# RELIABILITY AND VALIDITY OF THE FITNESSGRAM® PHYSICAL ACTIVITY ITEMS

Kaleigh San Miguel

Thesis Prepared for Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

August 2011

APPROVED:

Scott B. Martin, Major Professor
James R. Morrow, Jr., Committee Member
Allen W. Jackson, Committee Member and Chair of the Department of Kinesiology, Health Promotion, and Recreation
Jerry R. Thomas, Dean of the College of Education
James D. Meernik, Acting Dean of the Toulouse Graduate School

San Miguel, Kaleigh, Reliability and validity of the FITNESSGRAM® physical activity items. Master of Science (Kinesiology), August 2011, 33 pp., 9 tables, references, 25 titles.

Large-scale assessments of children and youth physical activity (PA) behaviors are regularly conducted in school settings. In addition to assessing actual fitness, the FITNESSGRAM® assesses self-reported PA behaviors for aerobic, strengthening, and flexibility activity within the past 7 days. The purpose of this study was to examine the reliability and validity of the three PA items. Participants included 1010 students in grades three through twelve and were either tested under a teacher – teacher condition, an expert - expert condition, a teacher – expert condition, or a trained teacher – expert condition. Comparisons of the responses to the PA items indicated adequate reliability for teachers, but the reliability improved with training. Likewise, the validities for teachers are moderate to fair; however, they improved when teachers received additional training.

TABLE OF CONTENTS

LIST OF TABLES

# INTRODUCTION

## Past Research

The prevalence of overweight children and adolescents in the United States has increased dramatically over the past fifty years (Daniels, Arnett, Eckel, Gidding, Hayman, & Kumanyika, 2005; Ogden & Carroll, 2010; Ogden, Carroll, & Flegal, 2008). It is estimated that 16% of school age children are overweight, twice the number of overweight children from twenty years ago (Centers for Disease Control and Prevention [CDC], 2007). The prevalence of obesity is one of the most imperative public health concerns facing the United States today (CDC, 2009). A feasible health goal is to expend more energy by increasing physical activity (PA) while decreasing energy intake (Hill, Wyatt, Reed, & Peters, 2003). In fact, healthy People 2010 identified overweight and obesity and PA as two of the top ten leading health indicators (LHIs; CDC, 2011). These health indicators were selected because of their potential to motivate action toward specific national targeted goals (see www.healthypeople.gov/2010/LHI/). Unfortunately, little progress toward the target goal of reducing the proportion of children and adolescents who are overweight or obese has occurred. One possible reason is that little time and effort have been devoted to regular fitness testing in schools over the past thirty years and is often replaced by self-reported PA measures. Recently PA measures have been under greater scrutiny for their lack of reliability and validity (Chinapaw, Mokkink, van Poppell, van Mechelen, & Terwee, 2010). Epidemiological studies have typically used subjective measures, such as the questionnaire, to assess PA in populations (Chinapaw, Slootmaker, Schuit, van Zuidam, & van Mechelen, 2009). PA questionnaires are easy to administer, non-reactive, relatively inexpensive and

accepted by study participants (Chinapaw et al., 2009). One area where the research is lacking is whether physical education teachers are able to appropriately administer self-reported PA items and if the data collected are reliable and valid.

The rates of overweight and obese children in Texas and other southern states are even more disturbing than nationwide rates (see http://www.cdc.gov/obesity/data/trends.html). Approximately one in three (35%) children in Texas is overweight or obese, which is more than double the national average (Texas Health Institute, 2006). Passing legislation is a major step to addressing the issue of overweight and obese children in Texas. Texas Senate Bill 19, which requires students in publicly funded elementary schools and middle schools to participate in PA, was enacted to improve the health of Texas children (Kelder, Springer, Barrosso, Smith, Sanchez, Ranji et al., 2009). The passage of Senate Bill 19 was significant because it was one of the first statewide efforts to target child health through mandated PA time and health education. In an attempt to reduce the number of overweight and obese children in Texas, the Texas State House of Representatives and Senate passed Senate Bill 530 (SB530) in 2007. This bill required each school district to assess the physical fitness of students on a yearly basis in Grade s three through twelve (see http://www.legis.state.tx.us/billlookup/text.aspx?LesSess=80R&Bill=SB530). Following this bill, schools began to implement a structured health program which included fitness testing in elementary, middle, and high schools. These efforts were viewed as positive attempts at making large scale changes (Kelder et al., 2009).

## FITNESSGRAM®

During that same year the Texas Education Agency (TEA) selected the FITNESSGRAM, a comprehensive physical fitness assessment battery for youth, as the evaluation tool to be used to accomplish this goal (Morrow et al., 2010). The TEA selected the FITNESSGRAM physical fitness battery as the required tool for the project because of its established test reliability and validity as well as its highly developed data tracking software (Morrow et al., 2010). The FITNESSGRAM physical fitness assessment includes aerobic fitness, muscle strength and endurance, flexibility, and body composition (Meredith & Welk, 2008). In addition to the FITNESSGRAM physical fitness assessment, some schools also use the FITNESSGRAM self-report PA questions which request students to recall their aerobic, strength and endurance, and flexibility activities during the past seven days.

Public schools throughout the state of Texas are required to submit their FITNESSGRAM data to TEA through customized data aggregation tools developed by The Cooper Institute (Morrow et al., 2010). These data are collected and submitted to TEA while preserving student confidentiality. Individual data are not tracked but school level data are available by grade and gender (Meredith & Welk, 2007).

An important factor to the comprehension and decision making associated with the test results is the measurement quality of the testing. The FITNESSGRAM battery is considered by many experts to be the most psychometrically sound assessment of fitness available for field-based testing in youth (Meredith & Welk, 2005; Plowman, Sterling, Corbin, Meredith, Welk, & Morrow, 2006). However, there is a gap in the

literature regarding information about the variability in psychometric properties (i.e., reliability and validity) of fitness and PA testing across various test settings.

Because previous studies have included data that have been collected by different physical education teachers in different school testing environments and grade levels across the state of Texas it is important to evaluate whether teachers are using the testing protocol appropriately and consistently. Therefore, reliability and validity are essential for interpretation and inference of data collection and test results. The reliability and validity of the FITNESSGRAM assessments are documented in the FITNESSGRAM Reference Guide (Meredith & Welk, 2007), but most of the published studies have been completed using small, convenience samples, whereas few have examined large-scale fitness testing (Hartman & Looney, 2003; Jackson, Morrow, Jensen, Jones, & Schultes, 1996; Joyner, 1997; Mahar, Rowe, Parker, Mahar, Dawson, & Holt, 1997; Patterson, Rethwisch, & Wiksten, 1997; Patterson, Wiksten, Ray, Flanders, & Sanphy, 1996). Examining the reliability and validity of large scale fitness and PA assessment would provide more insight for establishing proper protocol for conducting accurate testing.

The measurement of fitness and PA in children and adolescents has become an important field of interest and challenging enterprise because of its influence on health. There is a need to develop low cost, practical, and accurate measures of fitness and PA in children and adolescents; self-report PA is a promising methodology for large studies with these populations (Sallis, Buono, Roby, Micale, & Nelson, 1993). However, methodological problems have impeded research in PA epidemiology and a number of researchers have identified the need to improve the assessment of PA, particularly

among children (Cale, 1994, Chinapaw et al., 2010). To reduce the cost of large scale fitness testing, an increasing number of researchers have developed and used self-report measures for children and adolescent (see Chinapaw et al., 2010). It has been recommended that efforts be made to develop valid and reliable measures of self-reported aerobic activity for children and youth (see Belton, 2010; Chinapaw et al., 2010). PA levels of younger children revealed the need for a valid and reliable measure of PA that could be used to gauge the level to which children and adolescent are meeting current national PA guidelines (Belton, 2010). Although the National Institute of Health (NIH, 2010) recognized the role of self-report PA instruments to gather information on large numbers of people in a relatively small amount of time without incurring large costs, they need to be psychometrically sound (Chinapaw et al., 2010).

Morrow and colleagues (2010) examined the quality of large-scale physical fitness testing in Texas to determine if reliability and validity of the obtained test results from the FITNESSGRAM assessment items were related to potential confounding variables. Participants included students enrolled in school districts in the Dallas-Fort Worth metroplex, grades three through twelve. Individuals participated in the FITNESSGRAM assessment items and were tested on two separate occasions. The purpose of Morrow and colleagues' research was to report the psychometric characteristics of the data collected and compare these results to results obtained from highly trained individuals. Their results indicated that the validity of teacher-administered tests is good when compared to highly trained individuals. They also reported that reliability and validity of teacher-obtained health-related fitness measures is generally unrelated to potentially confounding student or school characteristics. The results from

their study offered potential insight about the quality and effectiveness of large-scale school-based testing (Morrow et al., 2010). Collected data reflect the capabilities of physical education teachers to conduct fitness tests and enhance the overall quality of large-scale physical fitness testing (Morrow & Ede, 2009). Morrow et al. (2010) concluded that teacher administered criterion-referenced, health-related physical fitness tests appear to be reliable and valid. However, since reliability and validity can increase with training, it is important for those considering large-scale testing to include plans and rationale for conducting widespread training on the administration of specific items to be tested (Morrow et al., 2010). Concurrently, it is important to determine if the three FITNESSGRAM PA items that examine aerobic, muscular strength and endurance, and flexibility during the past seven days are reliable and valid in a large-scale school-based testing environment, especially if states or schools receive less funding for actual fitness testing.

## Purpose

The primary purpose of the current investigation is to examine the psychometric characteristics (i.e., reliability and validity) of the test administrators who collect the self-report responses to the three FITNESSGRAM PA items. The data were collected during a typical school setting. Data collected by physical education teachers, teachers with additional training (trained teachers), and highly trained individuals (experts) will be compared to determine if the three FITNESSGRAM PA items are reliable and valid (in the same manner as the FITNESSGRAM physical fitness data were collected by Morrow et al., 2010). Based on Morrow et al.'s (2010) research it is hypothesized that the comparisons of the three FITNESSGRAM PA questions across the various testing

conditions will indicate sufficient reliability for teachers, but reliability will be better with individuals who have received additional training. While validity for data collected by teachers administering FITNESSGRAM PA questions will be low to moderate, the accuracy and validity of responses submitted by the participants will improve substantially when physical education teachers and administrators are trained. Data collection were collected through the four different groups assigned (Group 1; Teacher – Teacher; Group 2; Expert – Expert; Group 3, Teacher – Expert; Group 4; Trained Teacher – Expert). Therefore, it is hypothesized that reliability and validity of self-reported PA aerobic, muscular strength and endurance, and flexibility behaviors are better when individuals are trained. Another aim of this study was to determine whether students' self-reported aerobic and strengthening PA accurately reflected their aerobic and strengthening activity as categorized by the 2008 PA Guidelines (see http://www.health.gov/PAGuidelines/).

## METHODOLOGY

As part of a larger study, methodology was established and the data collection process was completed (see Morrow et al., 2010). Schools were selected based on their location (e.g., urban, suburban, or rural), ethnicity composition, SES, school level (elementary, middle, or high school), and previous year's Texas Assessment of Knowledge and Skills (TAKS) status (i.e., exemplary, recognized, acceptable, and unacceptable). Teacher characteristics included gender, experience, and the type of training that had been completed prior to administering the FITNESSGRAM.

### Participants

The number of participants include 1010 individuals (439 boys, 569 girls, and 2 who did not report their gender) of which 586 (58%) were from elementary schools (3rd and 5th grade) and 424 (42%) were from secondary schools (≥ 7th grade) (see Table 3). The recruitment process included contacting each school individually and receiving permission to work with schools from across the Dallas – Fort Worth metroplex districts to participate in this component of the project. Informed consent procedures were approved by the University of North Texas and The Cooper Institute IRBs.

### Instruments

The participants answered the three FITNESSGRAM PA questions. These questions ask participants to recall the past seven days of PA in three different activity areas: aerobic, muscular strength and endurance, and flexibility (see Table 1).

### Design and Procedures

The reliability and validity of fitness testing was evaluated using a multigroup (four group) design. Classrooms of students in Grades 3, 5, 7, and 9 through 12 were

assigned to one of four testing conditions to facilitate the evaluation of reliability and validity of the three FITNESSGRAM PA questions. The reliability of teacher administered tests was evaluated by having teacher testers execute exact measures two weeks apart (Group 1; Teacher – Teacher). Reliability was evaluated for expert testers by having a mobile expert team conduct duplicate assessments in matched classrooms using identical protocols (Group 2; Expert – Expert). The evaluation of validity was assessed by comparing teacher administered data to the trained expert testers (Group 3; Teacher – Expert). The trained teachers, who are knowledgeable in the FITNESSGRAM and administering self-reports PA testing, obtained data that were compared to expert testing to examine validity once more (Group 4; Trained Teacher – Expert). Because there may be variability in experience with PA and with individuals administering the FITNESSGRAM, trained teachers were encouraged to complete the on-line training, read the manual and view the accompanying DVD, and rereading of manuals with trained expert testers.

Expertly trained mobile data collection teams were used to evaluate the reliability and validity of data from the FITNESSGRAM PA questions collected. Expert training include reading the FITNESSGRAM training manual, reviewing an on-line DVD demonstration, completion of online FITNESSGRAM certification, and hands-on training at The Cooper Institute and the University of North Texas (both led by a FITNESSGRAM manager). Expertly trained testing teams of two to four people were sent to oversee fitness and PA testing in schools, traveling from school to school administering the FITNESSGRAM within days of teachers conducting the identical tests. Expert testers were set at the criterion measure.

Students' self-report data were coded and entered into a database according to the 2008 PA Guidelines. These guidelines recommend seven days per week of 60 minutes per day of aerobic activity. Within those seven days, at least three days of muscular strength and three days of bone strength activities should be implemented for children and adolescents (http://www.health.gov/PAGuidelines/). The data were coded to indicate whether students engaged in three or more days of aerobic, strength and endurance, and flexibility PA or less than three days of these activities per week.

The collection of FITNESSGRAM data took place during the students' scheduled physical education classes over a two week time period at each school. Sampling was completed in multiple classrooms at participating schools to guarantee there was a diverse sample of students and teachers collecting data for the four group comparisons. To accurately collect the data, the test-retest procedures for the three FITNESSGRAM PA questions were scheduled about two weeks following the initial assessments. Initial testers (teachers, trained teachers, or experts) conducted FITNESSGRAM testing and data collection, and then within fourteen days the retest group (teacher or expert) collected a second round of data.

## Statistical Analysis

The FITNESSGRAM PA results obtained from individual students were entered into a database and categorized by student ID number. The results from the first and second administration were then matched to examine the reliability and validity of the data collected. Subsequently, reliability was examined by evaluating the teachers' first administration of data collected (test) and comparing these data against the teachers' second administration (re – test) using Cronbach's Alpha. Using the same protocol,

experts' first administration of data collected (test) were compared to the experts' second administration of data collected (re – test). In addition, concurrent validity was examined by comparing the teacher and expert FITNESSGRAM PA assessment administration (test and re – test) using Pearson product moment correlations. To test validity once more, trained teacher data were compared to expert collected data (test and re – test) using Pearson product moment correlations. Raw scores were then classified and converted into two groups (participating in three or more days of PA, or less than three days of PA) and reliability and validity were estimated from contingency tables based on the 2008 PA Guidelines and evaluated with percent of agreement, modified kappa coefficient (Looney, 1989), phi coefficient, and chi square.

## RESULTS

Participants included 23 teachers (19 female, 4 male) and 1010 students (439 boys, 569 girls, and 2 who did not report their gender). Of the 1010 students tested at 21 different schools (21 elementary school classes, 11 middle school classes, and 4 high school classes), 49% of students were tested under the teacher – teacher condition, 22% were tested under the expert – expert condition, 20% were tested under the Teacher – expert condition, and 10% were tested under the Trained teacher – expert condition (see Table 4). There were 277 individuals tested in Grade 3, 309 participants tested from Grade 5, 294 individuals tested from Grade 7, and 130 students were tested in Grades 9 – 12 (see Table 5).

### Reliability

*Teacher – teacher*. Reliability for Group 1 was examined by assessing the teachers' first administration results and comparing these data against the teachers' second administration using Cronbach's alpha. Reliability of Group 1 for the first FITNESSGRAM physical activity question targeting aerobic physical activity was moderate ($r_{xx}$' = .67). The reliability of the data collected by teachers for the second question targeting strength and endurance yielded the highest reliability of the three questions ($r_{xx}$' = .77). Reliability for Question 3, focusing on flexibility, was consistent with Question 1 ($r_{xx}$' = .67). Comparisons of the three FITNESSGRAM PA items across the various conditions indicated adequate reliability for untrained teachers ($r_{xx}$' = .70; see Table 6).

*Expert – expert*. Reliability for Group 2 was evaluated by having a mobile expert team conduct duplicate assessments in matched classrooms using an identical

protocol. Expert test-retest data collected yielded higher reliabilities among the three FITNESSGRAM PA items than the teacher collected data. Expert collected data for Question 1, targeting aerobic activity, yielded the highest reliability of data collected by Group 2 ($r_{xx'}$ = .84). Reliability of data collected on the second question by experts, focusing on strength and endurance, was still highly reliable ($r_{xx'}$ = .81). Reliability for data collected on Question 3 targeting flexibility was also high ($r_{xx'}$ = .83). Therefore, reliabilities are substantially better with expert testers ($r_{xx'}$ = .82; see Table 6).

## Validity

*Teacher – expert*. Validity for Group 3 was examined by using concurrent validity and comparing the teacher and expert administration (Group 3) using Pearson product moment correlations. Validity of data collected for Question 1, focusing on aerobic PA, was moderate ($r_{xy'}$ = .49). Validity increased with the data collection of Question 2, looking at strength and endurance ($r_{xy'}$ = .64). Question 3's validity, looking at flexibility, was moderate as well ($r_{xy'}$ = .46). Therefore, validities for untrained teachers are moderate to fair (.49 < *r* < .64; see Table 7).

*Trained - teacher – expert*. Validity for Group 4 was significantly higher than Group 3 (.64 < *r* < .81), indicating that validity increases when teachers are trained. Validity for Question 1, targeting aerobic fitness, was high ($r_{xy'}$ = .81). The validity of data collected by Group 4 on Question 2, focusing on strength and endurance, was moderate ($r_{xy'}$ = .64). Group 4 administration on Question 3, focusing on flexibility, yielded moderately high results ($r_{xy'}$ = .72; see Table 7).

## Reliability and Validity According to Physical Activity Guidelines

Reliability and validity results for self-reported PA from the FITNESSGRAM items for aerobic, strengthening and endurance, and flexibility activities are presented in Tables 8 and 9. These data were coded and categorized according to the 2008 PA Guidelines (see http://www.health.gov/PAGuidelines/). The three activity items on the FITNESSGRAM asked respondents to indicate the number of days that they have engaged in aerobic, strengthening and endurance, and flexibility activities (separately) in the last 7 days. The data were divided according to the 2008 PA Guidelines which recommend at least seven days per week of aerobic activity, and three days per week of muscular strength and flexibility activities for children and adolescents.

Reliability is presented in Table 8 (teacher – teacher and expert – expert). Percent agreements are generally good, modified kappas generally indicated moderate agreement, and phi and chi square results are significant. Importantly, the reliability is noticeably better with expert testers (Group 2; $r_{xx}$' = .55, .53, and .57, respectively) than compared with teacher testers (Group 1; $r_{xx}$' = .38, .50, and .40). Reliability for teachers is adequate or acceptable for all three FITNESSGRAM PA items; however, reliability for expert testers is better than those for typical classroom teachers.

Validity statistics are presented in Table 9 (teacher – expert and trained teacher – expert). Typical classroom teachers are compared with expert testers (Group 3). These data indicate the lowest validity, suggesting teachers do not obtain comparable results to experts ($r_{xy'}$ = .34, .58, and .37). Trained teachers are compared to expert testers (Group 4) in Table 9. Group 4 results are substantially higher ($r_{xy'}$ = .71, .57, and .53) than teacher – expert collected data, indicating that teachers with training are able

to facilitate and administer the FITNESSGRAM PA items more accurately than those without training. These results suggest acceptable validity for the teachers but validity is considerably enhanced when teachers undergo training.

## DISCUSSION

The results from this investigation offer insights about the quality and effectiveness of large-scale school-based testing. Physical educators should use fitness testing in a manner that promotes regular physical activity so that students can achieve basic health standards (Silverman, Keating, & Phillips, 2008). The results have direct implications for policy and curricular decisions within Texas, but the large scope and representative nature of the study may also provide insight for public health researchers, school districts, and state politicians considering similar large scale testing endeavors. The current study examined the reliability and validity of the three FITNESSGRAM PA items self-reported to test administrators. This study also examined students' aerobic and strengthening self-reported PA as categorized as either meeting (3 or more days) or not meeting (less than three days) according to the 2008 PA Guidelines (see http://www.health.gov/PAGuidelines/).

Generally, the results of the current study that examines the reliability and validity of the administration of the three self-report FITNESSGRAM PA questions indicate that they are similar to reliabilities and validities reported in a related investigation of the FITNESSGRAM fitness items. Morrow et al. (2010) reported good criterion-related reliability and validity for individual FITNESSGRAM fitness items. In the current study, the FITNESSGRAM PA items resulted in moderate to high reliabilities and validities when using expert testers as the criterion measure. These findings are dissimilar to the conclusions made by Chinapaw et al. (2010) who examined previous research investigating children and adolescent self-report PA items. That is, Chinapaw et al. (2010) reported only a few assessments received positive ratings for either reliability or

validity when using accelerometers as the criterion measure and none showed both acceptable reliability and validity. The reliabilities and validities reported by Chinapaw et al. (2010) varied, with test-retest correlations ranging from 0.02 to 0.96 whereas the results of the current study revealed test-retest correlations that ranged from .49 to .84. Similarly, intraclass correlations (ICCs) reported by Chinapaw et al. (2009) for adolescents' test-retest reliability was fair to moderate ranging from 0.30 to 0.59, whereas the ICCs reported in this study were moderate to high, ranging from .50 to .73. The findings of the current study were similar to those reported by Sallis, Buono, Roby, Micale, and Nelson (1993) which were 0.77 for the seven day recall. Sallis and colleagues indicated that the seven day PA recall of children who are as young as those in fifth grade are of adequate reliability and validity to use in research. In the current study, the data from children in 3rd and 5th grades were included and also appears to be reliable and valid.

Reliability generally increases when expert testers conduct the questioning. Teacher test-retest data procedures were performed in the same manner and the reliability was good. However, based on the results comparing typical classroom teachers to those with more extensive training and expertise, it appears that test results are sufficiently more valid when administered by an individual with more training. Teachers do a satisfactory job of PA assessment but results are visibly more reliable when experts are involved, which is consistent with previous findings using the FITNESSGRAM physical fitness assessment (Morrow et al., 2010). Adequate training for teachers to complete FITNESSGRAM fitness and PA testing will likely increase the quality of the data.

The validity of the data collected also appeared to be positively affected by teachers and administrators who had undergone some type of additional training. While the validities for untrained teachers are fair to moderate, the validity of the collected data were better when teachers who had been trained administered the FITNESSGRAM PA questions. If the facilitators of self-report PA receive additional FITNESSGRAM training, then the reliability and validity of the test-retest data increases. That is to say, additional training increases the possibility that participants respond to the questions more consistently and more accurately across the two administrations, consequently providing more reliable and valid results.

Likewise, when data were coded and categorized according to the 2008 PAG, results showed that reliability and validity increased when administrators had previous training. When looking at teacher – teacher administered testing, percent agreement is slightly lower than expert – expert percent agreement (see Table 8). Generally, when teachers administered the self-report PA questions to students, they reported physical activity levels differently from one assessment to the next. Experts who administered the FITNESSGRAM PA questions generated higher agreement across both assessments. Trained teacher – expert comparisons (Group 4) also resulted in higher agreement than teacher – expert comparisons (see Table 9). In fact, the teacher – expert percent agreement is substantially lower than trained teacher – expert collected data. Again, these outcomes are anticipated because of the training the trained teachers received.

Although the results of the current study indicates that physical education teachers adequately administer self-report PA items to children and adolescents,

additional training appears to enhance the overall quality of large-scale self-report PA administration. These results provide insight about the benefit of receiving additional training. Teacher administered criterion-referenced, health-related PA questions appear to be reliable and valid and decision makers can place confidence in the decisions made based on such data when performed in conjunction with fitness testing.

## Strengths

A strength of the study is that the overall sample size is fairly large for estimating reliability and validity. Also, this study provides insight on future instructions and recommendations for administering large-scale PA items. For example, a potential strength of the FITNESSGRAM PA question administration is that it instructs participants to use a seven day recall period, which should allow participants to provide relatively accurate feedback. An additional strength of the FITNESSGRAM PA questions is that because there are only three items it is not time consuming. Also, the format of the FITNESSGRAM PA questions could be administered in any type of school or activity setting, allowing administrators to conduct this testing in a variety of locations.

## Limitations

A potential limitation of the current study is the accuracy of the participants' responses to the self-reported FITNESSGRAM PA questions. Although the increase in percent agreement is likely due to the proper training of the administrator who facilitated the testing, it could also be the case that under the teacher – teacher condition students may have accurately reflected their PA on both self-reports but increased or decreased their PA level from the first to the second test. Another limitation to this study is the conflicting school scheduling and school closures. Because of TAKS testing, blocked

scheduling, teacher conferences, spring break, holidays, and field trips, sometimes the test-retest procedures did not go as planned. Another interference with the data collection included bad weather days and ill-related sicknesses (i.e., number of flu cases which lead to swine flu school closures). Absences related to illness, cutting class, suspension, and changing schools possibly influenced the test-retest protocol. That is to say, these disruptions may have influenced the students' PA and how the questions were administered.

In addition to the previous limitations, teachers administering the FITNESSGRAM PA questions sometimes phrased the items differently from the first to the second administration. The different instructions may have caused the students to report incorrect data. For example, in one instance teachers instructing the individuals to recall their activity from the previous week prior which was associated with practicing the FITNESSGRAM fitness testing while during the other administration they did not mention examples of moderate and vigorous activity and previous PA during class. These types of statements may have influenced how the students' self-report their actual PA levels. In addition, the FITNESSGRAM PA questions were administered in conjunction with the FITNESSGRAM fitness testing which may have influenced the results. For example, if the FITNESSGRAM PA questions were administered prior to the FITNESSGRAM physical fitness testing, then students may provide responses that are more consistent with their regular PA as compared to it being associated with their PA within the previous seven days that included practicing for the FITNESSGRAM fitness testing. If the students regularly do the fitness activities associated with the FITNESSGRAM fitness testing then their responses may be more consistent and

accurate. Another issue with this study is the lack of student and teacher participation. That is, some were motivated to perform the assessment once but was less motivated to participate in testing a second time. Although the teachers received funding to participate, they may have thought it was too time consuming after they agreed to participate. In addition to time believing it was time consuming, some of the students may have viewed the PA and fitness testing as punishment, resulting in them not want to fully participate.

## Implications

Based on the current study, facilitators who receive additional training are likely to report more reliable and valid data when administering the FITNESSGRAM PA questions. Hence, the data submitted to the state of Texas are likely more reliable and valid with teachers who have undergone training. Because of the type of administration provided, measurements could be prone to error, causing problems with internal validity, and affecting the overall quality of the data collected. Therefore, it is imperative to implement some type of training to facilitators prior to the administration of the FITNESSGRAM PA questions. To increase the reliability and validity of the assessment it is also recommended that teachers and administrators prepare the students ahead of time to answer the PA questions (Meredith & Welk, 2007). It may be difficult for young children to accurately recall this information, therefore the FITNESSGRAM PA questions should be explained thoroughly by the test administrator. Administrators of the FITNESSGRAM PA questions also need to be confident in explaining the different types of PA (aerobic, muscular strength and endurance, and flexibility) and be able to illustrate how to count the number of activity days.

The three FITNESSGRAM PA questions do not directly align with the 2008 PA Guidelines. This may need to be taken into consideration in future FITNESSGRAM reference manuals. Likewise, future researchers may want to examine the influence of the written and verbal instructions of the three FITNESSGRAM PA items more systematically. When facilitators administrate the FITNESSGRAM PA items prior to fitness testing their instructions may be different versus administration during the actual fitness testing. A more detailed description on how to administer these questions may lead to more reliable and valid results. This could then influence the type of information teachers receive prior to administering the FITNESSGRAM PA questions, which would likely influence the format of the instructions within the FITNESSGRAM Manual.

## CONCLUSION

The FITNESSGRAM PA items appear to be adequately reliable and valid, but the results indicate that teachers' data collection may be better with additional training. Sufficient large-scale PA testing that collects information that leads to better physical fitness and health will not occur unless those involved receive proper knowledge, preparation, motivation, and support.

Table 1

*FITNESSGRAM Physical Activity Items*

| |
|---|
| For each of the following questions, think about what you have done during the past 7 days. |
| 1. On how many days were you physically active for a total of at least 60 minutes? This includes moderate activities (walking, slow bicycling, or outdoor play) as well as vigorous activities (jogging, active games, or active sports such as basketball, tennis, or soccer). (Add up all the time you spend in any kind of physical activity that increases your heart rate and makes you breathe hard some of the time).<br>0 Days 1 Day 2 Days 3 Days 4 Days 5 Days 6 Days 7 Days |
| 2. On how many days did you do exercises to strengthen or tone the muscles such as push-ups, sit-ups, or weight lifting?<br>0 Days 1 Day 2 Days 3 Days 4 Days 5 Days 6 Days 7 Days |
| 3. On how many days did you do stretching exercises to loosen up or relax the muscles? This includes exercises such as toe touches, knee bending, and leg stretching.<br>0 Days 1 Day 2 Days 3 Days 4 Days 5 Days 6 Days 7 Days |

Table 2

*Study Design for Assessing Reliability and Validity*

| Group | Initial Test | Retest | Measurement Quality |
|---|---|---|---|
| 1 | Teacher administered | Teacher administered | Reliability |
| 2 | Expert Team administered | Expert Team administered | Reliability |
| 3 | Teacher administered | Expert Team administered | Validity |
| 4 | Trained Teacher administered | Expert Team administered | Validity |

*Note*. For addition information about the design of the study see Morrow et al. (2010).

Table 3

*Distribution of Participants by School Level*

| School Level | Frequency | Percent |
|---|---|---|
| Elementary | 586 | 58 |
| Middle | 294 | 29 |
| High | 130 | 13 |
| Total | 1010 | 100 |

Table 4

*Distribution of Participants by Groups*

| Group | Participants | Percentage | Schools | Classes |
|---|---|---|---|---|
| Teacher – Teacher | 492 | 49% | 9 | 16 |
| Expert – Expert | 219 | 22% | 5 | 8 |
| Teacher – Expert | 202 | 20% | 4 | 7 |
| Trained Teacher – Expert | 97 | 10% | 3 | 5 |

Table 5

*Distribution of Participants by Grade Level*

| Grade Level | Frequency | Percent |
|---|---|---|
| 3 | 277 | 27 |
| 5 | 309 | 31 |
| 7 | 294 | 29 |
| 9 –12 | 130 | 13 |
| Total | 1010 | 100 |

Table 6

*Reliability: Teacher – Teacher Verses Expert – Expert*

| | Teacher – teacher (Group 1) | | | Expert – expert(Group 2) | | |
|---|---|---|---|---|---|---|
| Item | *n* | $r_{xx}'$ | Single Admin. | *n* | $r_{xx}'$ | Single Admin. |
| Aerobic Activities | 492 | .67 | .50 | 219 | .84 | .73 |
| Strength Activities | 492 | .77 | .63 | 219 | .81 | .69 |
| Flexibility Activities | 492 | .67 | .51 | 219 | .83 | .71 |

*Note*. Single Admin. = Estimate for Single Administration. Sample sizes change because of classroom sizes with specific comparisons.

Table 7

*Validity: Teacher – Expert Verses Trained Teacher – Expert*

| | Teacher – Expert(Group 3) | | Trained Teacher – Expert (Group 4) | |
|---|---|---|---|---|
| Item | *n* | $r_{xy}$ | *n* | $r_{xy}$ |
| Aerobic Activities | 202 | .49 | 97 | .81 |
| Strength Activities | 202 | .64 | 97 | .64 |
| Flexibility Activities | 202 | .46 | 97 | .72 |

*Note*. Sample sizes change because of classroom sizes with specific comparisons.

Table 8

*Reliability of Repeated Test Administration: Classified according to the 2008 PAG*

| Teacher – teacher (Group 1) | | | | | |
|---|---|---|---|---|---|
| Item | % Agreement | Modified Kappa | Phi | Chi Squared | *N* |
| Aerobic Activities | .85 | .70 | .38 | .001 | 492 |
| Strength Activities | .74 | .48 | .50 | .001 | 492 |
| Flexibility Activities | .76 | .52 | .40 | .001 | 492 |
| Expert – expert(Group 2) | | | | | |
| Item | % Agreement | Modified Kappa | Phi | Chi Squared | *N* |
| Aerobic Activities | .85 | .70 | .55 | .001 | 219 |
| Strength Activities | .77 | .54 | .53 | .001 | 219 |
| Flexibility Activities | .79 | .58 | .57 | .001 | 219 |

*Note.* Sample sizes change because of classroom sizes with specific comparisons.

Table 9

*Validity of Repeated Test Administration: Classified according to the 2008 PAG*

| Teacher – expert(Group 3) | | | | | |
|---|---|---|---|---|---|
| Item | % Agreement | Modified Kappa | Phi | Chi Squared | *N* |
| Aerobic Activities | .88 | .76 | .34 | .001 | 202 |
| Strength Activities | .81 | .62 | .58 | .001 | 202 |
| Flexibility Activities | .73 | .46 | .37 | .001 | 202 |
| Trained Teacher – expert(Group 4) | | | | | |
| Item | % Agreement | Modified Kappa | Phi | Chi Squared | *N* |
| Aerobic Activities | .89 | .78 | .71 | .001 | 97 |
| Strength Activities | .79 | .58 | .57 | .001 | 97 |
| Flexibility Activities | .77 | .54 | .53 | .001 | 97 |

*Note.* Sample sizes change because of classroom sizes with specific comparisons.

REFERENCES

Cale, L. (1994). Self-report measures of children's physical activity: Recommendations for future development and a new alternative measure. *Health Education Journal, 53*, 439-453.

Centers for Diesease Control Prevention. (2007). *Overweight and obesity*. Retrieved from http://www.cdc.gov/obesity/data/trends.html

Centers for Disease Control Prevention. (2009). *Childhood overweight and obesity.* Retrieved from http://www.cdc.gov/obesity/childhood/index.html

Centers for Disease Control Prevention. (2011). *Obesity and overweight: An overview.* Retrieved from http://www.cdc.gov/obesity/index.html

Chinapaw, M. J. M., Mokkink, L. B., van Poppell, M. N. M., van Mechelen, W., & Terwee, C. B. (2010). Physical activity questionnairs for youth: A systematic review of measurement properties. *Sports Medicine, 40*, 539-563.

Chinapaw, M. J. M., Slootmaker, S. M., Schuit, A. J., van Zuidam, M., & van Mechelen, W. (2009). Reliability and validity of the activity questionnaire for adults and adolescents (AQuAA). *BMC Medical Research Methodology, 9*, 58-75.

Daniels, S. R., Arnett, D. K., Eckel, R. H., Gidding, S. S., Hayman, L. L., Kumanyika, S. et al. (2005). Overweight in children and adolescents: Pathophysiology, consequences, prevention, and treatment. *Circulation, 111*, 1999-2012.

Hartman, J. G., & Looney, M. A. (2003). Norm-referenced and criterion-referenced reliability and validity of the back-saver sit-and-reach. *Measurement in Physical Education and Exercise Science, 7,* 71-87.

Hill, J. O., Wyatt, H. R., Reed, G. W., & Peters, J. C. (2003). Obesity and the environment: Where do we go from here? *Science, 299,* 853-855.

Jackson, A. W., Morrow, J. R., Jr., Jensen, R. L., Jones, N. A., & Schultes, S. S. (1996). Reliability of the Prudential FITNESSGRAM trunk lift test in young adults. *Research Quarterly for Exercise and Sport, 67*, 115-117.

Joyner, A. B. (1997). Reliability of criterion-referenced standards of the FITNESSGRAM PACER. *Georgia Association for Health, Physical Education, Recreation, & Dance Journal, 31,* 14-15.

Kelder, S. H., Springer, A. S., Barrosso, C. S.,Smith, C. L., Sanchez, E., Ranjit, N., et al. (2009). Implementation of Texas Senate Bill 19 to increase physical activity in elementary schools. *Journal of Public Health Policy, 30*(Suppl. 1), s221-s247.

Mahar, M. T., Rowe, D. A., Parker, C. R., Mahar, F. J., Dawson, D. M., & Holt, J. E. (1997). Criterion-referenced and norm-referenced agreement between the mile run/walk and PACER. *Measurement in Physical Education and Exercise Science, 1*, 245-258.

Meredith, M. D., & Welk, G. J. (2007). FITNESSGRAM®: *Test administration manual* (4th ed.). Champaign, IL: Human Kinetics.

Morrow, J. R., Jr., & Ede, A. (2009). Statewide physical fitness testing: A BIG “waist” or a BIG “waste.” *Research Quarterly for Exercise and Sport, 80*, 696-701.

Morrow, J. R. Jr., Martin, S. B., & Jackson, A. W. (2010). Reliability and validity of the FITNESSGRAM®: Quality of teacher-collected health-related fitness surveillance data. *Research Quarterly for Exercise and Sport, 81*(3), s24-s30.

National Institute of Health (2010). *Exercise and physical activity: Getting fit for life.* Retrieved from http://www.nia.nih.gov/HealthInformation/Publications/exercise.htm

Ogden, C. L. & Carroll, M. D. (2010). *Prevalence of obesity among children and adolescents: United States, trends 1963-1965 through 2007-2008.* Retrieved from http://www.cdc.gov/nchs/data/hestat/obesity_child_07_08/obesity_child_07_08.pdf

Ogden, C. L., Carroll, M. D., & Flegal, K. M. (2008). High body mass index for age among US children and adolescents, 2003-2006. *Journal of the American Medical Association, 299,* 2401-2405.

Patterson, P., Rethwisch, N., & Wiksten, D. (1997). Reliability of the trunk lift in high school boys and girls. *Measurement in Physical Education and Exercise Science, 1,* 145-151.

Patterson, P., Wiksten, D. L., Ray, L., Flanders, C., & Sanphy, D. (1996). The validity and reliability of the back saver sit-and-reach test in middle school girls and boys. *Research Quarterly for Exercise and Sport, 67*, 448-451.

Physical Activity Guidelines for Americans (2008). *2008 Physical activity guidelines for Americans.* Retrieved from http://www.health.gov/PAGuidelines/

Sallis, J. F., Buono, M. J., Roby, J. J., Micale, F. G., & Nelson, J. A. (1993). Seven-day recall and other physical activity self-reports in children and adolescents. *American in Science and Sport and Exercise, 25,* 99-108.

Silverman, S., Keating, X. D., & Phillips, S. R. (2008). A lasting impression: A pedagogical perspective on youth fitness testing. *Measurement in Physical Education and Exercise Science, 12,* 146-166.

Texas Legislature Online (n.d.). Retrieved from http://www.legis.state.tx.us/tlodocs/80R/billtext/html/SB00530F.htm