Research Quarterly for Exercise and Sport, *84*, 407–418, 2013

ISSN 0270-1367 print/ISSN 2168-3824 online
DOI: 10.1080/02701367.2013.844018

Routledge
Taylor & Francis Group

SPECIAL TOPICS: Effective Teaching in Physical Education

Measuring Teacher Effectiveness in Physical Education

Judith E. Rink
University of South Carolina

This article summarizes the research base on teacher effectiveness in physical education from a historical perspective and explores the implications of the recent emphasis on student performance and teacher observation systems to evaluate teachers for physical education. The problems and the potential positive effects of using student performance scores as well as establishing a comprehensive evaluation program are explored with supportive evidence that some level of accountability is necessary in our field to make significant change.

Keywords: accountability, program effectiveness, teacher evaluation

> The move to rethink how to evaluate a teacher's performance and explicitly tie assessments of teacher performance to student achievement marks an important shift in thinking about teacher quality. The demand for 'highly qualified' teachers is slowly but surely being replaced by a call for highly effective teachers. (National Council on Teacher Quality [NCTQ], 2011)

In a 1991 article on good teaching, Donald Cruickshank and Donald Haefele argued that there are many kinds of good teachers—some of them are effective at producing high levels of student performance and others are good for other reasons. Since the publication of that article, the education community has moved steadily toward the notion that good teaching is teaching that results in student achievement. A concern for teacher effectiveness largely follows the national standards and assessment movement designed to hold states, districts, schools, and teachers accountable for student performance on designated outcomes. Standards would define what every student should know and be able to do, curriculums would be designed to be aligned with the standards, and assessment would measure the extent to which students achieved the designated outcomes. Assessment of teacher effectiveness in this process naturally follows.The impetus for much of the reform in teacher evaluation has come as a result of the federal government's grant program to states known as the *Race to the Top Fund. Race to the Top* is part of the 2009 American Recovery and Reinvestment Act granting 11 states $4.35 million to reform education in their states, including a heavy emphasis on student achievement scores and accountability for student achievement (*American Recovery and Reinvestment Act of 2009*, 2009). As a result, states, including those not part of the federal program, have been systematically changing the criteria used to evaluate teachers to include student performance scores as part of required teacher evaluation programs. According to the NCTQ (2011), 30 states now require that teachers are evaluated at least in part on objective evidence of student learning. This represents support for the idea that our education system can be improved if we evaluate teachers on their effectiveness, therefore begging the question, "what is teacher effectiveness and how is it best measured in physical education?"

TEACHER EFFECTIVENESS RESEARCH: A HISTORICAL PERSPECTIVE

The study of teacher effectiveness is not new. Medley (1979) traced the development of conceptions of teacher effectiveness up to that point as: (a) the possessor of

Correspondence should be addressed to Judith E. Rink, 309 Stoney Pointe Circle, Chapin, SC 29036. E-mail: JRink@mailbox.sc.edu

desirable personal traits, (b) the user of effective methods, (c) the creator of a good classroom climate, (d) the master of a repertoire of competencies, and (e) the professional decision maker. Although it is not the purpose of this article to review the research on teacher effectiveness in education or physical education, it is important for the reader to have some perspective on how we have come to this point. The earliest research on teacher effectiveness in the classroom began in the 1940s with a somewhat futile search to link teacher characteristics to student learning. In 1974, Dunkin and Biddle established a model for the study of teaching and identified the constructs of teacher characteristics, student characteristics, process variables (including teacher and student behavior and characteristics), product variables, and the relationship between these constructs as primary targets for research on teaching (Dunkin & Biddle, 1974). The early studies focused on relationships between process variables (process–process studies).

The study of teaching shifted from process–process studies to process–product studies. What is important to note is that from a research perspective, effective teaching has consistently meant student learning. Given the difficulties with measuring student learning and using student learning to evaluate teachers, researchers were hoping to identify a *proxy* for student learning. Most of this search was for process variables that could be observed and had a high relationship to student learning outcomes. Brophy and Good (1984) identified and reviewed the research from the early 1970s to early 1980s. Their work focused on: (a) the opportunity to learn/content; (b) teacher expectations/role definitions/time allocations; (c) classroom management/student engaged time; (d) success level/academic learning time; (e) active instruction by the teacher; (f) group size; (g) presentation of information (structuring, sequencing, clarity, enthusiasm); (h) asking questions (difficulty level, cognitive level, wait time, selecting respondents, providing feedback); and (i) handling seatwork and homework assignments. Brophy and Good (1984) concluded that the decade had been productive to varying degrees in identifying effective teaching but that the quality of the research varied.

Effective teaching research in physical education was largely related to the work being done in the classroom. Each one of the research foci identified by Brophy and Good (1984) has a counterpart in the research on teaching in physical education. The search for the "silver bullet" even in physical education quickly shifted from more indirect teaching characteristics, such as teacher warmth, types of questions, praise, and flexibility, to those more consistent with direct teaching (task-oriented, structured learning experiences, student activity time, active monitoring, and feedback). The product measures of most of the literature in our field focused on motor skill learning.

Necessary but Not Sufficient

Research in physical education and motor learning identified several variables with a strong relationship to student motor skill performance. Among those identified in validation studies to be highly related to student learning outcomes was Academic Learning Time–Physical Education (ALT–PE)-motor engaged (Silverman, Devillier, & Ramírez, 1991), meaning the amount of time students spent in class engaged in motor activities related to the content. Poor management skills decreased ALT–PE. Clarity in task presentations (Werner & Rink, 1989) and the manner in which the teacher develops content (French et al., 1991; Gusthart & Springings, 1989; Masser, 1985; Rink, French, Werner, Lynn, & Mays, 1992) were also investigated and shown to have a relationship with students' motor skill learning. However, none of these variables could be characterized as the "silver bullet" that ensures student learning, in spite of efforts to refine their definitions (e.g., the transition from allocated time to on-task behavior, to motor-engaged time, to good practice). They became part of a body of knowledge that is probably best described as "necessary but not sufficient" conditions for learning. In other words, if you provide maximum practice time, you are not guaranteed learning, but if you do not provide enough practice time, it is likely that learning will not occur. Much of this research is synthesized in the works of Graber (2001), Silverman and Ennis (2003), and Silverman and Skonie (1997).

The difficulty in identifying the concept of effectiveness in teaching lies in the complexity of teaching. Many researchers perceived the process–product paradigm as an oversimplification of a very complex process that is largely a multifaceted interaction between the student, the teacher, the content, and other contextual variables. Attention would turn to the role of the student and content in the search to both understand the teaching/learning process and to be able to identify how best to ensure positive program outcomes. There would also be a shift from process–product studies to qualitative research methodologies (see, e.g., Hemphill, Templin, Richards, & Blankenship, 2012).

More recent research in our field has focused primarily on the student as the mediator of instruction and the processes involved in the dynamics of student motivation (e.g., perceived competence, self-efficacy, expectancy effects, and achievement goals; e.g., A. Chen, Martin, Ennis, & Sun, 2008; Sun & Chen, 2010; Zhang, Solmon, Kosma, Carson, & Gu, 2011). Although some of these mediating variables are beyond the control of the teacher, many are not (e.g., mastery vs. performance climate). The challenge is to link these variables first to teacher behavior and then to outcomes of instruction. With some exceptions, much of the more recent research has sought to understand the teaching process and the role of the student in that process, but researchers have struggled to tie that work directly to learning outcomes (Reeve & Halusic, 2009).

A second more recent focus on research on teaching in physical education has been the study of the effect of different orientations to teaching the content on student outcomes (e.g., sport education, teaching games for understanding, constructivist orientations; W. Chen, Rovegno, Cone, & Cone, 2012; French, Werner, Rink, Taylor, & Hussey, 1996; Penney, Clarke, Quill, & Kinchin, 2005; Sweeting & Rink, 1999). These studies have had mixed results. While the work on the student as the mediator of instruction has sought to understand the role of the student, this work has sought to understand the role of content and how it is delivered.

The work done in the paradigm of process–product studies has become part of the effective teaching literature in physical education and is used extensively to train teachers and observe teaching. The instructional skills identified in the process–product paradigm have become generic instructional skills, meaning they are necessary but not sufficient characteristics of effective teaching for most contexts in physical education where there is a learning objective (Rink, 2013). The work focusing on students and different approaches to teaching the content is just beginning to identify the implications for what the teacher should do. The teacher and what the teacher does are critical to whether the student learns or does not learn (Castelli & Rink, 2003; Gates Foundation, 2013a), which is why the current emphasis on the outcomes of instruction has refocused educators on the teacher and what students learn from the teacher.

PROBLEMS WITH STUDENT PERFORMANCE SCORES AS AN EVALUATIVE MEASURE

The education community has historically resisted using student performance scores to evaluate teachers for a variety of reasons. This is more problematic in programs lacking clearly defined outcomes and when the outcomes are not measureable by standardized tests. Most of the problems surrounding the use of student performance scores in evaluating teaching in physical education are associated with the following issues: a marginalized subject area; no consensus on what students should learn; a culture that does not value assessment; lack of program time and other barriers; unavailability of valid and reliable measures of student performance; diversity of student potential for learning; and the willingness of policymakers to invest resources to develop a valid and reliable evaluation of teacher effectiveness.

Marginalized Subject Area

Physical education has historically been a marginalized subject area in the education system. It is not that good physical education programs are not valued in many schools—they are just not valued to the same extent that "core" subjects are. The physical education profession tends to be saddled with the perceptions of policymakers whose personal experience with physical education was not positive. While the other "noncore" subject areas like art and music have a large political constituency, physical education does not. Health professionals, who could be potential advocates in the age of an "obesity" crisis, have been supportive if not collaborative.

Physical education has for the most part been kept away from the center of attention, influence, or power at both the national and state levels or has not sought to be included. In one sense, this has allowed good programs to become creative and to tailor what they do to the individual needs of their students. The result has not been a healthy neglect. It has protected programs from having to define or measure outcomes in a political environment where data count.

Consensus on What Students Should Learn/Outcomes

One of the major obstacles to evaluating teachers on student performance is defining what students should learn and the outcomes of teaching. Not only have teachers in physical education been given freedom to teach what they consider important and appropriate content for their students, but there is no consensus in the profession on what is important for students to learn (if anything). Although most of the teacher effectiveness research done in our field has made the assumption that motor skills are important learning outcomes, the current literature would suggest that may be a false or incomplete assumption (Metztler, McKenzie, van der Mars, Barrett-Williams, & Ellis, 2013). Although curriculum and learning expectations in the classroom have been more specifically identified, physical education teachers have had only loosely framed curriculums or none at all to direct what they teach.

The national content standards developed for physical education (National Association for Sport and Physical Education [NASPE], 1995, 2004) were a good step in defining the exit outcomes for programs but did little to identify grade-level outcomes until the recent 2013 version. There is consensus that the goal of programs should be the development of a physically active lifestyle. The national standards are designed to develop a physically active lifestyle, but the contribution of each of those standards to this goal is critical and yet to be determined. The work done with PE Metrics (NASPE, 2010, 2011), although not comprehensive of what is taught in most school programs, identified the most critical skills and knowledge to be taught in physical education from each of the standards and provided valid and reliable assessment materials that have the potential to be used as measures of student performance.

A Culture That Does Not Value Assessment

Most educators outside of the physical education field see assessment of student performance (both summative and formative) as an equal partner in the plan–teach–assess process of teaching. Many physical educators see assessment as time spent that can better be used for other purposes. Assessment of outcomes or the effectiveness of instruction is not part of our culture. Physical educators have not had to assess and practitioners largely do not value assessment as part of the teaching–learning process. The lack of consensus on what should be learned, lack of appropriate tools for measuring student learning, and the fact that few schools require teachers in our field to assess learning have all likely played a role in developing this culture.

Lack of Program Time and Other Barriers

One of the barriers to developing a national perspective or accountability for student learning in physical education is the great diversity in the amount of program time devoted to physical education—not only nationally but within the same state. It is difficult to hold teachers accountable for more than minimum expectations for learning when teachers do not have the time needed to teach for those expectations and when we have very little information on how much time it takes for students to become competent in an outcome. Likewise many physical education programs are faced with large classes, isolation, and a lack of administrator support (Ennis, 1992; Mackenzie, 1983).

Valid and Reliable Measures

Assessment materials in most core subjects are developed nationally by commercial companies whose expertise lies in the content and measurement and evaluation. One of the problems in physical education has traditionally been the lack of practical, reliable, and valid measures of program objectives other than fitness. The two volumes of PE Metrics (NASPE, 2010, 2011), one for the elementary level and one for the secondary level, took more than 10 years to complete and began with the identification of performance indicators for each standard. Assessment tasks were designed for a sample of the performance indicators and were not intended to be comprehensive of an indicator or the standard. The PE Metrics material does provide programs with valid and reliable measures. Because physical education has few permanent products, the motor performance assessments require video recording, which to some extent reduces their practicality from a teacher's perspective. The usefulness of the material is also reduced because the material only samples potential outcomes for broad standards, and it is quite likely that assessment materials would not be available for the all the content that programs would define as important outcomes.

Diversity of Student Potential for Learning

Physical educators can expect to have a range of motor skill abilities in their classes. Although most classroom teachers in a heterogeneous grouping are likely to have the same problem, physical educators have struggled to find instructional methodology that would meet the needs of such diverse groups. While most classroom teachers tend to teach to the middle, many physical educators are more likely to teach to the more skilled by moving on in their teaching when important steps in a progression have not been learned.

A major problem associated with using student performance scores in any evaluation of a teacher or a program is the potential of students to demonstrate growth. Students have different potentials for learning what is assessed. Teachers who work with high-ability students may be at a disadvantage on standardized tests simply because high-ability student scores top out, meaning the potential for gain is not there. Students at the other end of the continuum may not show a great deal of improvement because the measure is inappropriate for where they are in the content. Teachers are not in control of the many variables that may affect how a student performs, and this makes the use of absolute standardized test scores a real problem for identifying effective teachers.

Educators have tried to solve the problem of potential for learning by using pretest scores and establishing an "expected" score for a student. The expected score is then compared to the "real" score the student receives, and the difference becomes the residual score given to the teacher. Value-added modeling (VAM) uses student achievement data *over time* (preferably more than 1 year) to measure the learning gains students make (Sanders, 2006). VAM is not without its critics. Educators have attributed differences in teacher scores from year to year to differences in students and student behavior rather than what the teacher has done (Hill & Herlihy, 2011). There is also a big concern that the current research base is insufficient to support the use of value-added scores for high-stakes decisions and applications (McCaffrey, Koretz, Lockwood, & Hamilton, 2004). VAMs are considered to be fairer than simply comparing students' achievement scores or gain scores without considering potentially confounding context. Nevertheless, states and districts across the country forced to consider student learning in their evaluations of teachers have embraced VAM as a way to measure teacher effectiveness, either as the primary evaluation of teaching or as part of a more comprehensive system (NCTQ, 2011).

When a subject area is a nontested content area, meaning that standardized, national assessments are not available, not approved to be used, or not used, most states are using what are commonly referred to as "school scores" to evaluate teachers in these subject areas. School scores are a compilation of the academic scores of student achievement

across the tested content areas. For teachers who teach in nontested subject areas like physical education, this means that they are evaluated using student performance in content areas they do not teach.

POTENTIAL POSITIVE EFFECTS OF MEASURING STUDENT OUTCOMES

Many of the problems and issues associated with using student performance outcomes to evaluate physical education programs and/or teachers have been discussed in this article. In spite of the problems associated with the use of student performance as part of a teacher evaluation system, there may be merit in using student scores at least as part of a teacher evaluation system. This section focuses on the potential positive aspects of making the decision to do so.

Shared Vision

One of the positive effects of the standards and assessment movement is the potential to create a shared vision of what students should know and be able to do. Creating standards and holding educators accountable for outcomes related to those standards also has the potential for creating a national dialogue about what those outcomes might look like once developed in the school program. Since the original content standards for physical education were published in 1995 (NASPE), most states have either adopted or adapted the national standards, which can be considered at least some level of support for the notion of a shared vision. Unlike the academic subject areas in the school curriculum that rely largely on the selection of textbooks to define the content and its progression through the program, physical education has for the most part assumed that teachers in a school district can define the content of a good physical education program and develop it throughout the school program independently. This is particularly problematic when districts do not have supervision with content expertise in our field and has resulted in varied interpretations of the standards and less-than-effective programs. Although much of our literature in the past has assumed that teachers in the field do not support defining student outcomes or curriculum, there is little evidence that this is true (Fleming, 1998; Rink, Jones, Kirby, Mitchell, & Doutis, 2007).

The standards and assessment movement has always assumed that if you designate what should be taught and hold educators responsible for the content, they should be free to develop that content appropriately for their context. Standards were designed to describe minimal expectations for students rather than be inclusive of what a good program should teach. "Indicators" of the standards must be described when assessment material is designed to measure the standards. This narrows the inclusiveness of the assessment and therefore increases the potential for narrowing the curriculum (teaching to the test) but increases the probability that those minimal expectations will be achieved by students.

Advocacy

One of the unintended consequences of high-stakes assessment in the schools is the effect it has had on outcomes and content areas not part of the core academic subjects. What is not measured does not count. What does not count does not receive the support and resources needed and may even be eliminated from the school program. One way to maintain resources and get support, particularly for marginalized program areas, is to become part of the reform movement. Currently, that means clearly identifying outcomes, supplying policymakers with data on student achievement, and looking carefully at how we evaluate programs and teachers.

Some level of accountability for program and/or teacher effectiveness can help prevent program erosion and act as a very powerful advocacy tool. Teachers, programs, and schools do not want to be perceived as low-performing in any category or content area of a school program. They will do what they need to do to improve their status whether or not there are high-stakes consequences for not being good. For physical education, this means the potential for more program time, more teachers, more equipment, support for professional development, competent leadership, and some level of accountability for teachers.

Teacher Development

When teachers and schools are evaluated on student performance, the focus is on teaching for learning. This is true in the academic areas as well as physical education (Werner & Rink, 1989). One of the positive effects of collecting student performance data is that it motivates teachers to seek help in how to better facilitate student learning through professional development experiences and encourages districts to provide that support.

Accountability

Setting high standards for achievement has the potential to inspire greater effort (American Educational Research Association, 2010). Academic programs have traditionally had some level of accountability for student outcomes. Parents want their children to be able to read and do math. Schools want students to do well on national tests. On the other hand, physical education programs have had little accountability for student outcomes. Parents and many policymakers and administrators tend to be uninformed about the objectives of the program, other than opportunities to develop fitness. Program assessment is largely nonexistent. Lack of accountability has allowed programs to be free

from the negative effects of top-down supervision. More often than not, lack of accountability has reflected a status as an unimportant subject area, has protected poor teaching and poor programs, and has inhibited the incentive to do better.

THE SOUTH CAROLINA EXPERIENCE

One of the more comprehensive efforts to improve physical education programs through standards, assessment, and accountability began in South Carolina in 1994 with the publication of state standards to be followed by the development of assessment materials for those standards and the first mandated statewide data collection on student performance in 2000. The process involved building consensus, developing materials, establishing state policy, and extensive teacher development. The beginning stages of the program involved more than 100 professionals from all school levels and representatives from almost all of the teacher-training institutions in the state doing much to create that shared vision.

Five years of high school student performance data and 1 year of elementary and middle school data were collected, and 1 year of elementary and middle school data was collected before the economic recession forced the state to put on hold all assessment other than that federally mandated. The South Carolina experience can inform the discussion on collecting student performance data to evaluate teachers and/or programs. The process, program, and results of the 1st-year data collection appear as *a Journal of Teaching in Physical Education* monograph (Rink & Mitchell, 2003).

Teacher evaluation systems seek to measure student growth. This means you need to have some kind of premeasure of student learning. The decision was made by physical educators in the state to focus on program assessment and not to ask teachers already reluctant to give up class time for assessment to collect data twice. The South Carolina Physical Education Assessment Program (SCPEAP) was designed to collect data for a school every 3 years. The program did not collect pretest data and it sampled classes for each teacher, therefore making it far less than a comprehensive assessment program and subject to all the disadvantages of using student performance scores discussed earlier in this article. In spite of these weaknesses, the potential for positive change has been documented (Rink & Stewart, 2003).

Program Assessment or Teacher Evaluation?

One of the major decisions facing SCPEAP was how the data would be reported and for what purpose. The purpose of the program was always to create a shared vision of what good programs can and should be doing and to develop some level of accountability for doing so. Assessment would be accompanied by extensive teacher development programs. The program was designed to collect student performance data as an indicator of program effectiveness. Early on in meetings with high school teachers, it became clear that teachers wanted data reported by teacher as well. Good teachers wanted credit for their teaching and did not want their data combined with those in their departments who they felt were not doing a good job. Ultimately, reports were designed to include state-level data, school-level data, and teacher data for each of the four performance indicators. School reports were sent to teachers, principals, superintendents, and the South Carolina Department of Education.

Teacher effects in classroom literature have been reported to be large (Nye, Konstantopoulus, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Differences between teachers even in the same school teaching the same performance indicator in the SCPEAP data were also very large (Mitchell, Castelli, & Strainer, 2003). The effect of reporting teacher data to administrators was to increase teacher incentive to do well and to have the administrators question differences in student performance between teachers, which in itself acted as a mechanism for accountability.

Narrowing the Curriculum

Standardized assessment material was not available at the onset of SCPEAP. NASPE assessment materials were not published until 10 years after the program's onset. Performance indicators and assessment tasks would have to be developed to collect student performance data on the standards. The goal was to identify minimal expectations for student learning that could be achieved by all students with effective instruction. One of the major criticisms of standardized testing is that it narrows the curriculum to what is tested. Given the poor quality of many programs throughout the state, the assumption was made that if the assessments were good and representative of a good program, then the idea of assessment driving the curriculum was a good thing. Teachers will do whatever they need to do to ensure that their students will do well on the test (Linn, 2000). However, if the assessments are narrow representations of what should be a broad curriculum and assessment drives the curriculum, it is a bad thing. Grade-level task forces were established to work on both the development of the performance indicators and assessment tasks. College and university faculty across the state played a major role in developing the assessments and piloting all of the materials. Four high school indicators were developed as exit criteria for the high school program as follows:

- *Performance Indicator 1*: demonstrate competency in at least two movement forms.

- *Performance Indicator 2*: design and develop an appropriate physical fitness program to achieve a desired level of personal fitness.
- *Performance Indicator 3*: participate regularly in health-enhancing physical activity outside the physical education class.
- *Performance Indicator 4*: meet the gender and age-group health-related physical fitness standard as published by NASPE.

To prevent a narrowing of the curriculum, assessment rubrics for 23 different activities were developed at the high school level with the option for programs to submit others. The indicators and assessment materials were designed to give programs flexibility in the content.

Elementary and middle school materials were developed several years after the high school materials. The elementary level was the most problematic in terms of the implications for curriculum. At the lower grade levels, there are many essential fundamental motor skills and all could not be assessed without making the assessment impractical. Ultimately, the most critical skills in all three curricular areas—educational dance, gymnastics, and games—were determined, and assessment material was developed for those skills. The potential for narrowing the curriculum at the elementary level was high but would encourage the development of fundamental skills considered essential for the development of a physically active lifestyle (Stodden et al., 2008). The development of the middle school data was less problematic in that students at this level can combine skills in activities in one assessment task. Indicators would have to be designed so that programs had a choice for activities to teach, but the assessment tasks assessed students at the same level between activities.

Each of the performance indicators represented a significant curriculum change for most high school programs dominated by team sports and elementary programs that did not include one or more of the curriculum areas being assessed. In spite of the short duration of data collection, change in high school curricula as a result of the assessment program was documented (Fleming, 1998; Pebworth, 2006; Stanne, 1999; Wirszyla, 1998). This supports the positive effect that assessment of student performance accompanied by extensive teacher development can have on curriculum change. The standards and assessment materials for the state were used "voluntarily" before the assessment program was mandated. A study of the high- and low-performing high schools indicates that low-performing schools did not begin to make changes until the assessment was mandated, providing evidence that policy that creates some kind of accountability for change may be necessary if large-scale change is to be produced (Castelli & Rink, 2003).

Change in Student Learning

The goal of the assessment program was to increase student learning in the state standards. The 1st year of data collection at a school would act as a baseline to document change. Because schools were evaluated every 3 years, only the high school data were available to document change in student performance from one data collection to the next. Changes in school data and teacher data from the first time data were collected for a school to the second time indicated a significant change in student competency in three of the performance indicators: competency in at least two movement forms, the ability of students to design and develop an appropriate physical fitness program to achieve a desired level of personal fitness (Stewart & Mitchell, 2003), and student fitness levels. Performance Indicator 3—participate regularly in health-enhancing physical activity outside the physical education class—was high at the first data collection (Hall, French, Webster, Harvey, & Crollick, 2009; Heidorn, 2007).

Advocacy

Research done on teacher perceptions of the program at the high school level supports the idea that assessment can act as an advocate for programs. Before the mandated data collection, teachers saw the efforts to establish indicators and assessment as something being done "for them" and not "to them" (Fleming, 1998). Likewise, teachers who had done the assessment saw the assessment program as support for their programs within the school (Castelli & Rink, 2003). Many programs were able to get reduced class sizes and equipment they needed and were able to obtain scheduling concessions from guidance counselors. When physical education programs have expectations for what students should learn, the programs are put on more equal footing with other content areas within the school.

Level of Accountability

The South Carolina experience with state-level assessment was a positive one and would support the idea that high-stakes assessment is not needed to produce change but some level of accountability is needed. Reports were given to administrators. Some administrators created a dialogue with teachers on the content of the reports and some did not. The South Carolina Department of Education received reports for all the schools but applied no accountability for low-performing schools. In this experience, the idea that data are collected and reported to someone outside the gym was sufficient to create a level of accountability for change.

MULTIPLE MEASURES FOR TEACHER EVALUATION

What is clear in the literature is that most professionals looking at the evaluation of teachers support a multidimensional

model (Kane & Staiger, 2012). Although there is a clear trend toward using some kind of objective measure of student learning (30 states; NCTQ, 2011) and VAM has the best predictive value for identifying effectiveness (Kane & Staiger, 2012), few states have made student performance on standardized tests the only measure of teacher effectiveness or criterion for teacher evaluation.

The Measures of Effective Teaching (MET) project (Gates Foundation, 2013a) was a 3-year study designed to determine how to best identify and promote great teaching and concluded that student achievement gains and teacher observation together have the best predictive value (Gates Foundation, 2013b). Support for using each measure is based upon differences in the predictive power, reliability, and diagnostic usefulness of the measure.

An approach to identifying effective teachers and evaluating teachers using multiple measures is less likely to produce fluctuating ratings from 1 year to the next and is more likely to identify teachers who produce better outcomes. Although measuring the products of learning is critical for accountability, observation of teaching has become an essential component of such systems and is critical for helping teachers improve what they do so they can become effective. The problem with many existing observation tools used by many school districts and states is that the tools do not discriminate between effective and ineffective teachers. Observation instruments need to be comprehensive enough to capture a robust vision of effective teaching without becoming so extensive that they become unmanageable for observers. Many states and districts have begun the process of developing teacher observation tools that are generic and can be used across content areas (NCQT, 2011). The difficulty has been in designing observation systems that discriminate between effective and ineffective teaching across content areas and designing a system for the use of those tools that is a valid representation of a teacher's work.

The observation tools developed and/or tested by the Gates Foundation (2013a) for the classroom include both content-specific and generic approaches and clearly define the behaviors expected at multiple levels. The work of the Gates Foundation is not unlike the process–product studies conducted in education and physical education referred to at the beginning of this article. The intent in the design of the observation tool is to demonstrate a predictive value with student outcomes. The tools were shown to have a high relationship with student achievement, which means that the teachers who scored high on the behaviors designated on the observation tool produced a high level of student achievement. When used with other measures of teacher effectiveness, high levels of predictability existed.

An outline of the content used in the Framework for Teaching Evaluation Instrument (FFT; Danielson Group, 2013) is provided in Table 1 as an example of a generic tool found to be a reliable tool related to student learning in the academic areas. This is one of the most comprehensive tools, and many of the variables articulated in the tool are components of teacher observation tools now being used across the country. The tool itself provides indicators for the four domains (dimensions; only two of which are direct observation with students) and components (subcategories) with very specific descriptions, examples of each, and a comprehensive scoring rubric. The domains and components of the FFT instrument are certainly appropriate for evaluating physical education lessons. The problem is in the descriptors.

What is good grouping and management in the classroom is not necessarily good grouping and management in physical education. Teacher questioning is a critical skill used in the classroom to develop student understanding of the content. Rink (1979) argued that the critical unit for content development in physical education is the teacher movement task and the student movement response to that task. To use the FFT with any validity in physical education, the descriptors and examples would need to be changed and observers would need to be trained to discriminate the behaviors. The Physical Education and Lesson Observation Tool developed in Singapore is an example of a tool that attempted to do so. An example from that tool for management is presented in (Table 2).

NASPE (2007) developed a teacher observation tool that identifies several common constructs used in the evaluation of teaching in physical education including: instructional variables, evidence of student learning, management, class climate, and professionalism. Many of the components of the constructs are very highly inferred. For the tool be used effectively and discriminate effective teaching, each of the components of a construct would have to be defined much more clearly with specific examples. Rubrics with multiple levels of performance would also have to be designed. In physical education, it is likely that different curricula will emphasize different instructional arrangements and teaching behaviors. For example, it is likely that the evaluation criteria for a sport education lesson or a lesson in teaching games for understanding will have some characteristics of what is considered good instruction that are different from a lesson that utilizes direct instruction more exclusively. Professionals who argue that moderate-to-vigorous physical activity (MVPA) should receive a great deal of emphasis would certainly want to include that variable, although to use MVPA or any single variable exclusively as a measure of teacher effectiveness is not advisable unless it is the only outcome desired. In other words, aside from the generic variables identified in the FFT and most comprehensive tools, the objectives of a program would need to be considered in teacher observation tools.

Problems With Teacher Observation as a Measure of Effectiveness

One of the major problems with teacher observation tools is that they are designed primarily to observe the behavior of

TABLE 1
The Framework for Teaching Evaluation Instrument, 2013

DOMAIN 1: PLANNING AND PREPARATION
1A: Knowledge of Content and the Structure of the Discipline
- Knowledge of content and the structure of the discipline
- Knowledge of prerequisite relationships
- Knowledge of content-related pedagogy

1B: Knowledge of Students
- Knowledge of child and adolescent development
- Knowledge of the learning process
- Knowledge of students' skills, knowledge, and language proficiency
- Knowledge of students' interests and cultural heritage
- Knowledge of students' special needs

1C: Setting Instructional Outcomes
- Value, sequence, and alignment: Outcomes represent significant learning
- Clarity: Outcomes refer to what students will learn, not what they will do
- Balance: Outcomes reflect different types of learning, such as knowledge, conceptual understanding, and thinking skills
- Suitability for diverse students

1D: Demonstrating Knowledge of Resources
- Resources for classroom use align with learning outcomes.
- Resources to extend content knowledge and pedagogy of teacher
- Resources for students must be appropriately challenging.

1E: Designing Coherent Instruction
- Learning activities designed to engage students and advance them through the content
- Instructional materials and resources appropriate to the learning needs of the students
- Instructional groups to support student learning
- Lesson and unit structure to produce clear and sequenced lesson and unit structures to advance student learning

1F: Designing Student Assessments
- Congruence with instructional outcomes
- Criteria and standards must be clearly defined.
- Design of formative assessments as part of the instructional process
- Results of assessment guide future planning.

DOMAIN 2: THE CLASSROOM ENVIRONMENT
2A: Creating an Environment of Respect and Rapport
- Teacher interactions with students, including both words and actions
- Student interactions with other students

2B: Establishing a Culture for Learning
- Importance of the content and of learning
- Expectations for learning and achievement
- Student pride in work

2C: Managing Classroom Procedures
- Management of instructional groups
- Management of transitions
- Management of materials and supplies
- Performance of classroom routines

2D: Managing Student Behavior
- Expectations
- Monitoring of student behavior
- Response to student misbehavior

2E: Organizing Physical Space
- Safety and accessibility
- Arrangement of furniture and use of physical resources

DOMAIN 3: INSTRUCTION
3A: Communicating With Students
- Expectations for learning: goals communicated
- Directions for activities
- Explanations of content
- Use of oral and written language

(*continued*)

TABLE 1 – (*Continued*)

3B: Using Questioning and Discussion Techniques
- Quality of questions
- Discussion techniques
- Student participation

3C: Engaging Students in Learning
- Activities and assignments
- Grouping of students
- Instructional materials
- Structure and pacing

3D: Using Assessment in Instruction
- Assessment criteria
- Monitoring of student learning
- Feedback to students
- Student self-assessment and monitoring of progress

3E: Demonstrating Flexibility and Responsiveness
- Lesson adjustment
- Response to students
- Persistence

DOMAIN 4: PROFESSIONAL RESPONSIBILITY
4A: Reflecting on Teaching
- Accuracy
- Use in future teaching

4B: Maintaining Accurate Records
- Student completion of assignments
- Student progress in learning
- Noninstructional records

4C: Communicating With Families
- Information about the instructional program
- Information about individual students
- Engagement of families in the instructional program

4D: Participating in the Professional Community
- Relationships with colleagues
- Involvement in a culture of professional inquiry
- Service to the school
- Participation in school and district projects

4E: Growing and Developing Professionally
- Enhancement of content knowledge and pedagogical skill
- Receptivity to feedback from colleagues
- Service to the profession

4F: Showing Professionalism
- Integrity and ethical conduct
- Service to students
- Advocacy
- Decision making
- Compliance with school and district regulations

Notes. These are the subcategories of the instrument. In a few cases, they have been shortened for space purposes. They are described with examples of each and a comprehensive scoring rubric in the original instrument. Adapted with permission from *The Framework for Teaching Evaluation Instrument (FFT)*, by Danielson Group, 2013. Copyright 2013 by Danielson Group.

the teacher. The instructional process is an interactive process. The appropriateness of teacher behavior depends on what the students are doing and the appropriateness of that behavior to the content. Task presentations are effective when students do what the teacher has asked and when the information teachers give students is accurate. Feedback is effective when students can use that information to change

TABLE 2
Singapore Physical Education Lesson Observation Tool (PELOT)—Management Indicators for Physical Education

Lesson is managed to promote learning
Students are actively involved in tasks, with no delay in their participation when they report for lessons.
Routines and behavioral expectations are established and adhered to by students.
Students are physically safe.
The organization of space, equipment, and students supports instruction and allows for maximum practice.
Appropriate and sufficient equipment is used to maximize learning.
Task transitions promote maximum practice.
A positive learning environment is established.
On-task behavior is maintained.

Note. Adapted with permission from *Physical Education Lesson Observation Tool*, by the Ministry of Education, Singapore, 2013. Copyright 2013 by the Physical Education & Sports Teacher Academy, Ministry of Education, Singapore.

what they do. Progressions are effective when students experience success. Management is effective when students have high levels of quality practice time. Observing teaching without also observing the effect of teacher behavior on what the students do is problematic. Tools designed to look at what the teacher does exclusively are actually measuring many of the instructional variables identified earlier as "necessary but not sufficient" for effectiveness. They are likely to identify teachers with less-than-adequate instructional skills, but teachers who score high on the instruments may not necessarily produce the desired outcomes of that instruction.

Observing the effectiveness and appropriateness of what the teacher does requires observers who know the content area. Physical education is at a disadvantage in measuring products because we have no permanent products of the process of motor skills unless they are video-recorded. Physical education is at an advantage in that student behavior with the content is very observable. The best observation systems would consider both teacher behavior and student behavior in the context of the content. When the context and appropriateness of teacher behavior are considered, tools become more highly inferred and dependent on the competence of the observer. School districts without a supervisor who is a specialist in physical education are likely not to invest the time in developing instruments for our field, which may mean that tools with a manageable set of competencies fully developed for different performance levels may need to be more fully developed at the state or national level.

For a tool to be a reliable indicator of teacher effectiveness, three to four observations a year must be conducted (Gates Foundation, 2013b). The recommendations of NASPE for the evaluation of physical education teachers are consistent with the education literature (NASPE, 2007). Teachers should:

- be evaluated with standards, expectations, procedures, and rigor that parallel teachers of other curricular areas;
- be observed, assessed, and evaluated by trained evaluators;
- be observed multiple times during the academic year;
- be observed for the entire class period, from beginning to end;
- be observed and evaluated as part of a comprehensive assessment plan, which should include formal conferences, professional growth plans, etc.; and
- be accountable for student achievement of state standards in physical education or the NASPE (2004) in the absence of state standards (NASPE 2007).

CONCLUSION

Establishing a valid and reliable system to evaluate the effectiveness of physical education teachers will require resources beyond the principal making an occasional visit into the gym to observe a teacher using generic criteria with definitions more suited to classroom observations. If VAM is used, there will need to be a great deal of record-keeping, standardized assessments, and training in how to administer the assessments. Policymakers are not likely to be willing to make such an investment for a profession that does not value or work to make it so or without top-level policy that requires it (Kirby, 2005).

If physical education is to be a supported school program, physical educators must be willing to define their program outcomes and ways to measure those outcomes. They must also be willing to hold programs and practitioners accountable for effective teaching. Although having a shared vision of what those outcomes should be would do much to support advocacy for the profession, it is more important for programs to align outcomes, teacher evaluations, and student assessment with the designated outcomes of the program. It is unacceptable for students to be evaluated on content they have not had an opportunity to learn or for physical education teachers to be evaluated on content they are not expected to teach. The profession is best served by the development of both student assessment and teacher evaluation materials that match outcomes.

The current emphasis on measuring teacher effectiveness will impact physical education either positively or negatively depending on the extent to which physical educators become participants in the reform movement. Designing and conducting a quality teacher or program evaluation for physical education has the potential to be a significant impetus for change in our field, in spite of the identified measurement weaknesses in available tools and

processes. Quality evaluation programs should minimally include measures of student performance, and if possible, growth as well as observations of teaching. Evaluating teachers provides an incentive for change and a foundation for quality teacher development both at the preservice and in-service levels.

WHAT DOES THIS ARTICLE ADD?

This article synthesizes the current work being done in measuring teacher effectiveness and emphasizes the importance of developing and establishing valid and reliable tools and processes of evaluation for the field of physical education. Assessment of teacher effectiveness has the potential to improve the practice of physical education through the development of clear outcomes, student and teacher assessment tied to those outcomes, and accountability for the development of those outcomes.

REFERENCES

American Educational Research Association. (2010). *Position statement on high stakes assessment in pre-k–12 education*. Retrieved from http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx

American Recovery and Reinvestment Act of 2009 (ARRA). Pub. L. No. 111–5, § 14005-6, Title XIV.

Brophy, J., & Good, T. (1984). *Occasional paper #73*. East Lansing: Michigan State University, Institute for Research on Teaching.

Castelli, D., & Rink, J. (2003). A comparison of high and low performing secondary physical education programs. *Journal of Teaching in Physical Education*, *22*, 512–532.

Chen, A., Martin, R., Ennis, C. D., & Sun, H. (2008). Content specificity of expectancy beliefs and task values in elementary physical education. *Research Quarterly for Exercise and Sport*, *79*, 195–208.

Chen, W., Rovegno, I., Cone, T. P., & Cone, S. L. (2012). An accomplished teacher's use of scaffolding during a second grade unit on designing games. *Research Quarterly for Exercise and Sport*, *83*, 221–234.

Cruikshank, D., & Haefele, D. (1991). Good teachers, plural. *Educational Leadership*, *58*(5), 26–30.

Danielson Group. (2013). *The Framework for Teaching Evaluation Instrument (FFT)*. Princeton, NJ: Author.

Dunkin, M., & Biddle, B. (1974). *The study of teaching*. New York, NY: Holt, Rhinehart and Winston.

Ennis, C. (1992). Developing physical education curriculum based on learning goals. *Journal of Physical Education, Recreation and Dance*, *63* (7), 74–77.

Fleming, D. (1998). *The impact of state-mandated change and a systemic inservice training project on high school physical education curriculum* (Unpublished doctoral dissertation). Columbia: University of South Carolina.

French, K., Rink, J., Rickard, L., Mays, A., Lynn, S., & Werner, P. (1991). The effects of practice progressions on learning two volleyball skills. *Journal of Teaching in Physical Education*, *10*, 261–274.

French, K., Werner, P., Rink, J., Taylor, K., & Hussey, K. (1996). The effects of a 3-week unit of tactical, skill, or combined tactical and skill instruction on badminton performance of ninth-grade students. *Journal of Teaching in Physical Education*, *15*, 418–438.

Gates Foundation. (2013a). *Measures of Effective Teaching project* (Final research report). Retrieved from http://www.metproject.org/reports.php

Gates Foundation. (2013b). *MET report: Teacher observation less reliable than test scores*. Retrieved from http://www.metproject.org/reports.php

Graber, K. (2001). Research on teaching physical education. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed.) (pp. 491–519). Washington, DC: American Educational Research Association.

Gusthart, J., & Springings, E. (1989). Student learning as a measure of teaching effectiveness. *Journal of Teaching in Physical Education*, *8*, 298–311.

Hall, T. H., French, K. E., Webster, C. A., Harvey, R. L., & Crollick, J. (2009). South Carolina secondary physical education programs: Improvement across three years [Abstract]. *Research Quarterly for Exercise and Sport*, *80*(Suppl. 1), A–58.

Heidorn, B. (2007). *The effectiveness of an outside of school physical activity requirement for high school students* (Unpublished doctoral dissertation). University of South Carolina, Columbia.

Hemphill, M., Templin, T., Richards, K., & Blankenship, B. (2012). A content analysis of qualitative research in the *Journal of Teaching in Physical Education* from 1998 to 2008: Part one. *Journal of Teaching in Physical Education*, *31*, 279–281.

Hill, H., & Herlihy, C. (2011). *Prioritizing teaching quality in a new system of teacher evaluation*. Washington, DC: American Enterprise Institute Policy Studies. Retrieved from http://www.aei.org/policy/education/k-12

Kane, T. J., & Staiger, D. D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achivement gains*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://metroproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kirby, K. (2005). *High school principal's perceptions and support for a state physical education assessment program* (Unpublished doctoral dissertation). University of South Carolina, Columbia.

Linn, R. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16.

Mackenzie, D. (1983). Research for school improvement: An appraisal of some recent trends. *Educational Researcher*, *12*(4), 5–17.

Masser, L. (1985). The effect of refinement on student achievement in a fundamental motor skill in Grades K through 6. *Journal of Teaching in Physical Education*, *6*, 174–182.

McCaffrey, D., Koretz, D., Lockwood, J., & Hamilton, L. (2004). *Evaluating value-added models for teacher accountability*. New York, NY: Rand Corporation. Retrieved from http://www.rand.org/pubs/monographs/MG158.html

Medley, D. (1979). *Teacher competence and teacher effectiveness: A review of process product research*. New York, NY: American Association of Colleges for Teacher Education, Committee on Performance-Based Teacher Education.

Metztler, M., McKenzie, T., van der Mars, H., Barrett-Williams, S., & Ellis, R. (2013). Health Optimizing Physical Education (HOPE): A new curriculum for school programs—Part 2: Teacher knowledge and collaboration. *Journal of Physical Education, Recreation and Dance*, *84* (5), 25–56.

Mitchell, M., Castelli, D., & Strainer, S. (2003). Student performance data, school attributes and relationships. *Journal of Teaching in Physical Education*, *22*, 494–511.

National Association for Sport and Physical Education. (1995). *Moving into the future: National standards for physical education*. Reston, VA: Author.

National Association for Sport and Physical Education. (2004). *Moving into the future: National standards for physical education* (2nd ed.). Reston, VA: Author.

National Association for Sport and Physical Education. (2007). *Physical Education Tool*. Reston, VA: Author. Retrieved from http://www.aahperd.org/naspe/publications/TeachingTools/observepe.cfm

National Association for Sport and Physical Education. (2010). *PE Metrics™: Assessing National Standards 1–6 in elementary school*. Reston, VA: AAHPERD.

National Association for Sport and Physical Education. (2011). *PE Metrics™: Assessing National Standards 1–6 in secondary school*. Reston, VA: AAHPERD.

National Council on Teacher Quality. (2011). *CTQ State Teacher Policy Yearbook Brief Area 3: Identifying effective teachers*. Retrieved from http://www.nctq.org/reports.do?d=State+Policy&searchTerm=Identifying+Effective+Teachers

Nye, B., Konstantopoulus, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237–257.

Pebworth, K. (2006). *High school physical education curriculum in the state of South Carolina* (Unpublished doctoral dissertation). University of South Carolina, Columbia.

Penney, D., Clarke, G., Quill, M., & Kinchin, G. (2005). *Sport education in physical education: Research based practice*. Philadelphia, PA: Routledge.

Reeve, J., & Halusic, M. (2009). How K–12 teachers can put self-determination theory principles into practice. *Theory and Research in Education*, *7*, 145–154.

Rink, J. (1979). *Development of an observation system for content development in physical education* (Unpublished doctoral dissertation). The Ohio State University, Columbus.

Rink, J. (2013). *Teaching physical education for learning* (7th ed.). New York, NY: McGraw-Hill.

Rink, J., French, K., Werner, P., Lynn, S., & Mays, A. (1992). The influence of content development on the effectiveness of instruction. *Journal of Teaching in Physical Education*, *11*, 139–149.

Rink, J., Jones, L., Kirby, K., Mitchell, M., & Doutis, P. (2007). Teacher perceptions of a physical education statewide assessment program. *Research Quarterly for Exercise and Sport*, *78*, 204–215.

Rink, J., & Mitchell, M. (Eds.) (2003). State level assessment in physical education: The South Carolina experience [Monograph]. *Journal of Teaching in Physical Education*, *22*.

Rink, J., & Stewart, S. (2003). Insights and reflections on a state assessment program. *Journal of Teaching in Physical Education*, *22*, 573–588.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417–458.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*, 247–252.

Sanders, W. L. (2006, October). *Comparisons among various educational assessment value-added models*. Paper presented at The Power of Two—National Value-Added Conference, Columbus, OH. Retrieved from http://www.sas.com/resources/asset/vaconferencepaper.pdf

Silverman, S., Devillier, R., & Ramírez, T. (1991). The validity of Academic Learning Time–Physical Education (ALT–PE) as a process measure of student achievement. *Research Quarterly for Exercise and Sport*, *62*, 319–325.

Silverman, S. & Ennis, C. (Eds.). (2003). *Student learning in physical education: Applying research to enhance instruction*. Champaign, IL: Human Kinetics.

Silverman, S., & Skonie, R. (1997). Research on teaching physical education: Analysis of published research. *Journal of Teaching in Physical Education*, *16*, 300–311.

Stanne, K. (1999). *The effect of a varied and choice curriculum on the participation, perceptions and attitudes of females in physical education* (Unpublished doctoral dissertation). University of South Carolina, Columbia.

Stewart, S., & Mitchell, M. (2003). Instructional variables and high school students' knowledge and conceptions of health related fitness. *Journal of Teaching in Physical Education*, *22*, 533–551.

Stodden, D., Goodway, J., Langendorfer, S., Roberton, M., Rudisell, M., Garcia, L., & Garcia, E. (2008). A developmental perspective on the role of motor skill competence in physical activity: An emergent relationship. *Quest*, *60*, 290–306.

Sun, H., & Chen, A. (2010). An examination of sixth graders' self-determined motivation and learning in physical education. *Journal of Teaching in Physical Education*, *29*, 262–277.

Sweeting, T., & Rink, J. (1999). Effects of direct instruction and environmentally designed instruction on the process and product characteristics of a fundamental skill. *Journal of Teaching in Physical Education*, *18*, 216–233.

Werner, P., & Rink, J. (1989). Case studies of teacher effectiveness in physical education. *Journal of Teaching in Physical Education*, *4*, 280–297.

Wirszyla, J. (1998). *Case studies of state-mandated curriculum change in three high school physical education programs* (Unpublished doctoral dissertation). University of South Carolina, Columbia.

Zhang, T., Solmon, M. A., Kosma, M., Carson, R. L., & Gu, X. (2011). Need support, need satisfaction, intrinsic motivation, and physical activity participation among middle school students. *Journal of Teaching in Physical Education*, *30*, 51–68.