The Physical Educator Vol. 72 • pp. 179–191 • 2015

TEST AND MEASUREMENT

Addressing Educational Reform: Exploring PE Metrics as a System to Measure Student Achievement in Physical Education

Glenn Hushman, Carolyn Hushman, Kira Carbonneau

Abstract

The current educational reform movement in the United States is focused on measuring the effectiveness of teachers. One component of teacher effectiveness is student achievement. The effectiveness of using PE Metrics as a measure of student achievement in a physical activity setting with a low socioeconomic, culturally diverse population was examined in this study. Two raters scored second and fifth grade children (N = 90) on skills using the PE Metrics system. Skills assessed included skipping, galloping, dribble with hand jog, basketball, soccer, and overhand throw. An item analysis showed that teachers could use the majority of the items to discriminate skill level. Strengths and weaknesses of the PE Metrics system were discussed.

The lack of student achievement as measured by standardized tests has resulted in a renewed interest on the practices teachers employ to promote student learning. For more than two decades,

Glenn Hushman is an assistant professor, Department of Health, Exercise, and Sports Sciences, University of New Mexico. Carolyn Hushman assistant professor, Department of Individual, Family, and Community Education, University of New Mexico. Kira Carbonneau is post-doctoral fellow, College of Education, University of New Mexico. Please send author correspondence to ghushman@unm.edu

educational reform has been a topic of interest on a national level. Policy makers have implemented many systems to assess what students are learning, the most publicized model being No Child Left Behind (NCLB, 2002). An important finding from these systems of assessment is that "teachers are the fulcrum that determines whether any school initiative tips towards success or failure" (Darling-Hammond, 2009, p. 1). The best school predictor of student outcomes is teaching that is informed by high-quality practices, effective delivery of information, and quality systems to assess student achievement (Goldhaber, 2002; Wright, Horn, & Sanders, 1997). Researchers have supported this notion in recent years, such as Clotfelter, Ladd, and Vigdor (2007), who found quality teachers had a greater impact on achievement gains of students than parents and race combined.

Although systems are in place to observe teacher performance and student achievement, most school district and state government officials cannot pinpoint what differentiates a high-quality teacher from a low-quality teacher. This is due to an inefficient and unreliable methodology of how teacher assessment is conducted. For example, in the 2009 New Teacher Project study titled *The Widget Effect*, Weisberg, Sexton, Mulher, and Keeling found that the outcome of most evaluation systems is either satisfactory or unsatisfactory. Having only two final outcomes for an assessment is troubling, but what is more concerning is that 99% of teachers that are assessed with such systems earn a *satisfactory* rating. Weisberg et al. also suggested that when these teacher evaluation systems are linked to student achievement, they are only linked to standardized test scores, which sheds little light on theories and practices employed to measure teacher effectiveness. Even if these teacher evaluation models were valid, reliable, and fair, they still may not be used to assess teacher performance in unique settings such as physical education (PE), art, and music. Reform will need to include all teachers in all subject areas in the educational system for fairness. Therefore, this is an indication to begin the process of developing an assessment system to measure teacher effectiveness at all levels in PE. Measuring teacher effectiveness includes observation of teacher performance and documentation of student achievement. However, developing a system to measure student achievement is an important step in developing a system of teacher effectiveness in PE environments.

National Association of Sport and Physical Education

In 1983, the National Commission on Excellence in Education published a report titled *A Nation at Risk*, claiming the United States was becoming "academically obsolete" compared to other nations. The result was a nationwide effort to develop a system of standards that could be implemented to measure knowledge acquisition of school-aged students in specific academic subjects. The National Association for Sport and Physical Education (NASPE, 1995) responded to this call for reform by developing the first national content standards for PE. The framework for these standards is based on skill acquisition, knowledge development, and affective elements students would need to exhibit to stay physically active for a lifetime (see Table 1). However, although it has been suggested following these standards will lead to high-quality PE (Lambert, 2007), the numerous barriers of limited time allocation, low subject status, and inadequate resources have limited the application of standards to many PE settings (McKenzie & Lounsbery, 2009).

Table 1
National Standards for Physical Education

Standard 1: Demonstrates competency in motor skills and movement patterns needed to perform a variety of physical activities.
Standard 2: Demonstrates understanding of movement concepts, priniciples, strategies and tactics as they apply to the learning and performance of physical activities.
Standard 3: Participates regularly in physical activity.
Standard 4: Achieves and maintains a health-enhancing level of physical fitness.
Standard 5: Exhibits responsible personal and social behavior that respects self and others in physical activity settings.
Standard 6: Values physical activity for health, enjoyment, challenge, self-expression, and/or social interaction.

Note. From *Moving Into the Future: National Standards for Physical Education* (2nd ed.), by National Association for Sport and Physical Education, 1995, Reston, VA: Author.

Although content standards for PE have been developed in all 50 states, NASPE standards have not been fully adopted on a national level. Regardless of the challenges NASPE officials have faced implementing the content standards on a systematic level, they have moved forward with a system to measure student achievement as informed through the NASPE PE content standards.

Student Achievement in Physical Education

Students have traditionally been graded in PE on aspects other than ability (Johnson, 2008). For example, many assessment plans include awarding points for attendance, participation, and dressing appropriately. When teachers grade on elements other than motor ability, it is communicated to students that only showing up for class, displaying appropriate behavior, and the appearance of effort are needed to pass PE (Melagrano, 2007). In the current evidence-based educational environment, there is a need for documentation of student achievement on mastering content standards and less emphasis on managerial tasks such as attendance. Documentation of student achievement needs to be grounded in evidence of learning within the realms of cognitive, psychomotor, and affective domains. Therefore, such managerial tasks must be replaced by more legitimate factors for grade calculation if student achievement is to be measured on true learning objectives (Stiggins, 2001).

Most PE programs currently do not have an assessment system that results in valid and reliable scores with which to measure student achievement across a district or state. Therefore, district officials and teachers lack empirical data to measure the effectiveness of instruction. Thus, assessment measures that align with a national set of content standards will be needed to prove student learning occurs in PE environments.

PE Metrics

In 2000, a standards-based assessment of cognitive and psychomotor skills was developed by NASPE (2001) as an effective tool to measure change or growth in students' abilities in PE settings. This system of assessment was intended to be used to measure student achievement as based on the NASPE PE standards. The result was a published series of assessments called PE Metrics. Through a system of assessments, PE Metrics is used to measure competency in motor skill and movement patterns, understanding of movement concepts, understanding of characteristics of a physically active lifestyle, and knowledge of social responsibility as it relates to physical activity (NASPE, 2012). Fissette and Franck (2013) highlighted PE Metrics as a system that may be used to develop formative assessments to determine student growth. The use of PE Metrics as the primary assessment system in PE would result in pre- and postmodels of assessment, which teacher will be able to use to measure student achievement.

Although NASPE standards may be measured with PE Metrics, competence in psychomotor skills and movement patterns, as described in Standard 1, is the focal point of the assessment series. Standards 2 to 6 are focused on the cognitive domain, and students are tested through multiple-choice test questions. The assessment developers have provided empirical support for the reliability and validity of the scores from PE Metrics (Dyson et al., 2011; Fox et al., 2011; Zhu, Fox, et al., 2011; Zhu, Rink, et al., 2011). However, a replication of these results within different populations has yet to be published.

Although PE Metrics may be used to measure teacher quality and student achievement (NASPE, 2012), further research on PE Metrics must be conducted before this system can be confidently used as one part of measuring the effectiveness of teachers. One reason for this further research is the changing cultural climate in the United States. It is predicted the United States will be a predominantly majority-minority country by the year 2043 (U.S. Census Bureau, 2012). Because race and ethnicity are often determinants of a person's socioeconomic status, minorities often reside in low socioeconomic areas (House & Williams, 2000). Therefore, it is important to assess whether a system that measures student achievement will perform well with a low socioeconomic, culturally diverse population. PE Metrics has been validated by the developers in an area that lacks the predicted diversity of the United States. Therefore, there is a need to examine this instrument in settings with diverse populations. Although the long-term goal of this project is to investigate effective teaching practices in PE using student achievement and teacher observational tools, the first step is to investigate a standardized method of measuring student achievement. To this end, the purpose of this study was to (a) test the effectiveness of using PE Metrics with a low socioeconomic, culturally diverse population and (b) examine the test sensitivity of assessments for teachers to be able to discriminate among students with varying levels of abilities.

Method

Participants

Participants were 60 fifth (M = 9.9 years, SD = .52) and 30 second (M = 6.5 years, SD = .50) graders enrolled in a physical activity camp at a southwestern university. One enrollment requirement for the summer camp was that the campers had to come from

families with incomes near or below the poverty line, and therefore, all participants could be classified as coming from low socioeconomic status. Ethnically, the fifth grade participants self-reported being Hispanic or Latino (74%), Caucasian (10%), Native American (9%), African American (4%), and not identified (3%). The second grade participants reported being Hispanic or Latino (78%), Caucasian (13%), Native American (11%), African American (2%), and not identified (6%). This sample is the population in which further studies of teacher effectiveness would occur.

Procedure

During the gymnasium sports activity time of each camp, participants individually performed the targeted skills. Two raters watched the skill and scored participants on the corresponding PE Metrics rubric. Data collection occurred over several days after minimal amounts of training between the raters. All participants performed one skill before rating began on the next skill. Raters were former K–12 teachers with experience in PE and elementary settings. Interrater reliability for second grade skipping, galloping, and basketball dribbling was $r = .85$, $r = .82$, and $r = .96$, respectively. Interrater reliability for fifth grade basketball dribbling, soccer dribbling, and overhand throw was $r = .93$, $r = .96$, and $r = .94$, respectively.

Analysis

To determine the test sensitivity and effectiveness of using PE Metrics with a low socioeconomic, culturally diverse population, an item analysis was conducted on three motor skills within second (skipping, galloping, and dribbling with hand and ball) and fifth (basketball dribbling, soccer dribbling, and overhand throw) grade assessments. Test sensitivity is whether teachers can use a scale to discriminate among individuals with varying levels of abilities. Sensitivity of scale items of a given assessment was documented to deviate across populations (Ferketich, 1991), creating the need for assessments to be tested within and across many populations.

Results

Grade 2

The results of item analysis are provided in Table 2. Of specific interest to the current project was the item discrimination index and difficulty measure. The discrimination index, a measure of item sensitivity, is the tool teachers used to differentiate between students

who could efficiently demonstrate the assessed motor skill and students who had not developed or learned the motor skill. Results of the motor skills of 6- and 7-year-olds indicated that teachers cannot use the rubric for skipping to differentiate between ability levels with both of the skills within the rubric (form and consistency), resulting in a discrimination index of 0 and an item difficulty of 1, indicating that 100% of students within this population were able to competently complete this skill. The results of the motor skill of galloping, with a discrimination index of 0.26 and a difficulty measure of 0.86, indicated that teachers can use the rubric to discriminate between levels. This difficulty measure indicates that 86% of the students were able to meet the criteria for competence for galloping. Sufficient test sensitivity was revealed for the motor skill of dribbling with hand and ball, with results indicating a discrimination index of 0.73, 0.81, and 0.73 for the skills of form, spacing, and ball control, respectively. Difficulty measures of 40%, 30%, and 43% indicated students had competent form, competent spacing, and competent ball control, respectively.

Table 2
Results of Grade 2 Data

Item	*N*	*M*	*SD*	Discrimination index	Difficulty measure
Skipping					
Form	30	3.93	0.25	0	1.00
Consistency	30	3.86	0.34	0	1.00
Galloping					
Form	30	3.53	0.73	0.26	0.86
Consistency	30	3.70	0.59	0	0.93
Dribbling with hand jog					
Form	30	2.30	0.91	0.73	0.40
Spacing	30	2.26	1.01	0.81	0.30
Ball control	30	2.36	0.88	0.73	0.43

Grade 5

Results for Grade 5 are presented in Table 3. Grade 5 rubrics were more efficient for differentiating between students who could competently complete the motor skill and student who could not. However, some of the differentiation or variance between students' skill levels may be due to gender rather than skill ability. Interesting

patterns emerged when the results of the motor skills assessed were separated by gender. Overall, the motor skill of basketball had a discrimination index of 0.50, 0.66, and 0.75 for dribbling, passing, and receiving, respectively, with 75% of students being able to competently dribble, 65% being able to pass, and 72% being able to competently receive the basketball. When the results were separated by gender, 93% of males and only 64% of females were competent in dribbling. For the skills of passing and receiving, 90% of males and 48% of females were competent in passing and 90% of males and 53% of females competent in receiving. Results separated by gender are provided in Table 4. At the aggregated level, soccer motor skills on the whole, particularly foot motor skills, were more difficult for students, with results indicating only 54% of students being able to dribble a soccer ball competently, 35% able to pass, and 52% able to receive. The teachers could discriminate between those who were competent and those who could not complete the skill, with an index of 0.33 for dribbling a soccer ball, 0.33 for passing, and 0.50 for receiving. When results were separated by gender, males in general were rated more frequently as being competent in the motor skill than females, with 60% of males and only 41% of females being competent in dribbling a soccer ball. For passing and receiving, 35% and 65% of males, respectively, were rated as competent, with only 28% and 37% of females competent, respectively. Last, the results from the overhand pass revealed discrimination indices of 0.75 for form and 0.33 for accuracy, with 48% of students having competent form and 59% being competent in accuracy. When results were examined by gender ratings, 67% of males were competent in their form and 80% were competent in their accuracy of overhand throwing. For females, only 31% were competent in their form and 39% competent in their accuracy of overhand throwing.

Table 3
Results of Grade 5 Data

Item	*N*	*M*	*SD*	Discrimination index	Difficulty measure
Basketball					
Dribble	47	2.95	0.75	0.50	0.75
Pass	47	2.78	0.77	0.66	0.65
Receive	47	2.87	0.76	0.75	0.72

Table 3 (cont.)

Item	*N*	*M*	*SD*	Discrimination index	Difficulty measure
Soccer					
Dribble	42	2.54	0.73	0.33	0.54
Pass	42	2.16	0.79	0.33	0.35
Receive	42	2.50	0.70	0.50	0.52
Overhand throw					
Form	37	2.51	0.98	0.75	0.48
Accuracy	37	2.56	0.92	0.33	0.59

Table 4
Results of Grade 5 Data Separated by Gender

	Male				Female			
Item	*N*	*M*	*SD*	Difficulty measure	*N*	*M*	*SD*	Difficulty measure
Basketball								
Dribble	28	3.28*	0.59	0.93	19	2.66	0.79	0.64
Pass	28	3.14*	0.59	0.90	19	2.33	0.79	0.48
Receive	28	3.32*	0.66	0.90	19	2.52	0.74	0.53
Soccer								
Dribble	20	2.65	0.58	0.60	22	2.18	0.90	0.41
Pass	20	2.40*	0.59	0.35	22	1.90	0.92	0.37
Receive	20	2.70*	0.57	0.65	22	2.27	0.76	0.37
Overhand throw								
Form	24	2.91*	0.77	0.67	13	2.00	1.00	0.31
Accuracy	24	3.08*	0.71	0.80	13	2.23	0.92	0.39

*Male means significantly higher than female means at the alpha .05 level.

Discussion

PE Metrics is a beginning to the development of a standardized system to measure student achievement across PE settings. However, the results of this study show strengths and weaknesses of the PE Metrics system. The greatest strength is that PE Metrics is the first steps toward an evidence-based system to assess students in PE.

This begins a much-needed discussion of how to standardize methods of assessment in PE to meet new educational reform guidelines.

Moreover, evidence from this study strengthens the argument that reliable scores result from PE Metrics. Dyson et al. (2011) in their research on the performance of PE Metrics with elementary-aged students found through an item analysis that 57% of participants established a mean score between 2.2 and 2.8. In comparison, the item analysis in this study indicated that 64% of participants achieved a mean score between 2.2 and 2.8. This finding is supportive of Dyson et al.'s conclusion that the assessments provided by PE Metrics accurately capture student growth.

This study also indicates that individuals who have minimal experience in assessing student growth in PE may use PE Metrics. In this study, the raters had under 10 hr of training and were able to use the rubrics to produce reliable scores in real time. Although the developers of PE Metrics recommend tape-recording students performing these skills and scoring at a later time, the rubrics may be clear and concise enough to score participants in real time. This indicates PE Metrics may be used in a classroom environment with or without a videotaping mechanism, suggesting PE Metrics could be used in PE programs with limited resources.

Furthermore, one of the raters was an expert in PE and the second rater was an elementary teacher with minimal experience in PE settings. That both raters were able to produce consistent scores indicates that PE teachers of different levels of expertise may use PE Metrics to measure student achievement accurately in PE environments. Thus, beginning and veteran teachers may use PE Metrics as a method to assess student performance regardless of level of experience or expertise.

Given the small-scale nature of this study, there may be possible concerns regarding the variance in scores. For example, the second grade sample had just finished first grade and most students tested had already mastered the skills at the second grade level (100% skipping, 86% galloping). At the fifth grade level, variance seems to have occurred due to gender, suggesting a possible bias in the rubrics. This variance in gender could be a result of male participants playing the game of basketball more frequently than females. More research is needed to determine whether there was a gender bias within the PE Metrics assessment or males from low socioeconomic, culturally diverse settings play basketball more often than females. However, if these assessments are to be used as a

standardized method to measure student achievement, enhanced test sensitivity or modifications for female and male participants based on previous sport experience are essential to document change from one administration to the next.

Conclusion

PE Metrics has proven to be a reliable system to measure student achievement in PE settings. However, if PE Metrics is to be the primary vehicle to measure student achievement with intent to provide evidence of PE teacher effectiveness, more studies of PE Metrics in low socioeconomic, culturally diverse environments are needed to ensure a valid, fair, and objective system of assessment. Further studies should also be produced that focus more specifically on whether the results of PE Metrics are different based on gender.

References

Clotfelter, C., Ladd, H., & Vigdor, J. (2010). *How and why do teacher credentials matter for student achievement?* (NBER Working Papers 12828). Cambridge, MA: National Bureau of Economic Research.

Darling-Hammond, L. (2009). Recognizing and enhancing teacher effectiveness. *The International Journal of Educational and Psychological Assessment, 3*, 1–24.

Dyson, B., Placek, J. H., Graber, K. C., Fisette, J. L., Rink, J., Zhu, W., . . . Park, Y. (2011). Development of PE Metrics elementary assessments for national physical education standard 1. *Measurement in Physical Education and Exercise Science, 15*, 100–118.

Ferketich, S. (1991). Focus on psychometrics: Aspects of item analysis. *Research in Nursing & Health, 14*(2), 165–168.

Fisette, J., & Franck, M. (2013). How teachers can use PE Metrics for formative assessment. *Journal of Physical Education, Recreation, and Dance, 83*(5), 23–24.

Fox, C., Zhu, W., Park, Y., Fisette, J. L., Graber, K. C., Dyson, B., . . . Franck, M. (2011). Related psychometric issues and their resolutions during development of PE Metrics. *Measurement in Physical Education and Exercise Science, 15*, 138–154.

Goldhaber, D. (2002). The mystery of good teaching: Surveying the evidence on student achievement and teachers' characteristics. *Education Next, 2*(1), 129–145.

House, J. S., & Williams, D. R. (2000). Understanding and reducing socioeconomic and racial/ethnic disparities in health. In B. D. Smedley & S. L. Syme (Eds.), *Promoting health: Intervention strategies from social and behavioral research* (pp. 81–125). Washington, DC: National Academy Press.

Johnson, R. (2008). Overcoming resistance to achievement-based unit grading in secondary physical education. *Journal of Physical Education, Recreation, and Dance, 79*(4), 46–49.

Lambert, L. T. (2007). *Standards-based assessment of student learning: A comprehensive approach*. Reston, VA: National Association of Sport and Physical Education.

McKenzie, T. L., & Lounsbery, M. A. (2009). School physical education: The pill not taken. *American Journal of Lifestyle Medicine, 3*(3), 219–225.

Melagrano, V. (2007). Grading and report cards for standards-based physical education. *Journal of Physical Education, Recreation, and Dance, 78*(6), 45–53.

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

National Association for Sport and Physical Education. (2001). *Standards for initial programs in physical education teacher education*. Retrieved October 10, 2011, from http://www.aahperd.org/naspe/grants/accreditation/upload/standards_initial.pdf

National Association for Sport and Physical Education. (2012). *How can I demonstrate to my building principal that I am an effective physical education teacher?* Reston, VA: Author.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115, Stat. 1425 (2002).

Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

U.S. Census Bureau. (2012). US census bureau projections show a slower, older, more diverse nation a half century from now: 2012 [Press release]. Retrieved from http://www.census.gov/newsroom/releases/archives/population/cb12-243.html

Weisberg, D., Sexton, S., Mulher, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness* (2nd ed.). Retrieved from The New Teacher Project website: www.widgeteffect.org

Wright, S., Horn, S., & Sanders, W. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57–67.

Zhu, W., Fox, C., Park, Y., Fisette, J. L., Dyson, B., Graber, K. C., . . . Raynes, D. (2011). Development and calibration of an item bank for PE Metrics assessments: Standard 1. *Measurement in Physical Education and Exercise Science, 15*, 119–137.

Zhu, W., Rink, J., Placek, J. H., Graber, K. C., Fox, C., Fisette, J. L., . . . Raynes, D. (2011). PE Metrics: Background, testing theory, and methods. *Measurement in Physical Education and Exercise Science, 15*, 87–96.