2003

Teacher Competency Using Observational Scoring Rubrics

Lori Williams
University of South Carolina - Upstate

Judith E. Rink
University of South Carolina - Columbia, jrink@mailbox.sc.edu

Publication Info

Published in *Journal of Teaching in Physical Education*, Volume 22, Issue 5, 2003, pages 552-572.
http://journals.humankinetics.com/jtpe-contents

JOURNAL OF TEACHING IN PHYSICAL EDUCATION, 2003, **22**, 552-572

Chapter 5: Teacher Competency Using Observational Scoring Rubrics

Lori Williams
University of South Carolina – Spartanburg

Judith Rink
University of South Carolina – Columbia

The most recent wave of education reform is the standards, assessment and accountability movement (Walberg, Haertel, & Gerlach-Downie, 1994; Wolf, LeMahieu, & Eresh, 1992). The initial steps of this reform established national content standards in each of the content areas of the school program. The national standards define what students should know and be able to do in each content area and have been used by many states and districts either directly or to guide the development of state content standards. National content standards have created a shared vision of the role of each content area among professionals and the public (Fullan, 2001).

The development of standards has been followed by an increased emphasis on measuring the extent to which students meet the standards and holding teachers and schools responsible for students meeting those standards. The public, administrators, state and federal governments, and accrediting groups value assessment data as a means to evaluate students, programs, and reform efforts (Slavin, 2002; Walberg et al., 1994). Statewide assessments are used to provide evidence of student progress toward established standards and the quality of educational programs. Through high stakes assessment, teachers and programs are held accountable to the general public and policy makers for student achievement.

The establishment of content standards has also necessitated new strategies for examining and assessing educational outcomes. The standards, assessment and accountability movement relies heavily on the production of both valid and reliable data and has moved toward performance-based assessment. Performance assessment is tied to standards-based education and plays a key role in the reform movement (Marzano, Pickering, & McTighe, 1993). Performance assessment requires students to construct rather than select a response and has necessitated changes in the types of tasks we use to measure outcomes. In many cases, student achievement in more complex behavior is measured by means of observation and professional judgment (Stiggins, 1987). Most performance tasks are judged using observational records of some sort and most recently by the guidance of scoring

rubrics. There is some question whether observations of complex behavior can be done reliably (Linn, Baker, & Dunbar, 1991).

Work in physical education has followed the work of general education to the extent that content standards have been established. In 1995 the National Association for Sport and Physical Education (NASPE) published the landmark document, *Moving Into the Future: National Standards for Physical Education*. NASPE delineated seven content standards specifying what physically educated students should know and be able to do. The standards have provided consensus for what physical education students should learn and what teachers should teach. Also, the standards have provided a basis for student assessment and teacher accountability for student learning.

Unlike general education, physical education does not have a tradition of systematic evaluation or formal assessment of student performance (Bayless, 1978; Hensley, Lambert, Baumgartner, & Stillwell, 1987; Imwold, Rider, & Johnson, 1982; Kneer, 1986; Morrow, 1978; Veal, 1988). With the current education reform movement placing an emphasis on assessment and accountability, assessment is an area of concern for physical education. The educational climate is one that requires evidence of achievement (Slavin, 2002). The very survival of physical education in the public school system will depend, in part, on the methods we use to document student and teacher success (Wood, 1996).

The newer emphasis on performance-based assessment holds much promise for assessment in physical education particularly for the assessment of motor skills. Physical educators want to know if students can perform motor skills and use those motor skills in game and activity settings.

The viability of performance-based assessment as a formal measure of complex motor skill performance in physical education has not been investigated. As assessment plays a greater role in physical education, we need to know if teachers can accurately administer tests and assess student performance of complex behavior. Teachers' assessments form the greatest proportion of assessments experienced by students. What is not clear is whether teacher judgments are credible sources of reliable information for more formal and standardized testing. "To earn public confidence as a vital component (perhaps the basis) of a high-quality assessment system, teachers' judgments have to be supported by evidence and attain some comparability" (Mitchell, 1992, p. 133). The South Carolina Physical Education Assessment Program (SCPEAP) is one of the first statewide comprehensive assessment programs in physical education. SCPEAP relies heavily on teacher judgments in standardized observations and assessments of student performance.

The purpose of this study was to investigate the accuracy of teacher motor performance data submitted by teachers to SCPEAP. It was also the purpose of this study to describe teacher compliance with the testing protocols of the state assessment program, factors related to teacher's ability to use scoring rubrics accurately, and factors related to data compliance. Specific factors related to teacher accuracy and compliance included: (a) teacher gender, (b) training sessions attended, (c) Physical Education Institute (PEI) in-service sessions attended, (d) the activity assessed, and (e) the final student competency score given to a class.

Specific research questions guiding this study were: (a) Can teachers use scoring rubrics to accurately assess motor performance? (b) What factors affect teachers' ability to use scoring rubrics accurately? (c) What were the differences between teacher scores and monitoring committee scores? and, (d) What factors were related to data compliance?

In the SCPEAP program teachers videotape student performance in activity settings using standardized protocols for data collection (Appendix C). Teachers score student performance from the videotape using a standardized scoring rubric (Appendix C) and submit their scores and tape to the assessment program. A monitoring committee of teacher peers determines the compliance of teacher data with the protocols and accuracy of the teacher data in terms of agreement with the monitoring committee.

Methods

Class data sets submitted to SCPEAP by Cycle 1 high school teachers for the assessment of PI-1 (movement competence), assessment material generated by the monitoring committee, and SCPEAP records were analyzed to investigate teachers' ability to use scoring rubrics accurately to assess student motor performance and to follow testing protocols. Data were also analyzed to identify factors related to data accuracy and compliance.

Participants

Data for this study were drawn from all high schools completing the first round of the South Carolina state assessment program. There were 65 high schools designated as Cycle 1 high schools. These schools were geographically distributed throughout the state and represented a random selection of schools of different sizes and economic conditions. Sixty-two of these high schools submitted data to SCPEAP. One hundred fifty-nine (64 female, 95 male) physical education teachers teaching within these schools were required to submit assessment data. One hundred fifty-two (62 female, 90 male) teachers submitted assessment material for at least one movement form for PI-1 (movement competence).

Data Sources

Data were collected from a variety of sources. Data sources included Cycle 1 teacher assessment data submitted to SCPEAP in the Spring of 2001, monitoring committee assessment data, and SCPEAP records of unsubmitted data, teacher gender, training session attendance, PEI attendance, and final student competency scores for each class.

Teacher Assessment Data

Assessment data submitted by teachers included computer generated class roles, student exemption forms, videotapes of student performance, and summary score sheets of student scores done by the teacher. Of the 152 teachers submitting

data, 136 teachers submitted class data sets for two activities and 16 teachers submitted class data sets for only one activity. Seven teachers submitted no data for Performance Indicator-1 (movement competence). Overall, 288 class data sets were submitted for activities. The class data set was the unit of analysis for this study. By activity, these assessments included: basketball, 58; weight training, 47; badminton, 32; volleyball, 31; tennis, 30; softball, 13; flag football, 11; bowling, 10; golf, 9; soccer, 8; aerobic dance, 7; table tennis 8; track and field, 5; archery, 4; ultimate Frisbee, 5; square dance, 3; country western and line dance, 3; ballroom dance, 1; folk dance, 1; gymnastics, 1; and outdoor pursuits, 1.

Upon completion of videotaping student performance, teachers were to view the tape of student performance and assess each student's performance. Teachers used a four-point scoring rubric (see Appendix C for a sample rubric), which included qualitative criteria for performance at each of the four levels (0,1, 2, 3) to assess each student's level of performance. These scores were recorded on summary score sheets. A recorded score of "0" indicated little or rare demonstration of competence in the activity. A score of "1" indicated some evidence of competence. A score of "2" indicated the demonstration of competency in a movement activity and a score of "3" indicated a demonstration of performance above a level of competency in an activity. Individual student scores were averaged to determine a class score.

Monitoring Committee

The monitoring committee, which consisted of the High School Assessment Director of SCPEAP, Monitoring Committee Chairperson, K-12 physical education professionals and college/university faculty members, received the data submitted by the schools and worked in teams of four reviewers to determine its accuracy and compliance.

Determining data accuracy. Two reviewers from each team worked together to assess a 25% student sample, but no fewer than seven students, from each class of submitted scores to determine the accuracy of the teacher's data. A table of random numbers was used to establish which student scores on a teacher's summary score sheet would constitute the 25% sample. The two reviewers independently scored the selected random sample of students for a class and negotiated any differences in scoring between them. The accuracy of the teacher's data was determined by using simple percentage of agreement. If the teacher's scores and the two reviewer's scores reached an agreement level of at least 80% for the sampled data, the teacher's class data set for that assessment was accepted as reported. If the teacher scores and monitoring committee scores for a class data set did not reach an agreement of at least 80%, the reviewers took another 25% sample. If 80% agreement was reached for the 50% sample then the teacher's class data set for that assessment was accepted. If 80% agreement was not reached then the class data set was given to a second set of reviewers. The second set of reviewers pulled a 25% sample of students and followed the same procedures to determine agreement with the teacher's scores (e.g., going to a 50% sample if necessary). If an agreement of 80% was reached on either the 25% or 50% sample by the second set of reviewers, the data for that class were accepted as accurate. If agreement on the 50% sample was less than 80% the data were considered inaccurate and the data set received a score of 0 for that class.

Teacher accuracy in using the rubrics was determined using a discrimination of competency and noncompetency. Teacher scores of a 3 or 2 were considered representative of "competent" performance. Teacher scores of a 0 or 1 were considered "not competent" performance. This level of agreement was the standard the monitoring committee used to determine if a teacher's data would be accepted. This level of accuracy will be referred to as level one accuracy. The percentage of students competent in an activity (a student score of 3 or 2) was the class score used by SCPEAP to determine school scores.

Determining data compliance. Teachers were given and trained to use both general protocols for data collection and protocols written specifically for each activity. The team of two monitoring committee reviewers assigned to a class determined the compliance of the submitted data. Compliance refers to the adherence to both the general and activity specific data collection protocols. Class data sets were accepted or rejected according to the degree to which the data conformed to data collection protocols, in addition to the level of agreement. Class data sets in adherence to the protocols for data collection and assessment were accepted by the reviewers and deemed compliant. Part of the responsibility of the monitoring committee was to list any teacher violations of general or activity specific protocols on a log of data concerns for each class.

The class data set was accepted by the monitoring committee if the teacher scores were accurate and if there were no major violations of general or specific protocols. Class data sets with minor violations of protocols were accepted. Minor violations of protocols allowed the monitoring committee to assess the data for accuracy and did not affect the scores given classes. Minor violations of protocols included: no official roster included, students identified on the class roster but not on the videotape, teacher totaled scores for individual skills rather than giving an overall assessment of 0, 1, 2, or 3, or teacher provides instruction during assessment task.

Class data sets with major protocol violations were rejected by the monitoring committee as being noncompliant. These class data sets were in violation of data collection protocols to the extent that the monitoring committee could not assess the data. For example, if students were not introduced on the tape and not wearing a numbered pinnie/jersey, there was no way for the reviewers to identify students to obtain a 25% random sample. In addition, data were not accepted if a teacher's change in testing protocol would have affected student scores (e.g., using a volleyball trainer rather than an official volleyball).

The monitoring committee assessment data included monitoring committee score sheets that described the accuracy of the teacher's data and logs of data concerns. Monitoring committee score sheets contained scored samples of student performance and a percentage of agreement score that represented agreement between the monitoring committee and the teacher scores with the teacher data. A score sheet was completed for each activity assessed. The log of data concerns for each activity assessed contained monitoring committee members' concerns related to data compliance.

SCPEAP Records

Various SCPEAP records were used as data sources to identify factors that may be related to teachers' ability to use scoring rubrics accurately and teacher

compliance with data collection protocols. Data sources included records of teacher gender, training session attendance, Physical Education Institute (PEI) in-service session attendance, the records of data that teachers were required to submit, and the school reports designating class scores for a data set.

Teacher gender. Research has identified female teachers as more frequent users of formal assessments (Hensley et al., 1987; Imwold et al., 1982; Morrow, 1978). A study by Wirszyla (2002) also identified that male teachers may not be as supportive of the state physical education reform efforts. Teacher gender was identified in this study as a factor to investigate in relation to teacher's ability to use scoring rubrics accurately and teacher compliance with data collection protocols. SCPEAP records of teacher gender were used as the data source to identify the gender of each of the Cycle 1 teachers.

Training session attendance. Teacher training was selected as a factor potentially related to teacher accuracy and teacher compliance with data collection protocols. Several researchers have suggested the need for further study of the effect of training on rater reliability (Stuhlmann, Daniel, Dellinger, Denny, & Powers, 1999). SCPEAP records of teacher attendance at training sessions were used as a data source to identify the number of training sessions attended by each Cycle 1 teacher. A two-day training session, considered a full training session and a one-day make-up training session, also considered to be a full training session was provided for teachers.

Two teachers from each school were invited to attend the training sessions. Teachers who did attend the training session were instructed to assist teachers who did not attend the training at their school.

PEI attendance. The effect of Physical Education Institute (PEI) attendance on teacher's ability to use scoring rubrics accurately and teacher compliance with data collection protocols was selected for investigation. SCPEAP records of teacher attendance at the PEI sessions were used as a data source to identify the number of PEI sessions attended by each of the Cycle 1 teachers. There were 3 PEI sessions held for high school teachers the year of data collection.

School folders. The activity assessed was selected as a factor potentially related to teacher accuracy in the use of scoring rubrics and compliance with testing protocols. Differing levels of complexity inherent to various individual, dual, and team activities both in terms of the motor behavior observed as well as the complexity of the testing setting could affect observations. School folders were used as the data source to identify the activity assessed by each teacher. SCPEAP records included school folders for each of the Cycle 1 schools. Each folder was labeled with the school name and contained the assessment materials submitted by each teacher at that particular school, as well as monitoring committee score sheets and logs of data concerns for the activities assessed.

Final student competency score. Final student competency score was selected as a factor for investigation to determine if the skill level of the student is related to teachers' ability to accurately assess psychomotor performance. School folders were used as a data source to identify the number of students achieving competency for the activity in which they were assessed. Each school folder included SCPEAP records of final student competency scores for each school and each teacher by class.

Procedures

Levels of accuracy in the use of scoring rubrics and compliance with data collection protocols were established for each class data set submitted for PI-1 (movement competence). Factors related to teacher accuracy and compliance were identified by examining the relationship between accuracy and compliance and (a) teacher gender, (b) training sessions attended, (c) PEI sessions attended, (d) the activity assessed, and (e) final student competency score.

Accuracy

A comparison of teacher assessment data with the monitoring committee's data was used to determine teacher accuracy for each class data set submitted. Noncompliant/not accepted data were not scored for accuracy by the monitoring committee. Out of the 288 class data sets submitted, 74 class data sets were considered noncompliant/not accepted and were not included in the analysis of accuracy. Levels of accuracy were determined for 214 class data sets.

The scoring rubric used by teachers discriminated four levels of performance. The assessment program only needed information on whether students were competent or not competent at the activity. Therefore, the monitoring committee only assessed the accuracy of the teacher to discriminate competent or not competent performance (level one accuracy). Performance levels 0 and 1 were collapsed to represent an assessment of "not competent." Levels 2 and 3 were collapsed to represent an overall assessment of "competent." This level of accuracy, level one accuracy, was represented by the percentage of agreement score assigned by the monitoring committee. A mean score for percentage of agreement at level one accuracy was calculated for all classes scored by the monitoring committee and included in the database (n = 214).

In addition to the level of accuracy determined by the monitoring committee, issues of teacher accuracy were pursued to a higher level. Accuracy was determined for each class using all four levels of the rubric rather than collapsing the rubric into two levels. Level two accuracy represented a higher level of accuracy because agreement was determined by the performance level score using the four levels of the rubric rather than collapsed levels of competent and noncompetent. The researcher determined accuracy at this level by comparing the performance level score assigned to a student by the teacher to the performance level score assigned to a student by the monitoring committee. The same formula was used in determining level two accuracy as was used in determining accuracy at level one: agreements/total observations x 100. The level two percentage of agreement for each class data set was determined. A mean score for percentage agreement at level two accuracy was then calculated for each class.

Factors related to accuracy. To determine factors that affect teachers' ability to use scoring rubrics accurately an analysis was conducted to determine the relationship between teacher accuracy and: (a) teacher gender, (b) training sessions attended, (c) PEI sessions attended, (d) the activity assessed, and (e) final student competency score.

To determine the relationship between teacher accuracy and teacher gender, a 2 x 2 (accuracy level x gender) analysis of variance (ANOVA) was used to analyze teacher accuracy scores at level one and level two accuracy for females and males. An ANOVA was also used to determine if teacher accuracy was related to training. A 2 x 2 (accuracy level x training) ANOVA was used to analyze teacher scores at the two levels of accuracy and teacher attendance at training sessions. Due to the low number of teachers (n = 5) who attended one training session, teacher attendance at 2-day (full training) and 1-day (1/2 training) training sessions were collapsed into one level of training, representing those teachers who received training. The second level, no training, included teachers who did not attend training sessions (0 training).

A 2 x 2 (accuracy level x PEI attendance) ANOVA was used to analyze teacher scores at level one and level two accuracy and teacher attendance at PEI inservice sessions. The two levels of attendance at PEI sessions included no attendance (0 sessions) and attendance (1, 2, and 3 sessions). Due to low numbers of attendance at 2 (n = 17) and 3 (n = 3) PEI sessions, teacher attendance at 1, 2, and 3 PEI sessions was collapsed into one level, representing those who attended PEI sessions.

A final student competency score for each activity for each teacher was obtained from the school report. The percentage of a teacher's students for a class receiving a score of 2 or 3 for the activity assessed was recorded on the spreadsheet. Pearson r correlation coefficients were calculated to determine the relationship between percentage agreement scores at level one and level two accuracy and final student competency scores.

The activity assessed was also examined as a factor related to teacher accuracy. Mean scores were calculated for teacher accuracy at level one and level two for all activities that were accepted by the monitoring committee for assessment. Because of the low numbers of classes for some activities a descriptive analysis was conducted.

Differences in the Direction of Scores

To determine the differences between teacher scores and monitoring committee scores, the performance level scores assigned by the teacher and the performance level scores assigned by the monitoring committee for a class were compared for both accurate and inaccurate data submitted by the teacher. The percentage of competent scores (a score of 2 or 3) assigned by the teacher for a class was calculated. The percentage of competent scores assigned to a class by the monitoring committee was also calculated. A mean score was then calculated for the percentages of competent scores assigned to a class by the teacher and for the recorded percentages of competent scores assigned by the monitoring committee. Mean scores of the monitoring committee and the teacher were then compared.

To further investigate the differences in scores assigned by the teacher and the monitoring committee, disagreements at accuracy level two were analyzed. The percentage of disagreements due to the teacher assigning a higher performance level score was calculated. The percentage of disagreements that were due

to the monitoring committee assigning a higher performance level score was also calculated. A mean score was then calculated for the recorded teacher percentages and for the recorded monitoring committee percentages. Mean scores of the teachers and the monitoring committee were then compared.

Compliance

Three levels of compliance were established to describe the degree to which the teacher assessment data conformed to data collection protocols. The three levels of compliance included: compliant/accepted, noncompliant/accepted, and noncompliant/not accepted. Class data sets were assigned to one of these levels based upon the decisions of the monitoring committee as reported on the log of data concerns.

The first level of compliance, compliant/accepted, refers to submitted data that were consistent with data collection protocols established in the assessment materials and accepted by the monitoring committee. The second level, noncompliant/accepted, refers to submitted data that failed to observe all data collection protocols but the data were accepted by the monitoring committee for assessment. The third level, noncompliant/not accepted, refers to submitted data that did not adhere to major data collection protocols resulting in the data not being accepted by the monitoring committee. These data could not be assessed for accuracy due to the protocol violations impeding the assessment process as well as the impact the violations would have on student scores.

Each class of data submitted by a teacher was assigned to one of the three levels of compliance based on the decision log and final decision of the monitoring committee. The level of compliance: compliant/accepted, (C/A); noncompliant accepted, (NC/A); or noncompliant/ not accepted, (NC/NA), was recorded for each class by activity and teacher. The frequency of class data sets assigned to each of the three levels of compliance (C/A, NC/A, and NC/NA) was determined. The percentage of teachers with class data sets within each level of compliance was calculated.

Factors related to noncompliance. To determine factors related to teacher compliance with the testing protocols a descriptive analysis was conducted to determine the relationship between the level of compliance (C/A, NC/A, and NC/NA level) and (a) teacher gender, (b) training session attendance, (c) PEI attendance, and (d) the specific activity assessed. To determine if teacher gender was a factor related to data compliance, the frequency and percentage of C/A, NC/A and NC/NA data submitted by female and male teachers were calculated. Data were analyzed descriptively.

In examining the relationship between compliance and teacher training, the frequency and percentage of C/A, NC/A, and NC/NA data submitted by teachers who received no training and those who received training were determined. A descriptive analysis was conducted.

In determining if PEI attendance was a factor related to data compliance, the frequency and percentage of class data sets within each level of compliance submitted by teachers attending 0 and those attending 1, 2, and 3 PEI in-service sessions were calculated. Descriptive analysis was used to analyze data.

Levels of compliance were also examined by the specific activity that was assessed. The frequency and percentage of class data sets within an activity that were accepted for assessment (C/A or NC/A), and not accepted for assessment (NC/NA) were calculated. A descriptive analysis was conducted.

Issues related to noncompliance. Logs of data concerns completed by the monitoring committee for each activity were used to identify issues related to noncompliance. Concerns related to noncompliance were most often noted at the NC/A level as well as the NC/NA level. Comments recorded on the logs were identified, along with the teacher name and activity assessed, and then entered onto a spreadsheet. The researcher then coded these comments into categories, with each category having rules of inclusion. The categories describing noncompliance were: documentation, student identification, scoring, modification of the assessment task, and technical problems. The frequency and percentage of incidents occurring within each category were calculated. Data describing issues of noncompliance were analyzed descriptively.

Results

Results of the study are reported first for teacher accuracy using the scoring rubrics and factors related to accuracy and second for teacher compliance with the testing protocols and factors related to teacher compliance.

Accuracy

The accuracy with which teachers used scoring rubrics to assess student performance was determined at two levels. The percentage of agreement assigned by the monitoring committee, for all class data sets submitted and scored for accuracy for level one (collapsed rubric) accuracy ($n = 214$) was 86.89% ($SD = 17.60$). The mean score for level two accuracy (using all scores on the rubric), a higher level of accuracy, was 67.89% ($SD = 23.48$).

To determine differences between teacher scores and monitoring committee scores the percentage of competent scores assigned by both were calculated. The mean score for the percentage of competent scores assigned by the teacher was higher ($M = 72.22\%$, $SD = 25.94$) than the mean score for the percentage of competent scores assigned by the monitoring committee ($M = 67.31\%$, $SD = 29.22$).

In further investigation of the differences between scores assigned by the teacher and the monitoring committee, disagreements at accuracy level two were analyzed to determine the direction of the differences between the teacher scores and the monitoring committee scores. At level two accuracy, a larger percentage of disagreements between the monitoring committee and the teacher were due to the teacher assigning higher scores ($M = 58.40\%$, $SD = 45.29$) for student performance than the monitoring committee assigning higher scores ($M = 26.68\%$, $SD = 39.73$) for student performance. Teacher gender, training session attendance, PEI attendance, final student competency score, and the activity assessed were examined to determine factors related to teacher accuracy.

Teacher gender. Accuracy scores for females (n = 59) and males (n = 68) who submitted class data sets that were accepted by the monitoring committee to assess for accuracy were analyzed. ANOVA revealed no significant difference, $F(1, 213) = .358$, $p > .05$, between accuracy scores of females (M = 86.16, SD = 18.74) and males (M = 87.60, SD = 16.46) at level one accuracy. There was also no significant difference, $F(1, 213) = .020$, $p > .05$, between level two accuracy scores of females (M = 68.12, SD = 24.70) and males (M = 67.67, SD = 22.33).

Teacher training. ANOVA revealed no significant difference, $F(1, 213) = .370$, $p > .05$, between accuracy scores of trained teachers (M = 86.32, SD = 18.33) and untrained teachers (M = 87.83, SD = 16.40) at level one accuracy. There was also no significant difference, $F(1, 213) = 2.458$, $p > .05$, between level two accuracy scores of trained teachers (M = 69.85, SD = 23.99) and untrained teachers (M = 64.68, SD = 22.40).

Table 1 Mean Scores of Percentage Agreement for Levels of Accuracy by Activity

	Accuracy			
	Level 1		Level 2	
Activity	*M*	*SD*	*M*	*SD*
Ballroom dance	100.0		57.0	
Archery	100.0	.00	66.7	21.8
Square dance	100.0	.00	47.7	21.4
Gymnastics	100.0		43.0	
Aerobics	96.0	6.8	67.4	23.0
Table tennis	95.3	12.5	86.1	20.2
Country western & line dance	94.3	9.8	84.3	16.6
Bowling	93.1	13.1	80.0	19.8
Weight training	92.7	10.1	63.3	25.6
Flag football	92.1	10.4	71.0	12.2
Soccer	90.2	12.1	75.3	13.5
Basketball	89.4	14.1	73.2	23.2
Track & field	87.0	1.4	80.5	7.8
Folk dance	86.0		86.0	
Ultimate frisbee	83.8	12.5	53.3	6.9
Badminton	83.6	18.1	69.0	23.8
Golf	80.6	16.4	55.6	14.6
Tennis	80.0	22.7	60.6	25.8
Softball	78.7	15.3	61.3	23.5
Volleyball	77.6	26.5	66.5	26.8
Outdoor pursuits	–	–	–	–

Note. Outdoor pursuits was not accepted for assessment of data accuracy.

PEI attendance. ANOVA revealed no significant difference, $F(1, 213) = .286$, $p > .05$, between level one accuracy scores of teachers who attended PEI sessions ($M = 87.66$, $SD = 16.65$) and teachers who did not attend PEI sessions ($M = 86.38$, $SD = 18.25$). There was also no significant difference, $F(1, 213) = 2.167$, $p > .05$, between level two accuracy scores of teachers who attended PEI sessions ($M = 70.68$, $SD = 24.24$) and teachers who did not attend ($M = 66.05$, $SD = 22.85$).

Student competency score. The correlations between final student competency scores of a class and level one ($r = .652$, $p < .01$) and level two accuracy ($r = .382$, $p < .01$) of the teacher were both significant. However, the correlation between final student competency score and level one accuracy was higher.

Activity. Mean scores were calculated for accuracy at level one and level two for all activities that were accepted by the monitoring committee for assessment. Table 1 displays the mean scores for levels of accuracy for all activities. Mean scores for level one accuracy were greater for individual activities, such as ballroom dance, archery, square dance, gymnastics, and aerobics. However, level two accuracy mean scores for these same activities fell well below an acceptable level of 80%. With the exception of softball and volleyball mean scores for level one accuracy were 80% or higher for all activities. Outdoor pursuits was not accepted for assessment of data accuracy.

Mean scores for level two accuracy were 80% or greater for several individual and dual activities. Table tennis, country western and line dance, bowling, track and field, and folk dance had mean scores greater than 80% at both level one and level two accuracy.

Compliance

The levels of compliance investigated for the 152 teachers and 288 class data sets were compliant/accepted (C/A), noncompliant/accepted (NC/A), and

Table 2 Percentage of Class Data Sets Within Levels of Compliance by Teacher Gender, Training Session, and PEI Attendance

	Level of compliance		
	C/A	NC/A	NC/NA
Teacher gender			
Female	13.9	73.0	13.1
Male	6.0	59.0	34.9
Training session attendance			
No attendance	6.7	60.8	32.5
Attendance	11.3	67.9	20.8
PEI Attendance			
No attendance	4.2	64.0	31.7
Attendance	19.2	66.7	14.1

noncompliant/not accepted (NC/NA). Descriptive summaries revealed that 27 class data sets (9%) were identified as C/A, 187 class data sets (65%) were identified as NC/A, and 74 (26%) class data sets were identified as NC/NA. Sixteen per cent of the teacher's submitted data that was identified as C/A. Eighty percent of the teachers had one or more class data sets at the NC/A level, and 37% of the teachers had one or more class data sets at the NC/NA level. Seventy-four percent of the data was ultimately accepted (C/A and NC/A) including some or all of the data for 96% of the teachers.

Variables investigated related to data compliance included teacher gender, training session attendance, PEI attendance, and the specific activity assessed. Levels of compliance within each of these four variables were examined and are reported in Table 2.

Teacher gender. Descriptive summaries were calculated for females (n = 62) and males (n = 90) with class data sets belonging to each of the 3 levels of compliance. A larger percentage of compliant and accepted data were submitted by females than by males (13.9% females, 6.0% males). A larger percentage of accepted data with minor protocol violations were submitted by females than by males (73.0% females, 59.0% males). Males submitted a larger percentage of data that was not accepted by the monitoring committee to assess for accuracy (13.1% females, 34.9% males).

Teacher training. Table 2 displays descriptive summaries for the number of training sessions attended by each teacher (n = 152) and the level of compliance of the class data they submitted. Teachers who attended training sessions submitted the largest percentage of class data sets that were accepted by the monitoring committee to assess for accuracy. The monitoring committee accepted 79.2% of the class data sets submitted by trained teachers and 67.8% of the class data sets submitted by untrained teachers. Trained teachers also submitted a larger percentage of compliant/accepted data than those that did not attend training sessions (11.3% trained, 6.7% untrained).

PEI attendance. A larger percentage of class data sets submitted by teachers who attended PEI sessions (n = 50) was compliant (19.2%) and accepted (66.7%) in comparison to the percentage of class data sets submitted by those who did not attend (see Table 2). Compliant data were submitted by 4.2% of the teachers who did not attend PEI sessions (n = 102) and 64% of the class data sets submitted by these teachers was accepted. Teachers who did not attend PEI sessions also submitted a larger percentage of class data sets that were not accepted for assessment (31.7%) than teachers who attended PEI sessions (14.1%).

Activity. Class data sets were submitted for 21 different activities. Levels of compliance were examined by activity. Table 3 displays the percentage of class data sets within an activity that were accepted and/or not accepted for assessment. Accepted class data sets include those that were compliant (C/A) and those with minor protocol violations (NC/A). Noncompliant/not accepted (NC/NA) data were not accepted for assessment.

Individual activities, such as the various dance forms (i.e., aerobic, ballroom, country western and line, and square) and gymnastics, had the highest percentages of compliance and acceptance (see Table 3). Dual activities, such as tennis

Table 3 Compliance by Activity

Activity	% of Class data sets accepted	Frequency accepted	% of Class data sets not accepted	Frequency not accepted
Aerobics	100.0	7		
Ballroom dance	100.0	1		
Country western & line dance	100.0	3		
Folk dance	100.0	1		
Gymnastics	100.0	1		
Square dance	100.0	3		
Tennis	90.0	27	10.0	3
Golf	88.9	8	11.1	1
Table tennis	87.5	7	12.5	1
Volleyball	83.8	26	16.1	5
Basketball	82.7	48	17.2	10
Flag football	81.8	9	18.2	2
Ultimate frisbee	80.0	4	20.0	1
Archery	75.0	3	25.0	1
Badminton	71.9	23	8.1	9
Bowling	70.0	7	30.0	3
Weight training	51.0	24	48.9	23
Soccer	50.0	4	50.0	4
Softball	46.2	6	53.8	7
Track & field	40.0	2	60.0	3
Outdoor pursuits			100.0	1

and table tennis, were also among the activities with high percentages of acceptance. Several team activities, such as volleyball, basketball, flag football, and ultimate Frisbee, had an acceptance rate of at least 80%.

Some activities were almost equally divided with approximately 50% of the class data sets accepted and 50% not accepted (see Table 3). These activities included weight training, soccer, and softball. Outdoor pursuits and track and field were the two activities with the lowest percentage of compliance and acceptance. These were also low frequency activities, which means that only a few classes in these activities were included in the database.

Issues related to noncompliance. The monitoring committee identified the issues related to noncompliance for each of the class data sets that were categorized as either scoring, documentation, technical, modification of the assessment task, or student identification. These categories are described with the percentage of the noncompliant class data sets that were in violation of each incident in Table

Table 4 Categories and Percentage of Incidents of Noncompliance

Category	Percentage of incidents
Scoring	
Not all students on roster were assessed	29.2
No scores for student on videotape	27.1
Teacher totaled columns for overall score	13.2
Teacher averaged individual skill scores and did not report overall score as a whole number	6.6
Cognitive test scores were omitted from score sheet	5.9
Teacher did not give an overall score	5.6
Teacher did not use lower score of test/lift	3.5
Teacher reported overall score as a range	1.4
Incorrect scoring of cognitive test	.7
Teacher gave 0 for skills not observed	.7
Did not adhere to scoring rubric	.3
Documentation	
No Form A's for exempt students	35.1
No official roster	18.8
Roster/score sheet mismatch	8.3
Teacher did not send cognitive tests	4.2
No score sheet	1.4
Sent optional movement form	1.0
Technical	
Poor camera placement or angle	13.5
Could not hear names to identify students	10.8
Camera placed too far away	4.5
Court or lane markings not visible	2.1
Did not use VHS tape	2.1
Camera follows the ball	1.4
No video	1.0
Videotape broken upon receipt	.7
Modification of the Assessment Task	
Protocol was not followed	18.1
Teacher coaches or provides instruction	5.2
Numbers are not on the front and back of pinnies/jerseys	2.8
Student numbers were duplicated	2.4
Student number on video did not match number on score sheet	1.4
Students not wearing pinnies/jerseys	1.0
Student number changed after initial identification	1.0

4. The majority of noncompliant class data sets included more than one incident of noncompliance.

The category identified as scoring included the highest total of incidents of noncompliance (94.2% of the class data sets submitted) compared to the other categories. The scoring category represented problems teachers had with establishing and recording an overall student performance score of 0, 1, 2, or 3 for all students. The noncompliant incidents that occurred most often were not assessing all of the students on the roster and not reporting scores for students who appeared on the videotape. Within the scoring category, the process used to report the overall student score, such as totaling (13.2%) or averaging (6.6%) columns of individual skill scores, was also a frequent incident of noncompliance.

The documentation category included incidents related to not supplying documents or documenting information. Sixty-nine percent of the submitted data had noncompliance issues related to documentation problems. The documentation category of compliance issues contained the highest percentage for any single incident of noncompliance within a category. Teachers were required to submit a Form A exemption form for each student who was not tested but who did appear on the class roster. Student exemption forms were missing from 35.1% of the noncompliant class data sets. Official school rosters were also frequently left out of assessment materials. In many cases, the students on the official roster did not match the students on the score sheet.

The technical category of compliance issues included incidents of noncompliance that affected the quality of the videotape (see Table 4). The monitoring committee made every attempt to work with the tape sent by the teacher. However, if it was not possible to see student performance on the tape, data were considered noncompliant. Thirty-six percent of the submitted data had noncompliance issues related to technical problems. Most often the camera placement or angle was poor and did not allow for the viewing of the entire assessment area resulting in students not always being in view. For example, students could not always be seen serving during the volleyball or tennis assessment tasks. Poor sound quality, sometimes combined with background noise, often prevented the monitoring committee from hearing student names and numbers during introductions. There were several incidents of the camera being placed too far away to identify student numbers and to assess the skill level of the performance.

Teachers modified the assessment task for 23.3% of the data. Most of the teacher modifications of the assessment task beyond the protocols involved the teachers coaching students or providing instruction during the assessment task. There were 15 incidents of coaching or instruction occurring while students were engaged in the assessment task, which was also in violation of the protocols.

The compliance category of student identification included events of noncompliance that made the identification of students difficult for the monitoring committee (see Table 4). Twenty-three percent of the submitted data had noncompliance issues related to problems identifying students. Most often, students did not introduce themselves on the videotape by giving their first and last name and

pinnie/jersey number. There were also several incidents of student numbers being omitted from the score sheet, pinnies/jerseys not numbered on the front and back, and student numbers being duplicated or used more than one time. There were few incidents of students not wearing pinnies/jerseys or of student numbers changing after the introduction.

Discussion

The results of this study will be discussed first in terms of the findings related to data accuracy, data compliance, and the role of the specific activity in relation to both accuracy and compliance. The section ends with a discussion of the implications of the results of this study for practice.

Data Accuracy

The most significant finding of this study was that teachers could use scoring rubrics to accurately assess student motor performance when making a judgment about competent or not competent student performance. The mean score for percentage agreement for all classes (86.89%) between teacher and monitoring committee scores was sufficiently high to consider teachers as reliable observers and scorers. Furthermore, teachers could provide accurate assessments of student performance with minimal training. Limited training time did not allow for instruction in the use of assessment materials for all 21 possible activities that teachers could assess. It is unlikely that training programs will be able to train teachers specifically in every activity they will be expected to assess. Knowing that teachers can use good assessment materials reliably with minimal training should provide support for performance-based assessments requiring teacher observation.

It was more difficult for teachers to score accurately using four performance levels (a score of 0, 1, 2, or 3 on the rubric) than to discriminate between noncompetency (a score of 0 or 1 on the rubric) and competency (a score of 2 or 3 on the rubric). Teachers may be able to discriminate four levels of competency with additional experience and training. For this assessment program teachers used the four point scoring rubric but were only required to discriminate between competent and noncompetent performance.

Another significant finding of this study was the effect of skill level on teacher accuracy in scoring students. Teachers who were more accurate in scoring also scored students who were more competent. One possible explanation for this finding is that it is less difficult to score competent students. Another possible explanation is that teachers who are accurate assessors are also better teachers and produce more competent students.

Skillful observation is a critical instructional skill. Teachers who are more skilled observers have a greater ability to select and sequence a progression of tasks, taking students from one level of performance to the next. When observing student responses during movement tasks, good observers know what to look for. Skillful observation not only promotes accurate assessment of student performance,

but also provides information from which teachers can make decisions about what students are doing. This information is essential to teaching an activity. Clearly, training and expecting teachers to observe student performance has great potential for improving instruction and student performance.

This study did not find teacher training to be a significant factor related to teachers' ability to use scoring rubrics to accurately assess student motor performance. All schools (maximum of two teachers) were represented at the data collection training but not every teacher. Many teachers including those who did not attend the training session were accurate in assessing student performance. Teachers who did attend the training session were instructed to assist teachers at their school who did not attend the training. There are several possible explanations for the idea that teachers could use the assessment materials effectively. One is that the assessment materials, and support materials such as the CD that provided teachers with examples of performance for each activity and each level, were appropriately constructed to allow teachers to easily follow the protocols for implementing the assessment program. Another possible explanation is that the trained teachers were very effective in assisting other teachers at their school. Training may not have been a significant factor because the trained teachers did not receive training in all 21 activities. There may be generic skills or commonalities throughout the assessment program that enable the effective use and administration of the various assessments with good teacher-friendly materials.

Teacher gender was also not found to be significantly related to accuracy in scoring student performance. This is good news for this program. An initial study conducted after the state legislation specified the performance indicators and before the accountability and assessment piece of the program was in place, showed a marked difference in male and female support of the program (Fleming, 1998; Wirszyla, 2002). The fact that there were no differences in accuracy may mean that the accountability aspect of the program has been successful in beginning to bridge the gender gap.

The assessment program in South Carolina is dependent upon teachers being able to score student performance accurately. There have been few attempts to do formal testing on a wide scale with psychomotor performance, largely because there have not been permanent products of performance. When video recording is used, the assessment is dependent upon the ability of the observer to accurately use a set of criteria for making a judgment about performance. Having teachers make those judgments has advantages, not the least of which is that teachers and students get immediate feedback on the level of performance.

The major contribution of this study is the knowledge that accurate data can be collected by teachers. It is unclear whether these high levels of teacher accuracy could have been achieved without the accountability for accuracy provided by the monitoring committee. One suspects that accountability played a major role in focusing teachers on submitting accurate data (Walberg et al., 1994). It is also worth projecting that even higher levels of accuracy can be achieved with revisions in some of the testing materials for activities that did not have high levels of accuracy. It is also conceivable that levels of accuracy will improve as teachers use the materials beyond this first data collection.

Data Compliance

Other important findings of this study were related to teacher compliance with the testing protocols. The majority of teachers (96%) submitted assessment data for PI-1 (movement competence) within a level of compliance that was accepted by the monitoring committee for the assessment of data accuracy. However, within this majority, a larger percentage (80%) of teachers violated some aspect of the data collection protocols. Minor protocol violations were common. Student exemption forms were often omitted from class data sets and all of the students on the class list were not assessed. On the one hand this is discouraging and on the other hand encouraging. It is discouraging, because as assessment and accountability play greater roles in physical education, teachers need to be able to provide documentation and/or assessment scores for all students in their classes. These findings are encouraging because even with this being the first round of the state assessment program in which teachers were to collect and evaluate assessment data, the majority of the teachers submitted data that were accepted by the monitoring committee for the assessment of data accuracy. With experience, teachers should be expected to reduce the protocol violations. With increased accountability for compliance with the protocols teachers will also increase their level of compliance.

Two factors were identified that were related to teacher compliance, teacher training and gender. Teachers who received training submitted the largest percentage of compliant data as well as the largest percentage of data accepted by the monitoring committee for assessment. Teachers who attended training sessions received instruction on how to collect, evaluate, and submit assessment data. Teachers were also provided with hands on experience in setting up and videotaping assessment tasks. Teachers who did not come to training sessions had to use the written and CD materials provided for this purpose or get help from a teacher who did attend the training session. Teachers who attended training sessions also submitted the smallest percentage of noncompliant/not accepted data.

Teacher gender was also related to data compliance. Female teachers submitted a larger percentage (13.9%) of compliant data than males (6.0%), as well as a larger percentage (73.0%) of accepted data with minor protocol violations than males (60.8%). Female teachers may be more familiar with assessment protocols in general as past research has revealed that female teachers are more frequent users of formal assessments (Hensley et al., 1987; Imwold et al., 1982; Morrow, 1978). Females may also be more supportive of the assessment program (Wirszyla, 2002).

The Role of the Specific Activity

The activity assessed by the teacher was related to teacher accuracy and data compliance with testing protocols. Individual activities, such as ballroom dance, aerobics, country western and line dance, square dance, and gymnastics, were most often accurately assessed at level one accuracy and had the highest percentages of compliance. However, few class data sets were submitted for these activities and

the low frequency rate may have influenced mean accuracy scores and percentage of acceptance.

In comparison, team sports often had lower accuracy scores and percentage of compliance than individual activities. Observation becomes more difficult as the number of individuals, the amount of equipment, and space increase. Therefore, the testing protocols for team sports, such as softball, soccer, ultimate Frisbee, and flag football, become more complex and more difficult to observe than for individual activities such as aerobic dance, which is performed in personal space with a stationary step. Also, to allow viewing of movement within the playing area, the camera is placed further away from student performance for the taping of team activities than for individual activities. This increase in distance may not allow observers to see student performance as clearly during team activities.

Second to basketball, weight-training assessments were among the highest frequency of activities submitted. Approximately half of the weight-training assessments were accepted and half were not accepted. This is possibly due to the cognitive test component of the weight training assessment. Teachers were to administer and evaluate a cognitive test, supplied by SCPEAP, for weight training. The written test score for each student was to be included in the overall scoring of student performance. The cognitive test component of weight-training was problematic for teachers as tests were often incorrectly scored, not sent with assessment materials, or not correctly considered in the overall student score.

Implications for Practice

Historically, assessment has not been a regular practice in physical education. Physical education has most often been left out of efforts at all levels to document student achievement. The lack of assessment and accountability in physical education has contributed to poor programs and low student performance. State level assessment of physical education has potential for improving the quality of instruction and student performance in physical education. Assessment at this level relies heavily on the production of reliable data. Results of this study indicate that teachers can use performance-based assessments to accurately assess student performance and produce reliable data. Teachers can provide data that are an accurate reflection of student performance on indicators of standards when they are held accountable for the quality of that data.

Teacher involvement enhances the assessment process. Not only should teachers be involved with training and the use of assessment materials, but also with assessing their students' achievement. When teachers are involved in the assessment process their understanding of exactly what students should know and be able to do becomes clear. Teacher involvement can also promote the construction and quality of assessment materials to be used in statewide assessment. In South Carolina, state level assessments were constructed by teachers and administered by teachers. Teachers also collected and submitted data that were accurate and reliable.

In the South Carolina Physical Education Assessment Program teachers are held accountable for producing accurate data. Classes for which teachers do not

submit accurate data receive a zero for student performance. The role of the monitoring committee is critical to developing accountability for both data accuracy and data compliance. It is unlikely that high levels of teacher accuracy or compliance could have been achieved without the accountability part of the program. The work of the monitoring committee is a very time consuming and expensive part of the state program but essential to its success.

This is the first study to investigate teacher use of scoring rubrics to accurately assess student motor performance in physical education on a large scale. More work is needed to validate reliable teacher use of scoring rubrics to assess psychomotor performance. In particular, future studies are needed to examine the relationship between teacher accuracy and the complexity of the movement form. It is also essential that we know what kinds of materials and training best facilitate the effective use of assessment materials.