The Physical Educator Vol. 75 • pp. 582–616 • 2018
https://doi.org/10.18666/TPE-2018-V75-I4-8471

ASSESSMENT

Physical Education Meets Teacher Evaluation: Supporting Physical Educators in Formal Assessment of Student Learning Outcomes

Hans van der Mars, Jeff McNamee, Gay Timken

Abstract

Few physical educators engage in formal assessment of student learning outcomes. Recent trends in high-stakes teacher evaluation make formal assessment of learning progress/outcomes by physical educators a teaching function, central to a "well-rounded education" through recent federal legislation. This project targeted the following research questions: (a) Can physical educators increase their use of formal-formative assessment of student learning outcomes as a consequence of a professional development intervention that includes on-site coaching support? and (b) Can physical educators reliably collect formal-formative assessment data on student learning? Seven (4 females, 3 males; experience range: 4–15 years) licensed secondary school physical education teachers volunteered as participants. The intervention included professional development (PD) workshops, ongoing on-site coaching support, and prompting. A multiple baseline design (MBD) across subjects determined the efficacy of the intervention on teachers' use of formal-formative

Hans van der Mars is a professor and the Physical Education Program director, Mary Lou Fulton Teachers College, Arizona State University. Jeff McNamee is chair and professor of Human Performance, Department of Health, Human Performance, and Athletics, Linfield College. Gay Timken is a professor, Division Health and Exercise Science, Western Oregon University. Please send author correspondence to hans.vandermars@asu.edu

Acknowledgment: This research was funded in part by SHAPE America.

assessment practices. Postcheck follow-up observations were made during the following school year. Data were collected via the validated Systematic Observation of Formal Assessment of Students by Teachers (SOFAST). The reliability of teachers' use of formal assessment was established with a second trained observer simultaneously collecting identical data on the same aspects of student performance as the teacher. Data were plotted graphically across sessions and assessed with standard visual analysis criteria, which demonstrated the functional relationship between the intervention and teachers' use of formal-formative assessment of student outcomes. Standard interobserver agreement calculations determined the reliability of the teachers' assessments of students' MVPA levels versus outside observers' via the total interval method. All seven teachers increased their use of formal assessment of student performance and shifted focus of the assessment toward subject-matter performance. Their formal assessment observations were found to be reliable. Experienced secondary school physical education teachers can successfully and reliably employ formal-formative assessment of student performance outcomes.

Assessment of student learning outcomes is a central teaching function (e.g., NASPE Assessment Task Force, 2008; National Association for Sport and Physical Education [NASPE], 2002; Society of Health and Physical Educators [SHAPE America], 2010a, 2011; Steffen & Grosse, 2003; Stork, 2007) and a standard for beginning and advanced physical educators (SHAPE America, 2010b; National Board for Professional Teaching Standards, 2014). Yet Lund and Veal (2008) noted that school physical education programs lack a culture of assessment. This is problematic in today's climate of high-stakes teacher evaluation practices where teachers' job security is, in part, tied directly to their students' achievement. Despite the demonstrated problem of value-added models (i.e., lack of fairness, reliability and validity; e.g., American Educational Research Association, 2015; American Statistical Association, 2014; Amrein-Beardsley, 2008, 2012; Berliner, 2013, 2014; Lavigne, 2014; Pivovarova, Broatch, & Amrein-Beardsley, 2014), 43 states now require objective measures of student achievement to be included in teacher evaluations, and student growth is the preponderant criterion in teacher evaluations in 16 states (National Council on Teacher Quality, 2015). Moreover, school administrators feel ill-prepared to make use of such teacher

evaluation protocols in physical education contexts (Norris, van der Mars, Kulinna, & Beardsley, 2017).

The recent passing of the Every Student Succeeds Act (ESSA) could impact school physical education in several ways. First, physical education is now seen as a subject central to students' "well-rounded education," and thus (at least in theory) ESSA places it on the same level with other "core" subjects (i.e., Mathematics and English Language Arts). Second, states and school districts are still required to implement teacher and principal evaluation systems that are partially based on evidence of student achievement. Third, at least 20% of Title IV funding associated with ESSA and being distributed by states must fund "safe and healthy" schools, and 20% must go to subjects considered part of a "well-rounded education." Finally, physical education teachers are expected to develop credible evidence of student learning outcomes.

In K–12 education, the focus has shifted from "assessment of learning" to "assessment for learning" (e.g., Black & Wiliam, 1998a; Broadfoot & Black, 2004; Hay, 2006; Wiliam, Lee, Harrison, & Black, 2004). Hay (2006) argued that assessment has two central purposes: (a) assessment for accountability and (b) assessment for learning. Assessment for learning is associated with end-of-unit "summative" assessment and used primarily by teachers to assign final grades. Assessment for learning seeks to inform students (and teachers!) regarding their progress in learning the subject matter throughout a unit of instruction and school year (Black & Wiliam, 1998b). This is analogous to formative assessment and shifts the focus toward how teachers can best support students throughout the learning process (Broadfoot & Black, 2004). Throughout this paper, the terms *learning* and *performance* will be used interchangeably.

Several researchers have called for greater emphasis on formative assessment for learning where teachers collect data throughout the learning process on students' progress (e.g., Baker & Gordon, 2014; Hay, 2006; van der Mars & Harvey, 2010). This shift has come about as a result of an attempt to "align" curriculum, instruction, and assessment (Lund & Tannehill, 2014; Veal, 1992, 1995).

Few physical educators have integrated formal assessment for learning into their day-to-day teaching (Shepard, 2001). Moreover, they have reported that formal assessment of (or for) student

learning is too time consuming and has little value, and/or that they lack the necessary knowledge to perform such assessments (Kneer, 1986). Beyond the typical managerial indicators of attendance, dress, and on-time behavior, physical educators mostly use a mix of student attitude, participation, sportsmanship, and effort as primary performance indicators for grades (e.g., Desrosier, Genet-Volet, & Godbout, 1997; Hensley, Lambert, Baumgartner, & Stillwell, 1987; Imwold, Rider, & Johnson, 1982; Matanin & Tannehill, 1994; Tousignant & Siedentop, 1983). Moreover, physical educators dislike the nature of typical summative (i.e., end-of-unit) assessment (Pryor & Akwesi, 1998).

Greenwood and Maheady (1997) noted that the "inability to document meaningful changes in student performance has impeded our ability as teachers . . . to identify those instructional arrangements and practices that may be responsible for subsequent changes in learner performance" (p. 266). However, simultaneously performing regular instructional duties and assessing student performance throughout units of instruction is a challenging task that might be more within reach through continuing professional development. Continuing professional development that is ongoing and where (experienced) teachers have input and ownership, and work collaboratively with those providing the continuing professional development is key to the effectiveness thereof (e.g., Armour & Yelling, 2007; O'Sullivan & Deglau, 2006; Parker & Patton, 2016).

The use of technology is another performance standard for beginning and advanced physical educators (National Board for Professional Teaching Standards, 2014; SHAPE America, 2010b). However, physical education teachers in the United States and United Kingdom have been slower to adopt and utilize technology than their classroom counterparts (e.g., Thomas & Stratton, 2006; Vahey & Crawford, 2003).

Today's context of teacher evaluation makes it more imperative that physical education programs can (a) clearly articulate their intended outcomes and (b) provide evidence that students are learning something substantive regarding our subject matter, to constituents (including parents, school administrators, policy makers; Rink, 2007). Logically, this places ongoing formal-formative assessment of and for students' learning squarely as a central teaching function for physical education teachers. Therefore, this research

wanted to answer two research questions: (a) Will physical educators increase their use of formative-formal assessment of student learning as a consequence of a yearlong professional development program that included on-site coaching support? and (b) While teaching, can physical educators collect formal-formative assessment data on student learning at an acceptable level of reliability?

Method

Participants and Settings

This research project was approved by a university-based institutional review board. Seven licensed physical educators agreed to serve as participants (4 females, 3 males). Four teachers had developed an extensive record of professional involvement through presentations and participation in state and regional conferences and in professional organizations. Experience with the use of technology in their programs ranged from none to extensive (e.g., program website design). None had employed any handheld digital technology in their teaching. One middle school was located in an affluent suburban bedroom community near a large metropolitan city. A second suburban middle school was located in a suburban middle-class neighborhood. The remaining two middle schools and one high school were located in more rural communities.

In the middle schools, teachers employed a traditional multiactivity curriculum. The three teachers in one middle school program grouped their classes (class size range: 33–45) together in one gym for a common fitness activity. After that, students were free to select between multiple activities at various parts of campus and switch between them daily. The high school teacher employed a combination of Sport Education (Siedentop, Hastie, & van der Mars 2011) and Fitness for Life (Corbin & Lindsey, 2005).

Dependent Variables

The dependent variables included percentage of class time that teachers spent (a) formally and/or informally assessing students' performance and (b) doing other related teaching functions/behaviors (i.e., instructing, managing/organizing, and silently observing students). As well, students' moderate to vigorous physical activity (MVPA) behavior was assessed. Formal assessment was defined as

teachers recording information about student subject-matter performance using either paper and pen or an electronic assessment template on a handheld digital device (PDA) throughout the class period. Informal assessment was defined as the time spent providing (non-)verbal information to students in the form of positive and/or corrective feedback throughout the class period. The focus of the assessment could be on students' content, management, and social behavior.

It is important that data collected by teachers on their students' learning task performance are pertinent and credible (Williams & Rink, 2003). Therefore, data on students' in-class performance that teachers collected while instructing their classes were compared to those of trained outside observers. The students' MVPA level was used as the primary in-class performance indicator. MVPA is a key national health objective for school-age youth (e.g., Institute of Medicine, 2013; Sallis et al., 2012; U.S. Department of Health and Human Services, 2008, 2010, 2012). Students' MVPA levels can be assessed with relative ease at any time throughout the lesson, and this indicator cuts across much of the content taught in physical education (i.e., fitness, sport, rhythms).

Intervention

A combination of three all-day professional development workshops, along with on-site coaching support and prompting, served as the intervention. Teachers were provided with a digital technology tool (i.e., a PDA), which they were allowed to keep at the completion of the project, contingent on continued participation throughout the project.

The first workshop focused on (a) general principles and multiple purposes of assessment, with a specific emphasis on formal assessment conducted throughout units of instruction (Siedentop et al., 2011; van der Mars & Harvey, 2010); (b) formal assessment of physical activity as one student outcome indicator; (c) use of paper-and-pen-based formal assessment; (d) features and use of PDAs; and (e) live practice of paper-and-pen-based formal assessment of student PA levels in real physical education classes.

The second workshop focused on (a) formal assessment of students' skill execution (including overview of how to define "appropriate" and "inappropriate" skill execution), (b) electronic

versions of formal assessment templates, and (c) "live" practice exercises for teachers to formally assess students' PA levels using the PDA during regular physical education classes, and students' skill execution using paper and pen.

The third workshop included (a) reviews of previously discussed topics, overviews, and discussions of the project data to date; (b) formal assessment of more tactical aspects of gameplay (e.g., off-the-ball movement in invasion games, returning to base position in net/court games; Mitchell, Oslin, & Griffin, 2006), and (c) practice of PDA-based formal assessment of gameplay performance using 3- to 4-point gameplay performance scoring guides (Siedentop et al., 2011).

The overarching goal of the workshops was to focus teachers on formally assessing student performance in the psychomotor domain, with the introduction of specific assessment strategies for increasingly more complex indicators of student performance/learning. Initially, teachers learned to track their own students' MVPA levels. Furthermore, maximizing physical activity opportunities for students in physical education is an often-espoused program objective for many teachers. Depending on their interest and readiness, teachers were encouraged to include other targets for assessment, such as students' technique execution (e.g., serve, forehand shot, soccer pass), performance on tactical aspects of gameplay, and aspects of students' fair-play behavior.

The intervention sought to have teachers come to view formal assessment as a more manageable and acceptable teaching function, by emphasizing the "when," "how much," and "who" of assessment (van der Mars & Harvey, 2010). That is, assessment should be ongoing throughout most every lesson. This then allows teachers to make repeated observations over multiple lessons and make adjustments prior assessments of individual students, affords students multiple opportunities to demonstrate what and how well they can perform, and shifts assessment away from the dominant culture of "testing" students. Relative to "how much" to assess, teachers were encouraged to limit the scope of their formal assessment by focusing on one intended outcome in any one class. For example, if returning to base position between every stroke during badminton gameplay was a unit objective, that would be the sole formal assessment focus. In

terms of "who" to observe, teachers were directed to limit the number of students to be assessed per lesson.

On-site coaching support consisted of approximately two visits per 3 school weeks by one of the researchers. In addition, supplemental support was available through telephone- and e-mail-based communications. On-site coaching visits included (a) formal lesson observations, (b) postlesson feedback for teachers on their use of formal assessment (and any other pertinent lesson events), (c) encouragement, (d) discussions on what worked and did not work relative to engaging in formal assessments, (e) the use of PDAs once teachers started employing these, and (f) answering concerns and questions posed by teachers.

Across the seven participating teachers, the initial on-site coaching visits occurred after the first all-day workshop. For individual teachers, the initial coaching visits were staggered throughout the middle part of the school year and signified the start of the intervention phase for the teacher in question.

As part of the intervention, teachers wore a MotivAider (http://habitchange.com/) on their waist belt. The MotivAider provided prompts at set intervals (90 s or 120 s, depending on teachers' preference and level of comfort), reminding them to employ formal assessment. This resulted in teachers being prompted between 20 and 26 times per lesson. Teachers were instructed to ignore prompts that occurred during times where it would be inappropriate or inconvenient (e.g., during a demonstration, instructing an individual student or group of students, or attending to a safety issue).

Research Design

A variation of the multiple baseline design across subjects determined the effects of the intervention on teachers' assessment patterns (Cooper, Heron, & Heward, 2007. Multiple baseline designs have been used widely in educational settings for several decades in applied behavior analysis (ABA) research. Following the repeated measurement of the target behaviors under natural conditions (baseline sessions, noted as A), an intervention (B) is introduced across persons (or behaviors or settings) at varying time intervals and repeated observations of the target behaviors continue. A functional relationship between the intervention and the target behaviors is established if/when (a) the target behavior changes in the desired

direction only upon the implementation of the intervention and (b) the behavior of those who continue in baseline conditions does not change appreciably. In this study, because of scheduling issues (school calendars, etc.), the start of the intervention could not be staggered across each teacher. Postcheck follow-up observations of each teacher were made during the fall of the following school year. This allowed the researchers to determine generalization of the use of formal-formative assessment beyond the professional development program.

In some instances (e.g., first day of a unit of instruction, PDA left at home, PDA battery not charged, a scheduled student choice/free day), teachers announced before class that they would not allocate class time to formally assess student learning. Across all teachers, this occurred in nine of the 88 (10.2%) intervention sessions.

Data Collection

Data on teachers' assessment patterns practices were collected via the validated Systematic Observation of Formal Assessment of Students by Teachers (SOFAST; van der Mars, Timken, & McNamee, 2018). SOFAST is a three-level observation system where observers code (a) teachers' primary teaching functions (i.e., Assessment [formal and informal], Instruction, and Management), (b) the focus of teachers' assessment (i.e., content, management, and social behavior), and (c) the lesson context (i.e., Management, Fitness, Knowledge, Skill Practice, Game, or Other). SOFAST is an interval recording-based observation system using alternating 10-s "observe" and 10-s "record" intervals. A full description of SOFAST can be obtained from the lead author.

Students' in-class MVPA levels were assessed throughout the study, and the researchers determined whether the teachers' formal assessment data reached acceptable reliability levels by comparing teachers' data with independent observers' data. Teachers were asked to make a dichotomous decision when determining students' PA levels, based on the PA level categories of the System for Observing Fitness Instruction Time (SOFIT; McKenzie, Sallis, & Nader, 1991; "No-MVPA" = Lying Down + Sitting + Standing, "Yes-MVPA" = Walking and Very Active; Rowe, van der Mars,

Schuldheisz, & Fox, 2004; Williams & Rink, 2003). Class sessions used for this purpose were separate from those sessions used for determining the effects of the intervention. Teachers and a trained reliability observer collected data on MVPA levels of three randomly selected target students. Target students wore a yellow pinny (numbered 1 to 3), which ensured that both the teacher and the outside observer would observe the same students in the same order.

Both the teachers and the independent reliability observers used momentary time sampling with observation intervals lengths set by the teacher (range 90–120 s). MVPA (non-)occurrence was observed and immediately recorded at the end of each interval. Momentary time sampling is an appropriate and valid means of collecting data across a wide variety of behaviors (including PA), persons, and settings (e.g., Gunter, Venn, Patrick, Miller, & Kelly, 2003; Harrop & Daniels, 1986; McNamee & van der Mars, 2005; Saudargas & Zanolli, 1990; Test & Heward, 1984; van der Mars, 1989a). To ensure that both the teachers and reliability observers would observe and record students' MVPA levels at the same time, the researchers had both wear a MotivAider and synchronize the prompting devices on the outset of each lesson so that both would be prompted at the same time to observe and record the target students' PA levels.

Data Analysis

The researchers used visual analysis of graphically plotted data (the standard analytical approach used in ABA research) to determine the functional relationship between the intervention and the teachers' assessment patterns. They used the following visual analysis criteria: (a) data variability within and across phases, (b) data trends within and across phases, (c) data overlap between phases, and (d) changes in level between phases (Cooper et al., 2007).

For the second research question (i.e., reliability of the physical educators' assessments of students' MVPA levels), the researchers calculated standard interobserver agreement (IOA) percentages using the total interval method (van der Mars, 1989b). The 85% criterion was set as the minimum mean level of IOA percentage for the teachers' observations to be considered sufficiently reliable.

SOFAST Observer Reliability

The researchers established observer reliability for SOFAST observations by conducting at least one IOA check during both the baseline and intervention phase for each teacher, using the total interval method to calculate the IOA percentages. IOA percentages across SOFAST categories were at acceptable levels across all teachers and project phases (see Table 1). Of the 210 IOA percentages, eight (3.8%) were below 80%. All but one occurred during IOA checks conducted during baseline sessions, and these were a consequence of low behavior occurrences. Based on the IOA results, the observers were deemed reliable.

Results

Class Context Data

Class time distribution for SOFAST context categories is presented for all teachers across baseline and intervention phases. The data for each teacher are presented across the five tiers in Figure 1 (from top to bottom). Throughout the intervention, Amanda's classes spent the majority of class time in health-related fitness content (74.4%), as a consequence of Amanda teaching a girls-only Sport Education–based weight training course. Within the finite amount of class time available, this coincided with a significant decrease in the time spent on Skill Practice (-19.6%) and Game activities (18.9%). Beth's classes on average spent less time in Managerial- (–11.5%), Knowledge- (–10.6%), and Skill Practice–related activities (–16.1%) during the intervention, compared to the baseline sessions, but more time in Game-related (or Competition-related) content (21.5%).

Chuck's classes included a significant Fitness component (third tier). The reduced time spent in Management from baseline to intervention (-11.2%), coincided with an increase in time spent on Skill Practice content (14.3%).

Table 1

Total Interval Interobserver Agreement Percentages Across Teachers and Conditions

	Participating teachers													
	1		2		3		4		5		6		7	
Condition	**BL**	**INT**	**BL**	**INT**	**BL**	**INT**	**BL**	**INT**	**BL**	**INT**	**BL**	**INT**	**BL**	**INT**
Teacher functions														
Formal Assessment	87.50	87.50	100.00	100.00	93.75	100.00	91.60	96.30	98.00	80.60	100.00	100.00	97.40	98.10
Informal Assessment	92.30	92.30	88.00	93.40	90.15	95.20	94.50	80.50	90.30	81.50	94.70	83.30	92.90	92.40
Participation/ Demonstration	100.00	100.00	91.00	98.00	95.50	100.00	100.00	100.00	88.90	85.60	50.00	96.60	96.10	95.90
Knowledge	80.00	80.00	68.00	91.50	74.00	100.00	100.00	96.30	96.80	80.50	91.60	86.20	99.00	96.00
Management	90.30	90.30	94.00	99.40	92.15	97.20	98.70	96.60	99.00	98.90	100.00	100.00	100.00	93.20
Silent Observation	94.40	94.40	33.00	91.40	63.70	100.00	90.30	83.90	93.70	80.10	100.00	90.00	92.30	89.10
Teacher Function Average	90.75	90.75	79.00	95.62	84.88	98.73	95.85	92.27	94.45	84.53	89.38	92.68	96.28	94.12
Assessment Focus														
Content	100.00	100.00	86.00	99.20	93.00	97.20	96.30	98.40	90.20	98.20	98.70	90.30	98.40	90.40
Management	100.00	100.00	100.00	100.00	100.00	100.00	95.00	93.80	97.90	100.00	61.80	100.00	90.50	90.70
Social Behavior	50.00	50.00	100.00	100.00	75.00	100.00	100.00	100.00	93.40	92.80	100.00	100.00	100.00	100.00
Assessment Focus Average	83.33	83.33	95.33	99.73	89.33	99.07	97.10	97.40	93.83	97.00	86.83	96.77	96.30	93.70
Class Context														
Management	90.20	90.20	94.00	96.20	92.10	95.40	97.60	99.20	94.00	92.40	86.30	66.00	98.10	87.50
Knowledge	100.00	100.00	94.00	100.00	97.00	75.00	94.90	100.00	90.60	83.90	90.50	100.00	93.90	100.00
Fitness	100.00	100.00	100.00	100.00	100.00	100.00	100.00	83.80	100.00	100.00	100.00	100.00	95.50	92.70
Skill Practice	97.80	97.80	96.00	100.00	96.90	100.00	98.90	100.00	98.90	100.00	100.00	100.00	100.00	100.00
Game	96.60	96.60	100.00	94.00	98.30	96.10	92.90	99.80	92.90	93.40	100.00	100.00	94.50	92.10
Other	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	98.00	100.00	100.00	100.00	93.90	100.00
Class Context Average	97.43	97.43	97.33	98.37	97.38	94.42	97.38	97.13	95.73	94.95	96.13	94.33	95.98	95.38

Note. BL = Baseline IOA; INT = Intervention IOA.

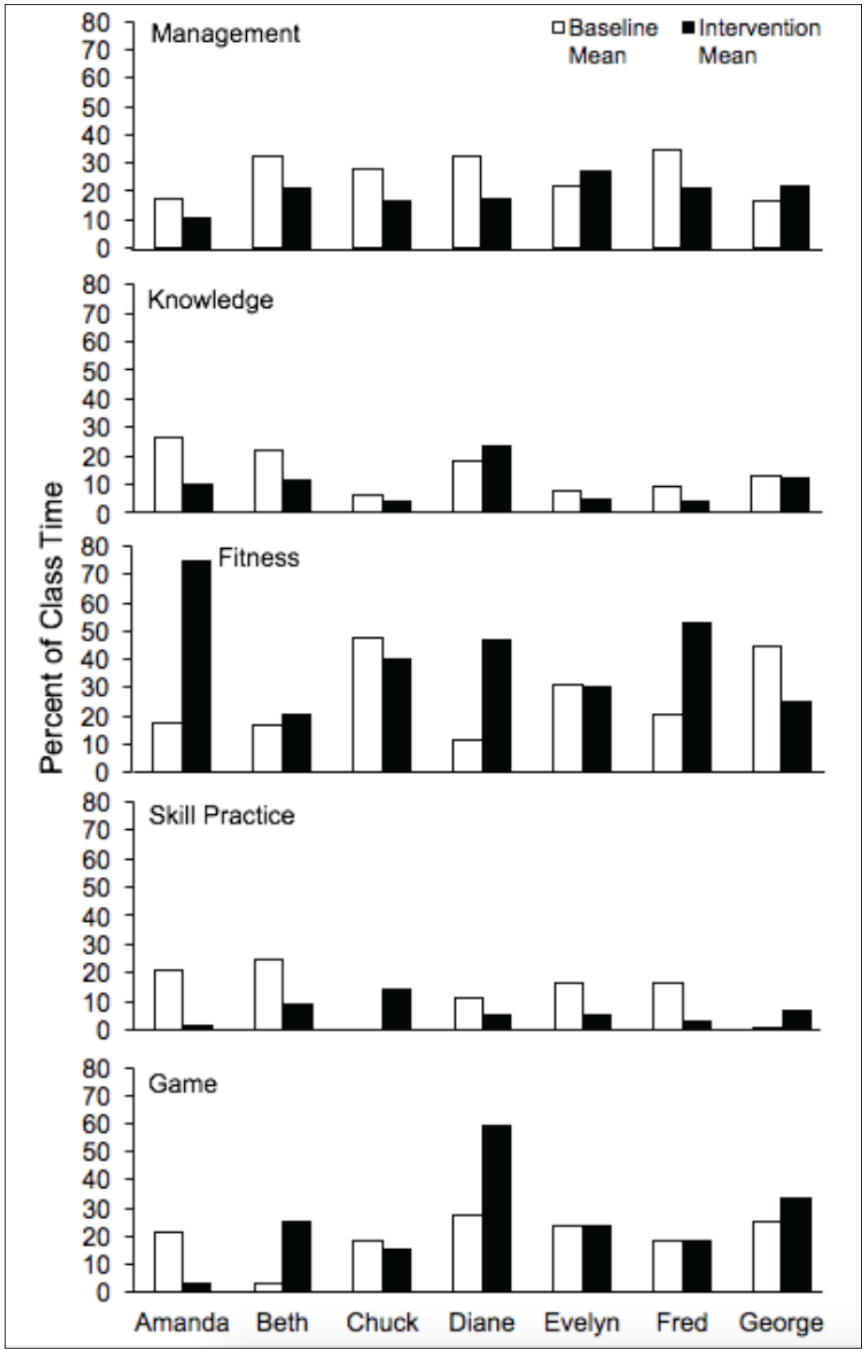

Figure 1. Mean percentage of class time during baseline and intervention across SOFAST class context categories.

Students in Diane's classes spent less time in Managerial tasks (−14.8%) and Game content (−22%) on average during intervention, compared to baseline sessions. This is related to the significant increase in fitness content (35.4%). Changes from baseline to intervention were less pronounced for students in Evelyn's classes, other than a slight increase in Management time (4.8%) and reduced time spent in Skill Practice (−11%). Students in Fred's program also experienced a significant increase (31.8%) in time spent on Fitness from baseline to intervention. This coincided with lower percentages for Management (−13.8%) and Skill Practice (−13.5%). Finally, students in George's classes experienced reduced class time in Fitness content (−19.5%) and increases in Managerial tasks (5.1%), Skill Practice (6.4%), and Game content (8.3%).

Teaching Functions Data

Figure 2 summarizes the data across the five teaching functions across baseline and intervention. On average, the class time spent engaged in Managerial activities during intervention, compared to baseline, was lower for Beth (−6.4%), Chuck (−9.5%), Diane (−20.9%), Evelyn (−7.0%), and Fred (−15.9%). Amanda's engagement in Managerial activities remained unchanged, while George increased his involvement by 5.8%.

As a group, teachers spent little time demonstrating/participating with students in physical activity during baseline. Most teachers (with the exception of Diane and George), lowered their demonstration/participation levels even further during intervention. Beth, Chuck, Evelyn, Fred, and George reduced the amount of time they spent instructing their students (i.e., content-related explanation and skill prompts) from baseline to intervention by 4.4%, 4.1%, 5.6%, 7.7%, and 7.5%, respectively. On average, silent observation of students by Chuck, Evelyn, and Fred increased during Intervention by 4.7%, 10.4%, and 3.7%, respectively.

Already low during baseline, silent observation decreased with Amanda and Beth (−5.6% and −2.4%, respectively). Finally, as a group, teachers spent a significant amount of their time assessing students' performance during baseline. This included the time spent in formal assessment. During intervention, except for George, all teachers increased their time on assessment between 7.1% and 23.2%.

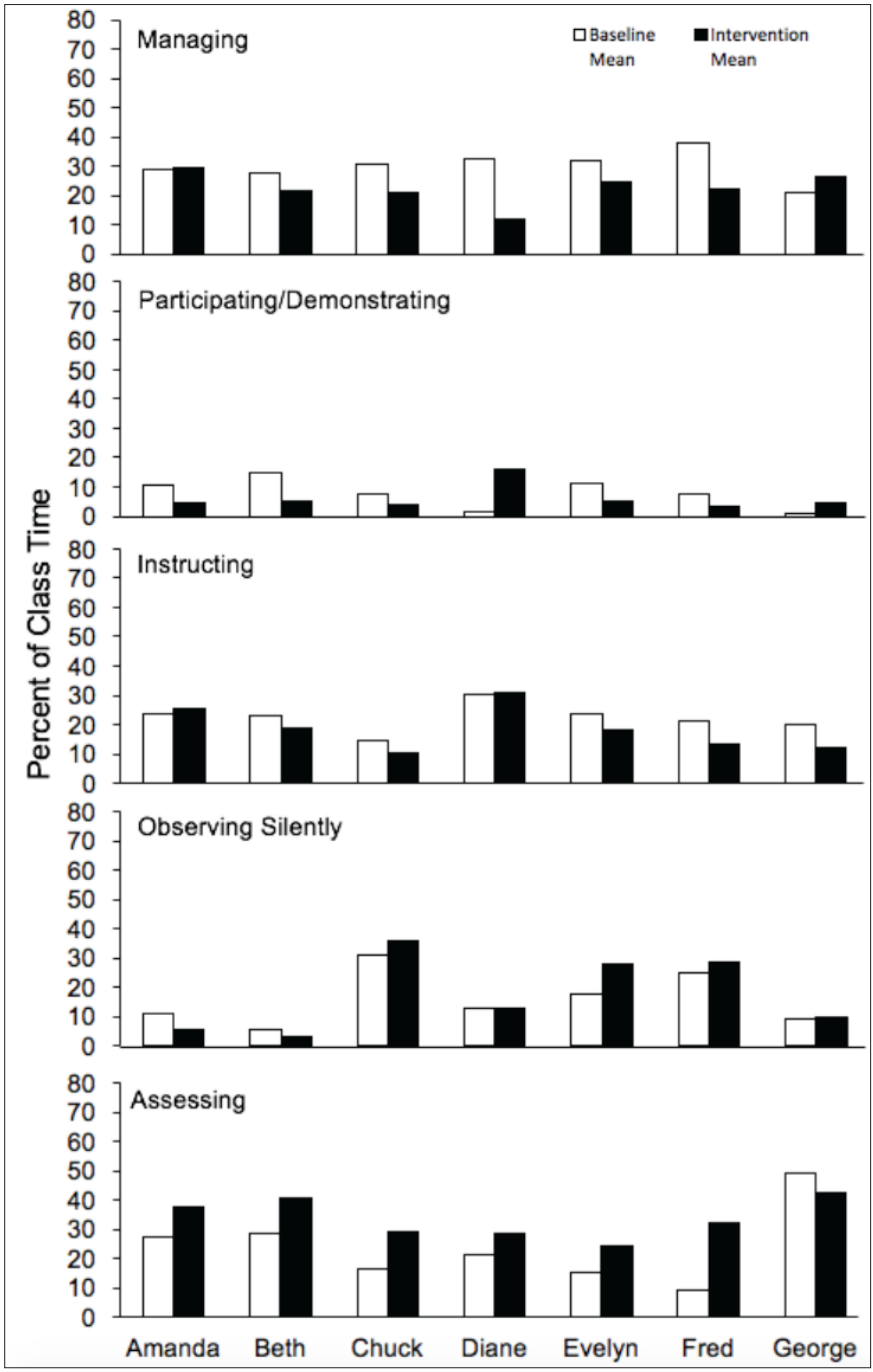

Figure 2. Mean percentage of class time during baseline and intervention across SOFAST teaching function categories.

Intervention Data

Figures 3 and 4 include data on the percentage of class time that each teacher allocated for formal-formative assessment. The start of the intervention for each participant is signified by the furthest left continuous vertical dashed line. On select days, teachers indicated they would not formally assess student performance for specific reasons, and these days are labeled with an asterisk above the session's data point. The furthest right vertical dashed line reflects the start of postcheck sessions across teachers.

During baseline (A), time spent on formal assessment of any kind was minimal across all teachers. The formal assessment that occured targeted Managerial aspects of students' performance (i.e., attendance, dress, on-time behavior) and was largely stable across teachers, with minimal variability and phase means ranging from .9% to 5% across teachers.

Immediate and appreciable increases occurred upon intervention (B). With the exception of the sessions marked with an asterisk, data overlap between baseline and intervention phases was minimal across all teachers.

Table 2 includes phase means and standard deviations for baseline, intervention, and postcheck sessions (intervention means and standard deviations include the sessions marked with an asterisk). It also includes data on the change in level from the final baseline session to the first intervention session. The change-in-level percentage ranged from a low of 8.1% (Evelyn) to a high of 27% (George).

Postcheck observations were made during the subsequent fall, and the researchers used them to determine whether teachers would sustain their use of formal-formative assessment of substantive student outcomes beyond the intervention. Figures 3 and 4 show encouraging results in that Beth, Diane, Evelyn, Fred, and George in at least some of the postcheck sessions engaged in formal assessment of student learning at levels that were at or above initial baseline levels.

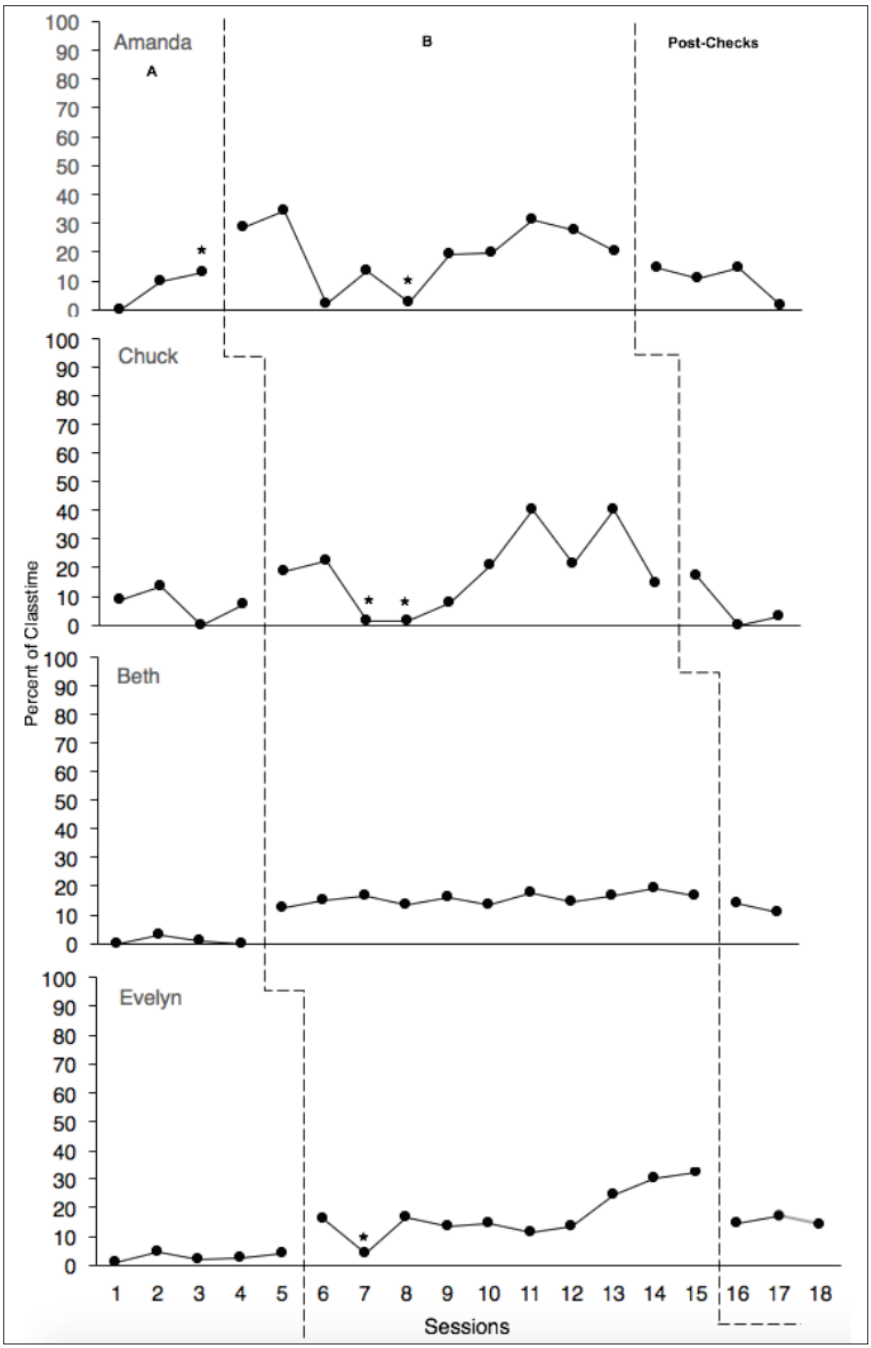

Figure 3. Mean percentage of class time spent on formative-formal assessment across conditions.

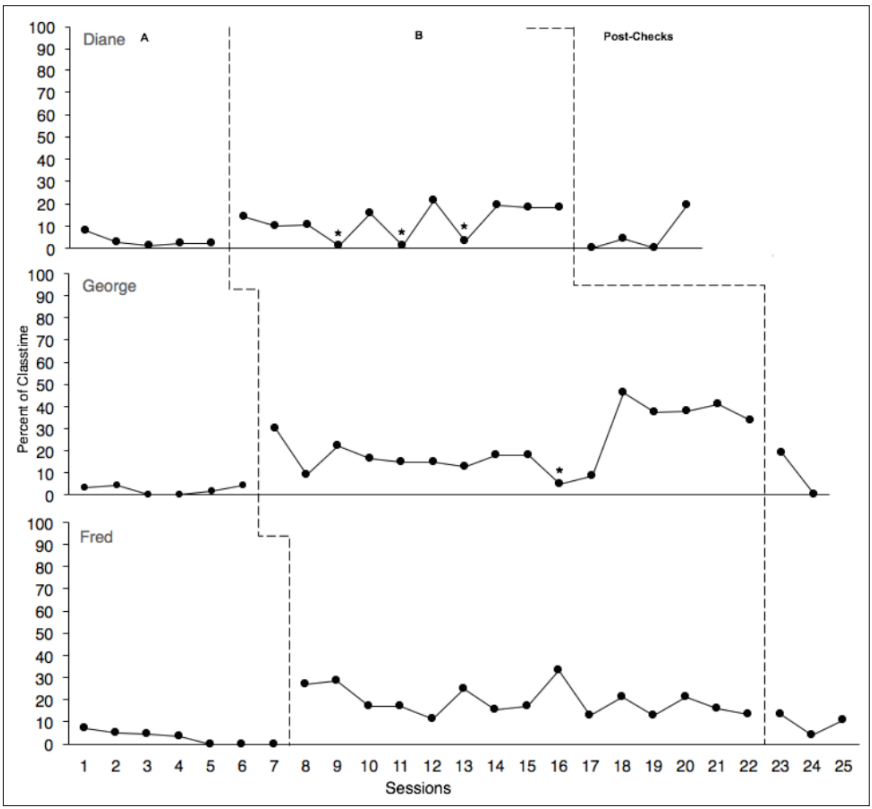

Figure 4. Mean percentage of class time spent on formative-formal assessment across conditions.

Table 2

Mean Percentage of Class Time Spent in Formative-Formal Assessment Across Phases

	Baseline		Change	Intervention		Postcheck	
Teacher	*M*	(*SD*)	in level[a]	*M*	(*SD*)	*M*	(*SD*)
Amanda	5.5	(5.6)	15.2	19.4	(10.7)	8.9	(8.6)
Beth	0.9	(1.3)	12.5	15.5	(2.4)	15.5	(2.7)
Chuck	5.7	(5.3)	9.4	18.8	(13.7)	6.7	(9.3)
Diane	3.0	(1.5)	12.0	19.1	(7.8)	15.3	(1.7)
Evelyn	2.7	(2.3)	8.1	12.1	(7.5)	5.9	(9.1)
Fred	2.8	(2.1)	25.3	22.9	(13.5)	9.6	(13.5)
George	2.8	(2.8)	27.0	19.3	(6.6)	9.2	(4.7)

[a]Difference between first intervention session and final baseline session.

Assessment Focus

Figure 5 presents the shift in teachers' formal-formative assessment focus between Content, Management, or Social behavior task performance. During baseline, teachers, except for Amanda and Chuck, focused exclusively on students' performance on Management-related expectations (i.e., attendance, dress, being on time), as shown in the white bars. Upon the start of the intervention, all teachers shifted their formative-formal assessment to students' performance on Content learning tasks. Furthermore, they maintained this emphasis during postcheck sessions.

During intervention, several teachers voiced interest in increasing their efforts in formally assessing students' overall class conduct (i.e., personal and social behavior), as they viewed this area as key in terms of what their program sought to accomplish. Although teachers may have kept records on this outside of class, as can be seen in Figure 2, during the 88 intervention sessions across all participating teachers, none included any within-lesson formal assessment of students' general class conduct.

Reliability of Teachers' Observations of Student Performance

The teachers' reliability in observing their students' MVPA was checked between six and eight times throughout the project's intervention phase. When a teacher used his or her PDA approach for recording student data, the outside observer would also use a PDA. If the teacher used a paper-and-pencil approach, then the outside observer used the same approach for data collection. Figure 6 includes the IOA data, including the mean, standard deviation, and range of IOA percentages. Four of the teachers used their PDA during the IOA sessions. Their mean IOA percentages ranged from 88.6% (Diane) to 96% (Beth). Except for Fred, whose data had the largest variability, all teachers met the 85% IOA criterion set prior to the study, with Fred's data having the largest variability.

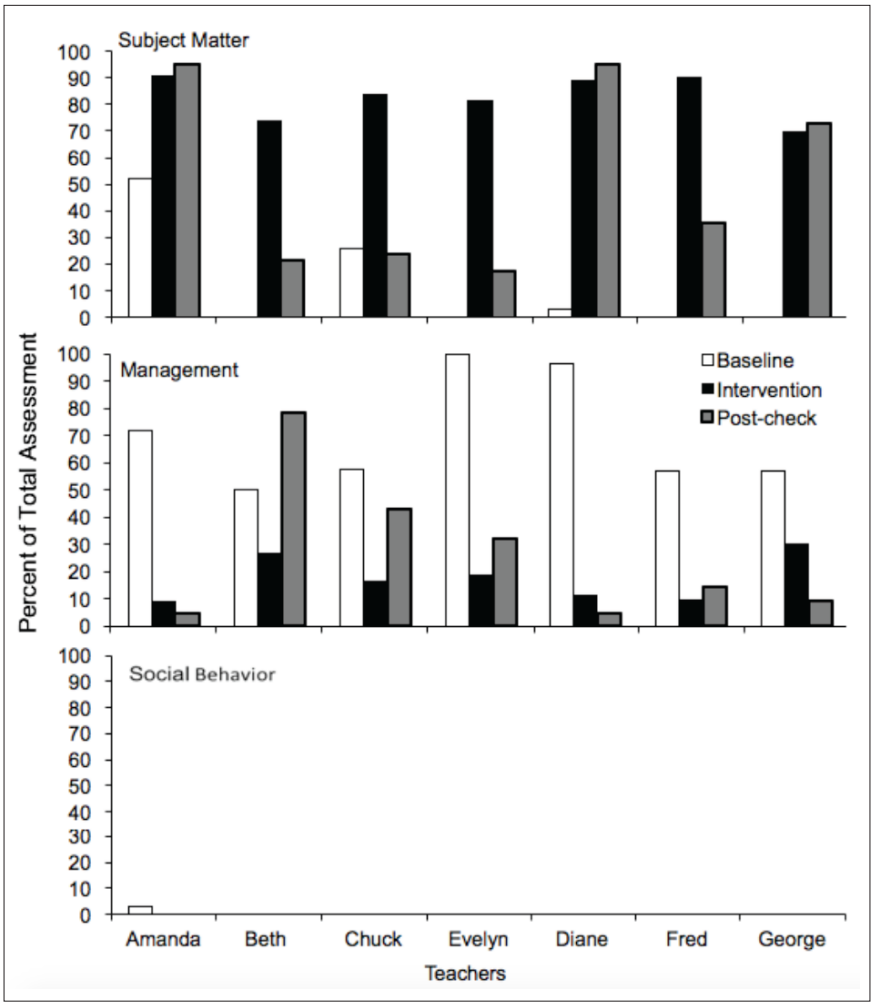

Figure 5. Mean percentage of formal-formative assessment of students' management, subject matter, and social behavior performance across conditions.

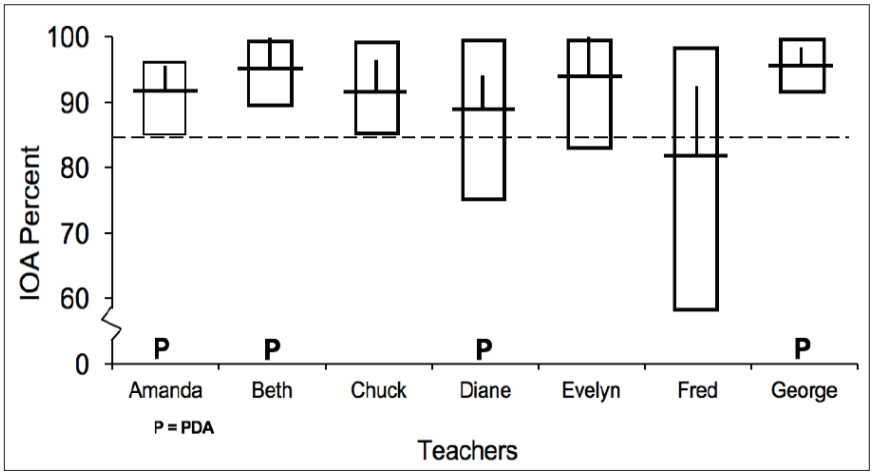

Figure 6. Mean interobserver agreement percentages (with *SD* and ranges) across teachers.

Discussion

Professional Development Intervention

The combination of professional development workshops, on-site coaching, and prompting enabled teachers to increase teachers' formal-formative assessment of substantive student performance. The changes in the assessment patterns and focus did not occur until the start of the intervention, were immediate and substantial, and were sustained throughout the intervention, as well as during post-checks. This reflects a functional relationship of the intervention with the teachers' assessment practices.

Although no formal analysis was made of which specific intervention component (workshops, on-site coaching, and prompting) was the more critical component, the workshops and the on-site coaching components were interdependent in the sense that without the workshops, the on-site coaching would have lacked context. Similarly, if the workshops had been the sole intervention without opportunities to practice the skills associated, progress might have been more difficult. The workshops offered teachers important background on the role and importance of assessment in teaching physical education and how assessment has multiple functions

beyond determining grades. Workshops also afforded teachers the opportunity to learn about and discuss with colleagues/peers concepts and issues related to assessment, the importance of which was documented by Armour and Yelling (2007).

Early on during the intervention, teachers tracked students' MVPA levels (three per lesson). This performance indicator was chosen because of the relative ease with which it can be assessed. It also allowed teachers to become more comfortable engaging in ongoing formal-formative assessment as a daily teaching function. While possibly viewed as a narrow indicator of student learning outcomes, student MVPA reflects a focus on national health objectives (e.g., Institute of Medicine, 2013; Sallis et al., 2012; U.S. Department of Health and Human Services, 2010), as well as a program outcome, oftentimes espoused by physical educators themselves.

Choice, ownership, and teacher-centeredness during professional development efforts are deemed cornerstones of effective professional development (e.g., Armour & Yelling, 2007; O'Sullivan & Deglau, 2006; Parker & Patton, 2016). From that perspective, the decision of what to assess was left to the teachers after the initial workshop. The researchers merely offered suggestions and support when asked (e.g., through provision of assessment tools). As the intervention phase unfolded, teachers were introduced to a variety of standards-based assessment templates that (a) covered a wider range of student outcome indicators and (b) allowed for true authentic assessment (i.e., assessment of student performance in conditions that reflect real-world settings). Four of the teachers broadened the scope of what they chose to assess by focusing on student learning indicators such as skill execution, gameplay performance from a tactical perspective, and fair play behavior. This is a desirable developmental step in the process of making assessment more of a habit than a nuisance.

Comments (as well as actions) from most teachers during postlesson discussions reflect that they came to view assessment as distinctly broader than the assigning of grades, and point to an increased interest in assessment for learning (Hay, 2006). An example of evidence for this occurred toward the end of the intervention when George's recording of a student's performance became a trigger to provide his students with either technique- or tactics-focused

prompts or feedback. While the combined use of formal and informal assessment did not become a prevalent pattern among all teachers, it does suggest that physical educators, with practice, can indeed weave formal assessment of student performance throughout during their instruction. This supports Desrosiers et al.'s (1997) argument that authentic assessment should be integrated within the teaching–learning process, be shared with the students, and have a formative focus.

Balancing the central teaching functions of instruction, management, and monitoring of students (Siedentop & Tannehill, 2000) is a complex endeavor that teachers engage in every day and every lesson. Given this context, making formal-formative assessment a more habitual/daily function was assumed to be a difficult process, as teachers have reported that formal assessment of student learning is too time consuming and lacks relevance (Hay, 2006; Kneer, 1986). Interestingly, physical educators spend a significant amount of class time silently observing students (i.e., 20–35% of total class time; Siedentop & Tannehill, 2000). However, to date, little is known about what teachers look for, what they think about, and/or what they plan for when silently observing the environment and students. Logically, then, at least part of this time spent in silent observation might lend itself to more focused and deliberate observations and subsequent recording of data on student performance.

As Figure 2 shows, each teacher took a slightly different approach to shifting the balance across the various teaching functions, to create time for themselves to formally assess their students. For example, most teachers spent less time in managerial tasks during intervention than during baseline. Surprisingly, the three teachers with the highest percentage of class time spent silently observing students during baseline spent even more time on this during the intervention. Yet even they managed to build in class time for the assessment function.

As part of the intervention, teachers were introduced to employing a digital handheld device to record student performance data. Compared to their classroom counterparts, physical educators have been found to be more resistant to employing (or at least slower to employ) technology in their teaching (e.g., Thomas & Stratton, 2006; Vahey & Crawford, 2003). The transition to using a PDA

differed for the participating teachers. Amanda, Beth, Diane, and George became consistent users of a PDA. For the others, employing formal-formative assessment via paper and pencil was sufficiently overwhelming, and using the PDA likely made the practice of formal assessment more difficult. Consequently, their PDA use was more intermittent. This pattern was similar to findings reported previously (e.g., Thomas & Stratton, 2006; Vahey & Crawford, 2003).

The intervention was successful for at least three reasons. First, throughout the workshops and on-site coaching, the researchers framed the process of formal assessment as ongoing, limited in scope, and focused on only a few students per lesson, and then offered teachers opportunities to practice the techniques associated with this type of formal-formative assessment. This may have helped the participating teachers view the assessment function as more manageable, and perhaps more acceptable.

Second, the structure of the professional development intervention went well beyond merely introducing a set of general principles of assessment and letting teachers figure out for themselves how to translate these into practice (Black & Wiliam, 1998b). It included opportunities to practice and on-site coaching support. Furthermore, during the workshops, teachers shared positive and negative experiences in designing and using their own assessment tools. Such active engagement is regarded a key component of quality professional development for teachers (e.g., O'Sullivan & Deglau, 2006; Parker & Patton, 2016), and it helped create a sense of partnership among teachers and with project leaders.

The contribution of the periodic prompts (coming between 90 s and 120 s) throughout the lessons cannot be underestimated. Prompts are a critical tool in the early stages of learning most every new skill (Cooper et al., 2007). The MotivAider kept the teachers focused on practicing a skill that is essentially new, relative to other more established teaching skills. It likely also contributed to the relatively stable trends and limited variability in the intervention phase data across all teachers (other than the sessions marked with an asterisk), which points to the effectiveness of this specific intervention component. During informal postlesson discussions, several teachers noted that with some practice and experience, the process of formal-formative assessment became easier. For example, George

noted that during gameplay portions of his badminton unit, he managed to assess as many as 14 students on their return to base position following each stroke. By the end of the intervention phase, George had phased out the use of the prompting device, yet maintained a steady level of formal assessment.

Reliability of Teachers' Observations of Student Performance

The complex task of orchestrating teaching functions of instruction, management, and assessment warranted an assessment of whether teachers would be reliable in their formal assessment of students' MVPA. Across all the teachers, just over 2,500 on-the-spot assessment decisions were made during the IOA sessions. Of those the assessments the teachers made, well over 2,100 instances matched those of the outside reliability observers. Six of the seven teachers met the preset 85% IOA criterion. These percentages are in line with those reported by Williams and Rink (2003). These levels of agreement are more than acceptable and offer confidence that teachers, with proper training and support, can reliably assess students' physical activity behavior on the fly. Furthermore, physical activity behavior can vary in duration and has a clear start and end. Thus, if the momentary sampling of the two observers is off by as little as 1 s (which is not unlikely), teachers might differ in their judgment of whether the student was engaged in MVPA.

The strengths of this study include (a) the efficacy of longer term (yearlong) professional development support to demonstrate that secondary physical educators are more than capable of employing formal-formative assessment throughout their lessons, (b) the shift in teachers' assessment from being focused primarily on students' managerial performance to subject matter–specific performance, and (c) the sustained level of such assessment beyond the intervention phase. Moreover, the professional development–focused intervention employed in this study is likely a key to building a culture where ongoing formal-formative assessment of student performance becomes an accepted part of physical education teaching practices.

This study was not without limitations. First, it lacked formal reliability checks on teachers' assessment of the more complex student performance indicators (e.g., tactical performance in gameplay).

All participating teachers were introduced to several three-level scoring rubrics that included gameplay performance indicators (e.g., Volleyball, Pickleball, and Basketball). However, the amount of workshop time available (approximately 24 hr spread over 3 days) allowed only for the introduction to the assessment templates and limited video-based and live observation practice opportunities. All seven teachers were encouraged to "try them out" in their own settings, but only four did so. For each activity, the assessment scoring guides included several performance indicators (e.g., court coverage, guarding/marking) from which teachers could choose. Each indicator included descriptors of observable gameplay behaviors at three performance levels (progressing, meets, exceeds). Compared to traditional, de-contextualized skills tests, these scoring guides reflect more authentic process indicators of progress and learning. While no IOA check were conducted, we believe that if teachers have the needed content knowledge and target specific student outcomes, they can, with practice, become skillful in assessing such outcomes.

Second, the small sample prevents generalization to all middle school physical educators. However, future systematic replications can add to this study's evidence. Third, even though there is now evidence that teachers can effectively and reliably assess students' substantive outcomes, and that they can sustain this beyond the intervention phase, there is no guarantee that they will in all cases. Physical education's policy landscape is potentially changing for the better (e.g., ESSA) with improved support for physical education, which is now regarded essential to students being well educated. However, until explicit program outcome expectations are in place at the state level, along with expectations for formal assessment of such outcomes by physical educators, it is more likely that the formal assessment strategies targeted in this study will occur largely because of the level of professionalism by individual teachers.

A final limitation is the constraints put on the researchers in scheduling the start and end of the intervention phase. For a multiple baseline design, the goal is to vary the lag time of the start for each teacher. Because of school schedules, the lag times were limited in a number of cases. However, this was offset by the immediate and substantial changes in assessment patterns across teachers.

Practical Implications

School physical education programs currently enjoy substantial support from outside the field (e.g., Centers for Disease Control and Prevention, 1997, 2001; Pate et al., 2006) in terms of their role in reversing the overweight/obesity trends among children and youth. This is evidenced by the emergence of national guidelines and recommendations and the advances in policy development and legislative efforts (e.g., ESSA) specific to physical education. Physical education cannot afford to claim its benefits and importance without being able to provide credible evidence of what it accomplishes. School physical education programs are part of a (publicly) funded school system. Thus, they bear the responsibility to demonstrate that continuing this investment is justified (Rink, 2007).

While the ultimate impact of such policy and legislative efforts are yet unknown, there is evidence that physical education programs are ill-prepared to present evidence that they have appreciable impact. For a functional culture of assessment to emerge in physical education, preservice PETE programs and professional development programs for already certified teachers must increase efforts in equipping current and future teachers with the skills, knowledge, and dispositions necessary to make assessment of student learning a teaching function that is viewed as a normal part of daily work.

The central message from this study is that formal-formative assessment of students' subject-matter performance is within reach for physical educators. The results also reinforce the need for professional development to be ongoing and long term. It should also allow for ample practice in developing the desired formal-formative assessment of learning skills and for active involvement by participating teachers in shaping specific features of such efforts (Armour & Yelling, 2007). Furthermore, these results are more likely to occur if a clear context is provided on the need and multiple purposes of assessment and on the link between assessment and instructional goals.

Conclusion

Focused and ongoing professional development that includes on-site coaching helps experienced secondary physical education teachers to (a) infuse formal-formative assessment of student in-class

performance as a primary teaching function and (b) shift the focus of such assessment efforts toward substantive learning outcomes.

References

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44,* 448–452. https://doi.org/10.3102/0013189X15618385

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Retrieved from http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System (EVAAS). *Educational Researcher, 37*(2), 65–75. https://doi.org/10.3102/0013189X08316420

Amrein-Beardsley, A. (2012). Value-added measures in education: The best of the alternatives is simply not good enough. *Teachers College Record, 114*. Retrieved from https://www.tcrecord.org/

Armour, K. M., & Yelling, M. (2007). Effective professional development for physical education teachers: The role of informal, collaborative learning. *Journal of Teaching in Physical Education, 26*, 177–200. https://doi.org/10.1123/jtpe.26.2.177

Baker, E. L., & Gordon, E. W. (2014). From the assessment of education to the assessment for education: Policy and futures. *Teachers College Record, 116,* 1–24.

Berliner, D. C. (2013). Problems with value-added evaluations of teachers? Let me count the ways! *Teacher Educator, 48*, 235–243. https://doi.org/10.1080/08878730.2013.827496

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record, 116,* 1–31. Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17293

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*, 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80,* 139–144.

Broadfoot, P., & Black, P. (2004). Redefining assessment: The first ten years of 'Assessment in Education.' *Assessment in Education, 11,* 7–27. https://doi.org/10.1080/0969594042000208976

Centers for Disease Control and Prevention. (1997). Guidelines for school and community programs to promote lifelong physical activity among young people. *Morbidity and Mortality Weekly Report, 46*(RR-6), 1–37.

Centers for Disease Control and Prevention. (2001). Increasing physical activity: A report on recommendations of the Task Force on Community Preventive Services. *Morbidity and Mortality Weekly Report, 50*(RR18), 1–16.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Columbus, OH: Merrill.

Corbin, C. B., & Lindsey, R. (2005). *Fitness for life* (5th ed.). Champaign, IL: Human Kinetics.

Desrosiers, P., Genet-Volet, Y., & Godbout, P. (1997). Teachers' assessment practices viewed through the instruments used in physical education classes. *Journal of Teaching in Physical Education*, *16*, 211–228. https://doi.org/10.1123/jtpe.16.2.211

Doolittle, S. (1996). Practical assessment for physical education teachers. *Journal of Physical Education, Recreation, and Dance, 67*(8), 35–38. https://doi.org/10.1080/07303084.1996.10604835

Greenwood, C. R., & Maheady, L. (1997). Measurable change in student performance: Forgotten standard in teacher preparation. *Teacher Education and Special Education, 20,* 265–275. https://doi.org/10.1177/088840649702000307

Gunter, P. L., Venn, M. L., Patrick, J., Miller, K. A., & Kelly, L. (2003). Efficacy of using momentary time samples to determine on-task behavior of students with emotional/behavioral disorders. *Education and Treatment of Children, 26*, 400–412.

Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis, 19,* 73–77. https://doi.org/10.1901/jaba.1986.19-73

Hay, P. J. (2006). Assessment for learning in physical education. In D. Kirk, D. Macdonald, M. M. O'Sullivan (Eds.), *Handbook of physical education* (pp. 312–325). London, England: Sage. https://doi.org/10.4135/9781848608009.n18

Hensley, L., Lambert, L., Baumgartner, T., & Stillwell, T. (1987). Is evaluation worth the effort? *Journal of Physical Education, Recreation, and Dance, 58*(6), 59–62. https://doi.org/10.1080/07303084.1987.10609577

Imwold, C. H., Rider, R. A., & Johnson, D. J. (1982). The use of evaluations in public school physical education programs. *Journal of Teaching in Physical Education, 2,* 13–18. https://doi.org/10.1123/jtpe.2.1.13

Institute of Medicine. (2013). *Educating the student body: Taking physical activity and physical education to school*. Washington, DC: National Academies Press.

Kneer, M. (1986). A description of physical education instructional theory/practice gap in selected secondary schools. *Journal of Teaching in Physical Education, 5,* 91–106. https://doi.org/10.1123/jtpe.5.2.91

Lavigne, A. L. (2014). Exploring the intended and unintended consequences of high-stakes teacher evaluation on schools, teachers, and students. *Teachers College Record, 116*. Retrieved from http://www.tcrecord.org/content.asp?contentid=17294

Lund, J. (1992). Assessment and accountability in secondary physical education. *Quest, 44,* 352–360. https://doi.org/10.1080/00336297.1992.10484061

Lund, J., & Tannehill, D. (Eds.). (2014). *Standards-based curriculum development* (3rd ed.). Sudbury, MA: Jones and Bartlett.

Lund, J., & Veal, M. L. (2008). Measuring pupil learning—How do student teachers assess within instructional models? *Journal of Teaching in Physical Education, 27,* 487–511. https://doi.org/10.1123/jtpe.27.4.487

Matanin, W. C., & Tannehill, D. (1994). Assessment and grading in physical education. *Journal of Teaching in Physical Education, 13,* 395–405. https://doi.org/10.1123/jtpe.13.4.395

McKenzie, T. L., Sallis, J. F., & Nader, P. R. (1991). SOFIT: System for Observing Fitness Instruction Time. *Journal of Teaching in Physical Education, 11,* 195–205. https://doi.org/10.1123/jtpe.11.2.195

McNamee, J., & van der Mars, H. (2005). Accuracy of momentary time sampling: A comparison of varying interval lengths using SOFIT. *Journal of Teaching in Physical Education, 24*, 282–292. https://doi.org/10.1123/jtpe.24.3.282

Mitchell, S. A., Oslin, J. L., & Griffin, L. L. (2013). *Teaching sport concepts and skills: A tactical games approach for ages 7 to 18* (3rd ed.). Champaign, IL: Human Kinetics.

NASPE Assessment Task Force. (2008). *PE Metrics: Assessing the standards, Standard 1 Elementary*. Reston, VA: Author.

National Association for Sport and Physical Education. (2002). *Authentic assessment of physical activity for H.S. students*. Reston, VA: Author.

National Association for Sport and Physical Education. (2010). *National standards and guidelines for physical education teacher education* (3rd ed.). Reston, VA: Author.

National Board for Professional Teaching Standards. (2014). *Physical education standards for teachers of students ages 3–18+* (2nd ed.). Retrieved from http://boardcertifiedteachers.org/sites/default/files/ECYA-PE.pdf

National Council on Teacher Quality. (2015). *2015 state teacher policy yearbook: National summary*. Retrieved from http://www.nctq.org/dmsView/2015_State_Teacher_Policy_Yearbook_National_Summary_NCTQ_Report

National Physical Activity Plan Alliance. (2016). *National physical activity plan*. Retrieved from http://physicalactivityplan.org/docs/2016NPAP_Finalforwebsite.pdf

Norris, J., van der Mars, H., Kulinna, P., & Beardsley, A. (2017). Administrators' perceptions of physical education teacher evaluation. *Physical Educator, 74*, 730–756. https://doi.org/10.18666/TPE-2017-V74-I4-7468

O'Sullivan, M., & Deglau, D. (2006). Chapter 7: Principles of professional development. *Journal of Teaching in Physical Education, 25*, 441–449. https://doi.org/10.1123/jtpe.25.4.441

Parker, M., & Patton, K. (2016). What research tells us about continuing professional development for physical education teachers. In C. Ennis (Ed.), *Routledge handbook of physical education pedagogies* (pp. 447–460). New York, NY: Routledge.

Pate, R. R., Davis, M. G., Robinson, T. N., Stone, E. J., McKenzie, T. L., & Young, J. C. (2006). Promoting physical activity in children and youth: A leadership role for schools. *Circulation, 114,* 1214–1224. https://doi.org/10.1161/CIRCULATIONAHA.106.177052

Pivovarova, M., Broatch, J., & Amrein-Beardsley, A. (2014). Chetty, et al. on the American Statistical Association's recent position statement on value-added models (VAMs): Five points of contention. *Teachers College Record, 2014.* Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17633

Pryor, J., & Akwesi, C. (1998). Assessment in Ghana and England: Putting reform to the test of practice. *Compare, 28*, 263–275. https://doi.org/10.1080/0305792980280304

Rink, J. (2007). PE teaching: It's ALL about outcomes [Editorial]. *PELinks4u, 9*(7). Retrieved from http://www.pelinks4u.org/archives/070107.htm

Rowe, P., van der Mars, H., Schuldheisz, J., & Fox, S. (2004). Measuring physical activity in physical education: Validating SOFIT for use with high school students. *Journal of Teaching in Physical Education, 23*, 235–251. https://doi.org/10.1123/jtpe.23.3.235

Sallis, J. F., McKenzie, T. L., Beets, M. W., Beighle, A., Erwin, H., & Lee, S. (2012). Physical education's role in public health: Steps forward and backward over 20 years and HOPE for the future. *Research Quarterly for Exercise and Sport, 83*, 125–135. https://doi.org/10.1080/02701367.2012.10599842

Saudargas, R. A., & Zanolli, K. (1990). Momentary time sampling as an estimate of percentage time: A field validation. *Journal of Applied Behavior Analysis, 23,* 533–537. https://doi.org/10.1901/jaba.1990.23-533

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 1066–1101). Washington, DC: American Educational Research Association.

Siedentop, D., Hastie, P. A., & van der Mars, H. (2011). *Complete guide to Sport Education* (2nd ed.). Champaign, IL: Human Kinetics.

Siedentop, D., & Tannehill, D. (2000). *Developing teaching skills in physical education* (4th ed.). Mountain View, CA: Mayfield.

Society of Health and Physical Educators. (2010a). *PE metrics: Assessing national standards 1-6 in elementary school* (2nd ed.). Champaign, IL: Human Kinetics.

Society of Health and Physical Educators. (2010b). *National standards and guidelines for physical education teacher education* (3rd ed.). Champaign, IL: Human Kinetics.

Society of Health and Physical Educators. (2011). *PE metrics: Assessing national standards 1–6 in secondary school.* Champaign, IL: Human Kinetics.

Society of Health and Physical Educators. (2014). *National standards and grade-level outcomes for K–12 physical education*. Champaign, IL: Human Kinetics.

Society of Health and Physical Educators. (2015a). *Updated standard 1 assessments for PE metrics ebook: Assessing standards 1–6 in elementary school*. Champaign, IL: Human Kinetics.

Society of Health and Physical Educators. (2015b). *Updated standard 1 assessments for PE metrics ebook: Assessing standards 1–6 in secondary school.* Champaign, IL: Human Kinetics.

Steffen, J., & Grosse, S. J. (2003). *Assessment in outdoor adventure physical education*. Reston, VA: National Association for Sport and Physical Education.

Stork, S. (2007). *Assessing gymnastics in elementary physical education*. Reston VA: National Association for Sport and Physical Education.

Test, D., & Heward, W. L. (1984). Accuracy of momentary time sampling: A comparison of fixed and variable-interval observation schedules. In W. L. Heward, T. Heron, D. Hill, & J. Trap-Porter (Eds.), *Focus on behavior analysis in education* (pp. 177–196). Columbus, OH: Merrill.

Thomas, A., & Stratton, G. (2006). What we are really doing with ICT in physical education: A national audit of equipment, use, teacher attitudes, support, and training. *British Journal of Educational Technology, 37,* 617–632. https://doi.org/10.1111/j.1467-8535.2006.00520.x

Tousignant, M., & Siedentop, D. (1983). A qualitative analysis of task structures in required secondary physical education classes. *Journal of Teaching in Physical Education, 3*, 47–57. https://doi.org/10.1123/jtpe.3.1.47

U.S. Department of Health and Human Services. (2008). *2008 physical activity guidelines for Americans.* Retrieved from http://health.gov/paguidelines/pdf/paguide.pdf

U.S. Department of Health and Human Services. (2010). Physical activity. Retrieved from https://www.healthypeople.gov/2020/topics-objectives/topic/physical-activity

U.S. Department of Health and Human Services. (2012). *Physical activity guidelines for Americans midcourse report*: *Strategies to increase physical activity among youth.* Retrieved from http://health.gov/paguidelines/midcourse/pag-mid-course-report-final.pdf

Vahey, P., & Crawford, V. (2003). *Learning with handhelds: Findings from classroom research.* Menlo Park, CA: SRI International.

van der Mars, H. (1989a). Basic recording tactics. In P. W. Darst, D. Zakrajsek, & V. H. Mancini, *Analyzing physical education and sport instruction* (2nd ed., pp. 19–53). Champaign, IL: Human Kinetics.

van der Mars, H. (1989b). Observer reliability: Issues and procedures. In P. W. Darst, D. Zakrajsek, & V. H. Mancini, *Analyzing physical education and sport instruction* (2nd ed., pp. 54–80). Champaign, IL: Human Kinetics.

van der Mars, H., & Harvey, S. (2010). Teaching and assessing racquet games using "Play Practice"—Part 2. *Journal of Physical Education, Recreation, and Dance, 81*(5), 35–43, 56. https://doi.org/10.1080/07303084.2010.10598478

van der Mars, H., Timken, G., & McNamee, J. (2018). Systematic Observation of Formal Assessment of Students by Teachers (SOFAST). *Physical Educator, 75,* 341–373. https://doi.org/10.18666/TPE-2018-V75-I3-8113

Veal, M. L. (1992). The role of assessment in secondary physical education: A pedagogical view. *Journal of Physical Education, Recreation, and Dance, 63*(7), 88–92. https://doi.org/10.1080/07303084.1992.10609932

Veal, M. L. (1995). Assessment as an instructional tool. *Strategies, 8*(6), 10–15. https://doi.org/10.1080/08924562.1995.10592045

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on achievement. *Assessment in Education, 11*(1), 49–65. https://doi.org/10.1080/0969594042000208994

Williams, L., & Rink, J. (2003). Teacher competency using observational scoring rubrics. *Journal of Teaching in Physical Education, 22*, 552–572. https://doi.org/10.1123/jtpe.22.5.552