hardly adequate to dissuade those interested in using evaluation to improve schools. However, one cannot ignore the time and money required to conduct meaningful teacher evaluations. Without a well-funded and well-designed evaluation, a partial or compromised teacher evaluation can result, causing personnel and political repercussions.

Ideally, the basic tenet driving teacher evaluation and accountability is the intent of improving students’ experiences in schools. The evidence, it appears, indicates that if this is the end-point, then spending “time and energy in areas other than teacher evaluation” (Murphy et al., 2013, p. 352) is perhaps a more efficient way of reaching it. There is little doubt that expending significant funds is required for the assessment of teacher effectiveness. In the United States, spending on education is often a topic that follows political party lines and possibilities for consensus are low. Inevitably, policy makers find compromise by agreeing to use the easiest and most cost-effective manner of conducting evaluations, often leading to the use of invalid and unreliable instruments and hiring untrained evaluators to conduct the evaluations. Both lead to poor data and have far-reaching implications for teachers, schools, and, most importantly, children.

**Models of teacher evaluation**

Educational reforms in the last 50 years focusing on teacher quality have been concurrent with an increase in federal policies related to teacher effectiveness (Cohen-Vogel, 2005). Consequently, the improvement of teacher evaluation systems has received amplified attention as a way to increase teacher effectiveness and student achievement.

There has been a recent push at the national level in the United States for school reform with an emphasis on teacher accountability and quality. The federal government’s introduction of ‘No Child Left Behind’ in 2002 and the ‘Race to the Top’ (RTTT) competition in 2009 intensified the focus on improving accuracy of teacher evaluation systems to measure teacher effectiveness. Outcomes of these systems are used for high-stakes teacher decisions such as tenure, salary increases, and dismissal. As a result, holistic teacher evaluation systems have emerged employing standards-based, multi-dimensional, and formative approaches using student outcome data. Since individual states and school districts have freedom to define and quantify teacher effectiveness, many have developed their own teacher evaluation systems based on established, modified, or locally developed models.

Primarily, teacher evaluation systems are composed of two types of models: (a) *performance-based* and (b) *student growth*. *Performance-based* models evaluate teacher quality with a focus on teacher-controlled variables such as teaching skills, dispositions, professional development, and assessment of student learning. *Student growth* models are statistical models that use norm-referenced statistics and may use non-teacher controllable factors associated with student achievement (e.g., student’s background and school characteristics) to calculate student growth. In essence, *performance-based* models evaluate *process* while *student growth* models evaluate *outcomes* to measure teacher effectiveness. The following discussion on teacher evaluation models in the United States and internationally will provide a description and examples of these models.
Teacher accountability

The performance-based models emerging in recent decades include teacher performance assessments and standards-based evaluations. Teacher Performance Assessments began in 1987 with the National Board for Professional Teaching Standards’ (NBPTS) evaluations for accomplished teachers (Darling-Hammond, 2010). Teachers voluntarily chose to participate in these evaluations and collected and organized evidence of their performance in a portfolio with lesson plans, student learning evidence, and video-recorded teaching lessons with commentaries. Trained experts scored these artifacts with rubrics that evaluated dimensions of teaching. California adopted this method to assign additional pay for teachers who had attained NBPTS certification. The Council of Chief State School Officers (CCSSO) following the NBPTS criteria created the Interstate New Teacher Assessment and Support Consortium (InTASC) standards. The InTASC standards have been incorporated in an evaluation approach used by the Council for the Accreditation of Educator Preparation (CAEP) for teacher education programs accreditation.

Standards-based evaluations are based on teaching standards grounded in teaching and learning research literature (Darling-Hammond, 2010). One example is the Danielson Framework which is a set of research-based instructional components aligned with CCSSO’s InTASC standards. It can be used to evaluate teachers’ planning and preparation, classroom environment, instruction, and professional responsibilities. Another example is the Marzano Teacher Evaluation Model which can be used to evaluate teachers’ classroom strategies and behaviors, planning and preparing, reflection, collegiality, and professionalism. Both are designed to increase conversations among practitioners for enhancing teaching skills while also being used for evaluation, professional development, mentoring, and coaching. These instruments utilize indicators such as analysis of lesson and unit plans, classroom observations, teacher participation in the community, student feedback, and teacher conversations to provide research-based evidence of teacher practices associated with student outcomes.

In addition to using models that evaluate teacher performance, teacher evaluation systems incorporate student growth models in their assessment of teacher effectiveness. The most commonly used models include Value-Added Models (VAMs) and Student Growth Percentiles (SGPs). Federally funded Race to the Top initiatives require states to adopt data on student growth as a measure of teacher effectiveness. Thus, government policy and funding have been instrumental in driving recent use of student growth models in teacher evaluation systems (Sledge & Pazey, 2013).

VAMs’ objective is to measure teachers’ yearly contributions to student learning and achievement based on standardized tests while accounting for student background and other factors related to student achievement beyond the teacher’s control (e.g., attendance and economic status). These models use hierarchical linear modeling to predict individual student scores using the students’ previous scores while accounting for specific variables and then compare the prediction to the student’s actual score and other students’ scores with similar performances. The difference between the scores is attributed to the impact, or value that the teacher ‘added’ to the students’ actual scores. States have assigned various values to VAMs in teacher evaluation. Arizona requires that VAMs must be weighted between 35 and 50% of the teacher evaluation while Colorado, Florida, and Idaho require a minimum of 50% (Baker, Oluwole, & Green, 2013).

SGPs differ from VAMs in that they do not account for student differences and student background. SGPs use a statistical method that calculates the relationship between a student’s current and previous scores in comparison to the student’s peers. Thus, if a student has a SGP of 90, it means that the student had more growth than 90% of similar students.

Internationally, countries have opted to use performance-based models alone or have complemented these models with student growth models. It appears that while the United States uses standardized assessments to measure educational quality, other countries are focusing on instructional practices and curriculum quality, using multiple measures (mainly formative evaluations)
B. A. McCullick et al.

to improve teaching instead of punishing educators or schools for low performance (Conley, 2015). Stewart (2011) noted that, in countries that employing high-performance measures such as Canada and Finland, teachers and principals are encouraged to discuss student progress, while annual teacher evaluations in Singapore and China use multiple measures such as classroom observations, community involvement, and contributions to their school. These countries rely on school-level accountability systems based on a variety of “school improvement goals, professional contributions, and indicators of student well-being” (p. 19) instead of relying solely on student achievement data. In other high-achieving countries such as South Korea, little attention is given to teacher evaluations (Darling-Hammond, 2010). Instead, these countries focus on effectiveness in the teacher preparation, support, and development to increase student achievement.

Other countries have developed student growth model-based teacher evaluation systems similar to those in the United States. In Chile, the discussion about teacher quality is centered on teacher characteristics related to students’ achievement in standardized tests and evaluation of teacher competencies (Santelices, Galleguillos, Gonzáles, & Taut, 2015). Lingard (2010) indicated that a “global trend – represented by the Anglo-American [student outcomes] model – [which] has been towards standardization” (p. 139) has driven policy changes in Australia and England, but not in other countries such as Finland, Scotland, and Wales.

**Issues associated with teacher evaluation models**

Researchers continue to explore the use of performance-based and student growth models in teacher evaluation systems (e.g., Goldhaber & Hansen, 2010; Kane & Staiger, 2012). In addition to questions surrounding their effectiveness and accuracy, findings suggest the majority of evaluations do not identify areas for teacher improvement. School districts do not consistently use results to provide beginning teachers with additional help and poor teacher performance, in general, is unaddressed (Weisberg, Sexton, Mulhern, & Keeling, 2009). Issues such as these have contributed to doubt undermining practitioners’ belief in these systems. Consensus among policy makers, practitioners, and researchers is that the majority of teacher evaluation systems do not provide school district administrators with assistance in personnel decision-making or in teacher improvement (Darling-Hammond et al., 2012; Strunk, Weinstein, & Makkonen, 2014).

**Performance-based** models are grounded in research connecting teacher performance to student achievement. Used in teacher evaluation systems for many years, these models confirm that if teachers demonstrate/possess (or not) characteristics linked to student learning, they may be deemed (in)effective. Although there is little doubt these methods can be effective (e.g., Cantrell & Kane, 2013; see www.metproject.org), other issues related to accuracy (due to human error) and teacher impact (attribution of student learning to the teacher) indicate teachers are only one factor impacting student achievement. Models do not consider multiple variables impacting student learning such as home and community challenges and support, peer culture and achievement, prior schooling and teachers, and other school factors such as instructional time, class size, and curriculum materials (e.g., Darling-Hammond et al., 2012).

**Student growth** models have received increased attention in recent years due to policy changes based on beliefs that these models can be used effectively to evaluate teacher quality. In fact, student growth models may be the hottest issue in education and are being used in teacher evaluations resulting in high-stakes decisions such as dismissal, tenure, and financial reward. VAMs have been the most commonly used student growth model in teacher evaluation systems. Researchers and statisticians have recognized the benefits of VAMs (e.g., Ballou, Sanders, & Wright, 2004; Goldhaber & Hansen, 2010). Theoretically, VAMs can indicate the impact an
individual teacher had on an individual student, and consequently, the teacher’s effectiveness. Conversely, researchers and scholars are raising issues around the use of VAMs in the evaluation of teachers (e.g., Darling-Hammond et al., 2012; Mathis, 2013; Sahlberg, 2011). According to Gabriel and Lester (2013), four methodological issues dominate the literature: (a) error rates, (b) test effects, (c) model biases, and (d) inherent problems of achievement tests.

Studies have indicated VAMs’ error rates make them inconsistent (e.g., Schochet & Chiang, 2010). Sass (2008) explored the changes of VAM scores of teachers in Florida and California who were in the bottom 20%. In the following year, between 20% to less than 35% of those teachers stayed in the same category, 35% to around 45% moved up to average level while 20% to over 40% scored above average (highest category).

Research on test effects has pointed out that the type of test used, time of the year taken, and stakes (or perceived importance) of the test can significantly change teachers’ VAM scores (Bill & Melinda Gates Foundation, 2010). Papay’s (2011) study indicated that almost 50% of teachers would have received a different category in their evaluation if the test itself, or time taken, were different. In a review of research examining potential bias in VAMs, Gabriel and Lester (2013) cited finding that minority teachers have more unstable general VAMs scores. These findings may be due to inclusion of variables in the model such as demographics, previous tests scores, and even the nonrandomized selection of students to teachers. The authors also pointed out that the use of students’ achievement test scores in VAMs assumes that these tests are an accurate measure of student learning and achievement, ignoring the inherent issues associated with standardized tests.

The issues surrounding VAMs result in a cautious approach to their use. Considering that at least one in four teachers may be misevaluated, critics argue that restraint is essential, especially in high-stakes decisions. Relying solely on VAM models could lead to ethical, legal, and practical issues, as VAMs are highly unreliable from test to test, classroom to classroom, and from year to year (Mathis, 2013). Sahlberg (2011) summarized this position stating that using only student test scores to evaluate individual teachers was inappropriate. Although VAM models have raised issues about their use in high-stakes decisions, supporters continue to call for more research to validate VAMs for teacher evaluation (Ballou et al., 2004).

Another growth model that is receiving attention is Student Growth Percentiles (SGPs). According to Baker et al. (2013), the popularity of SGPs is due to a lack of investigations on their validity and their design does not estimate teacher effectiveness. In fact, the authors warned, “student growth percentiles are not intended for attribution of responsibility for student progress to either the teacher or the school” (p. 8). Proponents of SGPs indicate that this model provides a simple and easy indication to stakeholders (parents, teachers, and administrators) of how much improvement a student had from year to year when compared to the students’ peers, which not only displays how much a student improved, but also what is the range of improvement expected for students with the same level of performance (e.g., Betebenner, 2011). The few available studies on SGPs have indicated that they should not be used in teacher evaluations because SGPs do not consider factors that are beyond a teacher’s control, and consequently, unfair for rewarding teacher performance or making high-stakes decisions (Ehlert, Koedel, Parsons, & Podgursky, 2012; Goldhaber & Walch, 2012).

As the number of states adopting student growth models into their teacher evaluation systems is increasing, there is a growing number of research studies investigating the relationship and consistency between performance-based and student growth models. The literature can be classified into studies exploring this relationship in (a) research contexts and (b) practice (Strunk et al., 2014). This differentiation highlights the methodologies and assumptions used in the studies and the implications for stakeholders involved in the use of these
measures. Strunk et al. (2014) noted the studies using research contexts have used sample sizes from 24 (Grossman, Loeb, Cohen, & Wickoff, 2013) to 3000 teachers (MET project) measuring multiple evaluation tools and have found moderate to low correlations between observation-based measures and VAMs. These studies have found a positive and moderate correlation between standardized tests and observational measures, due to issues associated with these tools that include observers’ consistency, types of measures used, and student bias toward teachers (Strunk et al., 2014). Recently, Morgan Hodge, Trepinski, and Anderson (2014) investigated the stability of teacher performance (observation evaluations) and effectiveness (standardized test scores) over a five-year period. The results indicated that neither evaluation method was highly stable over multiple years while the relationship, though small, between the methods was relatively stable over time.

In an attempt to solve issues associated with teacher evaluation systems, scholars and researchers have proposed different approaches. Taken together, these initiatives suggest a comprehensive system using multiple measures to create a fair and accurate evaluation of teacher effectiveness. Formative approaches focused on improving teachers’ weaknesses and teaching strategies are best balanced with summative approaches; although measures that de-emphasize test scores require more work, they may potentially have greater impact on instruction and improvement of education (Mathis, 2013). According to Weisberg et al. (2009), evaluation systems should: (a) be comprehensive with clear performance standards, multiple rating options, and regular feedback to teachers; (b) have trained administrators and evaluators who are regularly monitored and held accountable for using evaluations effectively; (c) use evaluations in high-stakes decisions such as reward, promotion, and dismissal; and (d) adopt low-stakes options in dismissal policies. Darling-Hammond et al. (2012) suggested most effective teacher evaluation systems are based on professional standards – with multiple observations, multiple data sources, implemented by expert evaluators, and used to provide timely and meaningful feedback to teachers. All key stakeholders (e.g., principal and teachers) must be involved in the development process. The most effective systems are specific for each teacher grade level, subject area, and experience levels (Clayton, 2014).

Overall, the literature has specified that more research is essential to improve our understanding of present evaluation systems and models. Additionally, scholars indicate that, due to the varied contexts of individual schools, there is no consensus on one evaluation system that may work across all grade levels and subjects. Despite this, the literature points to the need for meaningful changes to teacher evaluation systems to make them successful in accurately and fairly evaluating teacher effectiveness.

Physical education teacher evaluation and implications

Teacher evaluation in PE has been available for many years. Evaluators have relied on performance-based models using general evaluation tools (e.g., Danielson Framework for Teaching and Marzano Teacher Evaluation Model) and specific tools such as the PE Teacher Evaluation Tool (National Association for Sport and Physical Education [NASPE], 2007) to assess teacher practice and evaluate teacher effectiveness. Similar to other teacher evaluations, issues include the accuracy of the measure and knowledge of the evaluator about the subject. The fact remains that PE is the only school subject that requires and teaches performance of movement physical skills. This exacerbates the issue and suggests that most principals do not have the expertise or time to learn PE-specific evaluation tools. Combined with a lack of priority, it is no surprise that there has been little emphasis given to the accurate evaluation of PE teachers.
Teacher accountability

Recent incentives and renewed support for the use of student achievement data in the evaluation of teachers may soon extend to PE. The problem is that currently, due in part to the fact PE is not considered a core subject, with some exceptions (e.g., PE Metrics, NASPE, 2010, 2011) standardized testing in PE is rare. Fitness testing is the most commonly used among physical educators, even as criticism of its accuracy and negative effects on teaching practices remain (Mercier & Doolittle, 2013). In the same way classroom teachers may teach students to the test, physical educators could train students to do well on fitness testing (see Cale, Chapter 27, this volume). Furthermore, the use of fitness tests can negatively affect students’ perceptions toward physical activity. Most importantly, fitness testing does not necessarily represent students’ knowledge, skills, and dispositions. Welk (2008) noted a limitation of fitness testing is that it measures the product or outcome rather than the underlying behaviors that promote physical activity and fitness. While it can help students develop physical activity and fitness goals, it should be only one tool to help assess students’ dispositions and knowledge towards physical activity and, consequently, teacher effectiveness.

Another alternative that could assist administrators with PE teacher evaluation and PE’s pursuit of legitimacy is the creation of achievement data by the PE teachers themselves. This practice is hardly novel; as Mercier and Doolittle (2013) suggested, assessments have been the “third leg” (p. 38) of PE pedagogy (curriculum, instruction, and assessment) and support for their use by organizations including NASPE, NBPTS, and CAEP have addressed the use of assessments for instructional improvement and standard alignment. Mercier and Doolittle’s work provides guidelines for specific summative assessments (e.g., observations, written tests, and physical activity log with parental signature) that can be used to measure student achievement and provide evidence of teaching and program quality. Chief among them, student-learning objectives (SLOs) should guide selection of summative assessments that would demonstrate student achievement on goals for the unit, semester, or year. Sources of this type of assessments include PE Metrics (NASPE, 2010, 2011) and state level assessments designed by scholars. To make these assessments effective, several steps are needed: (a) a supportive environment between teachers and administrators; (b) teacher training and digital data tracking and recording; (c) effective communication with parents and students, especially if results are used for grade and school records; and (d) stakeholder understanding and value for the assessment process.

Lack of knowledge and concern from administration

Regardless of instructional level, building level principals have a significant impact on the culture and function of their respective school(s). Further, scholars have found that principal leadership has the ability to impact teacher efficacy (Soehner & Ryan, 2011). For physical educators, this can negatively impact their feelings and perceptions of their workplace as studies have shown many principals do not value physical education (Stringer, 2004). Stevens-Smith (2007) found that nearly 25% of the principals in her study described physical education as a low-priority course.

Further, most principals lack the knowledge about what constitutes a quality physical education program. They are unaware of state and national standards for physical education programming and teaching and, as a result, are ill-equipped to effectively evaluate physical educators’ teaching behaviors in the gym or curriculum. This is compounded by administrators’ personal, often negative, experiences as students in physical education.

Referred to as the ‘apprenticeship of observation’ (Lortie, 1975), administrators spent countless hours developing their conceptions of physical education and physical education teachers likely immersed in programs that could be described as ‘roll the ball out’ or ‘busy, happy,
good.' As a result, their notions of physical education subject matter and those who teach it are grounded in these images and impressions leading to inaccurate ideas about the physical educator working in their building under their supervision. With respect to evaluation and assessment, this might mean that principals enter the gym only looking to determine if children are well managed and physically active. As a result, many PE teachers’ evaluations and assessments of teaching behavior by these principals may lack value, fail to contribute to teacher development, and might be characterized as inaccurate.

**Impact on physical education and its teachers**

Principal leadership has a significant impact on the well-being and motivation of teachers (Eyal & Roth, 2011). As any teacher will tell you, a principal can make a significant difference in the climate of a building, the morale of teachers and staff, and the overall culture of a school. Further, principal leadership has the potential to influence the status of PE in schools and may impact how others in school view the value of physical education (Blankenship & Coleman, 2009). Michigan teachers in a study by Sun, Youngs, Yang, Chu, and Zhao (2012) reported receiving very few evaluations and assessments by their principals. Of the 88 principals included, only 26% evaluated their teachers each year, 40% of the teachers reported being evaluated every three years, and 6% reported that their principal had never evaluated them. If physical educators are not observed by administration, it is unlikely they will receive meaningful feedback to enhance their teaching. As a result, a physical educator’s development, growth, and improvement is dependent upon their abilities to reflect, self-assess, and seek out opportunities to connect with other teachers in their building or other physical educators in their school district.

**Future directions**

A chapter on this topic would be incomplete without addressing the increased need for advocacy associated with teacher evaluation. We will present recommendations relative to the respective constituents who would implement them. These constituents are (a) P-12 physical education teachers, and (b) PETE faculty and researchers.

**P-12 physical education teachers**

Physical educators should actively engage in their own advocacy efforts at the grass roots level. Physical educators should collect and share formative and summative assessments to document student performance and achievement. Assessments inform teacher decisions regarding curriculum, lesson planning, and lesson progression. Sharing this information within the school with other teachers, parents, administration, and students is one additional powerful advocacy tool. Professional organizations and scholars repeatedly urge practitioners to use and share data on student achievement to demonstrate the benefits of quality physical education. Given the current context of schools and educational policy, it is imperative that physical educators demonstrate the valuable place of PE and effective teaching in their schools.

**PETE faculty and researchers**

PETE faculty and researchers are essential to addressing the issue of teacher effectiveness and accountability. While not their sole province, teacher educators are usually in the best position
to provide principals with specific evaluation materials geared toward measuring PE teacher effectiveness.

PETE faculty can assist school administrators to select and use effective evaluation tools and blend information from evaluation methods to produce reliable and valid teacher evaluations. Professional organizations such as SHAPE America and the Special Interest Group (SIG) of the American Educational Research Association (AERA) can develop relationships with principals, state legislators, and state school boards to address problems with teacher evaluation.

On a more local level, PETE faculty can make valuable contributions by nurturing relationships with school district administrators (not just PE teachers) to educate them and enhance awareness of quality physical education and reliable and valid assessments of teacher effectiveness in PE. Further, PE teachers who are more aware and have a deeper understanding of the impact of assessment and evaluation on their careers are better prepared to advocate for themselves and their programs. Programmatically, PETE programs should:

• ensure graduates can self-analyze and discuss additional evaluation resources to gain effective and valuable feedback about their teaching;
• increase instruction on multiple ways to assess students in PE school programs;
• ensure that pre-service teachers develop and apply assessments during their student teaching experience;
• train corps of mentor teachers to assist pre-service teachers in developing and using assessments during their student teaching experience and early field experiences.

PETE scholars who are engaged in research also have a large part to play. Currently, the body of work on issues surrounding teacher evaluation, effectiveness, and accountability in PE is limited. Future studies can enhance our knowledge of principals’ perceptions of PE and encourage them to increase their knowledge of PE. In terms of PE-specific advocacy strategies, the field is nearly barren and the following areas of scholarship would be worthy of attention:

• identification of strategies necessary for improved evaluation of PE teachers by administration and significant others in the workplace;
• analyses of different evaluation systems implemented for PE teachers;
• the effects of this emphasis on PETE programs and PETE students’ assessment;
• exemplar studies of PETE programs’ efforts to increase emphasis on assessment;
• the relationship between teacher use of assessment and principal evaluation.

While the current state of teacher evaluation and accountability presents unique challenges for physical education and those who teach it, strategies presented here may provide a beginning to improving this process. Effective teacher evaluation has the potential to improve teaching practices, student learning, and demonstrate PE teachers’ pedagogical skill.

**Reflective questions for discussion**

1. Discuss the major concerns surrounding teacher assessment.
2. Of the approaches to assessment discussed in this chapter, which would be the most appropriate for PE?
3. What types of assessment information can PE teachers provide to principals to assist in this process?
4. If you taught a workshop for principals to teach them about ‘best practices’ in teacher evaluation in PE, what topics and experiences would you include in the workshop?
5. Once teachers have been assessed how can this information be used to assist teachers to improve their performance and enhance student learning?

References
Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project three-year study (policy and practice brief). Measures of Effective Teaching (MET) project, Bill & Melinda Gates Foundation.
Gabriel, R., & Lester, J. N. (2013). Sentinels guarding the grail: Value-added measurement and the quest for education reform. Education Policy Analysis Archives, 21(9). This article is part of EPAA/AAPE’s Special Issue on Value-Added: What America’s Policymakers Need to Know and Understand, Guest Edited by Dr. Audrey Amrein-Beardsley and Assistant Editors Dr. Clarin Collins, Dr. Sarah Polasky, and Ed Sloat. Retrieved from http://epaa.asu.edu/ojs/article/view/1165.
Teacher accountability


