The Physical Educator Vol. 71 • pp. 594–609 • 2014

TEST AND MEASUREMENT

Validity of Alternative Fitnessgram Upper Body Tests of Muscular Strength and Endurance Among Seventh and Eighth Grade Males and Females

Kalani Hobayan, Debra Patterson, Clay Sherman, Lenny Wiersma

Abstract

In a society in which obesity levels have tripled in the past 30 years, the importance of increased fitness levels within the academic setting has become even more critical. The purpose of this study was to investigate the validity of alternative Fitnessgram upper body tests of muscular strength and endurance among seventh and eighth grade males and females. The recommended test item used to assess students is the 90° Push-Up (90° PSU). However, the Fitnessgram provides alternative assessments to measure upper body strength: the Modified Pull-Up (MPU) and the Flexed Arm Hang (FAH). Adolescent males and females (N = 123) in seventh and eighth grades were administered all three muscular strength and endurance assessments. Males exemplified minimal acceptability for the 90° PSU–MPU and 90° PSU–FAH comparisons. Similar to the results for males, results for females indicated unacceptable reliability estimates for the 90° PSU–MPU and 90° PSU–FAH comparisons. As a result of this study, it is imperative that physical

Kalani Hobayan is a part-time lecturer, California State University Fullerton. Debra Patterson is a professor, Department of Kinesiology, California State University Fullerton. Clay Sherman is a professor, Department of Kinesiology, California State University Fullerton. Lenny Wiersma is a professor, Department of Kinesiology, California State University Fullerton. Please send author correspondence to khobayan23@gmail.com

educators and other administrators are aware that implementing the Fitnessgram alternative assessments of muscular strength and endurance may hinder and/or alter an adolescent's healthy fitness zone classification. Thus, future research regarding the muscular strength and endurance test items will ultimately promote a higher level of confidence among practitioners when using the test items interchangeably.

In a society in which obesity levels have tripled in the past 30 years, the importance of increased fitness levels within the academic setting has become even more critical (Morrow & Ede, 2009). Physical educators have indicated that school administration needs to stress the relevance of physical fitness testing in the same manner other academic tests are emphasized in schools (Martin, Ede, Morrow, & Jackson, 2010). Thus, heavy emphasis is being placed upon the Fitnessgram and other similar test batteries (e.g., The President's Challenge) to assess students' fitness levels within physical education.

The Fitnessgram health-related physical fitness test was developed by The Cooper Institute for Aerobics Research in 1999 and is currently endorsed by the Society of Health and Physical Educators (Shape America; Sherman & Barfield, 2006). The Fitnessgram test is used to determine whether students' physical fitness levels are acceptable for their appropriate age level and gender. The scores that students obtain are evaluated against norm- and criterion-referenced standards (Sherman & Barfield, 2006). The recommended test item used to assess muscular strength and endurance is the 90° Push-Up (90° PSU). However, the Fitnessgram includes alternative assessments for the 90° PSU. The Modified Pull-Up (MPU) and Flexed Arm Hang (FAH) are included in the Fitnessgram protocol as acceptable alternatives for muscular strength and endurance (Sherman & Barfield, 2006). The underlying assumption is that regardless of the assessment being initiated by the physical educator, students will ultimately obtain the same Healthy Fitness Zone (HFZ) classification.

To address this issue, researchers have investigated the reliability and validity of the tests of muscular strength and endurance. In 1993, Pate, Burgess, Woods, Ross, and Baumgartner implied that field tests of muscular strength and endurance were accurate measures of relative strength as opposed to absolute strength. Similar to Pate et al., McManis, Baumgartner, and Wuest (2000) focused their

attention on the reliability and objectivity of the 90° PSU. They revealed glimpses of acceptable reliability among the recommended test by obtaining reliability estimates (for high school and elementary males and females) that ranged from .50 to .86. In 1990, Cotten evaluated the factors associated with the MPU. Specifically, in his evaluation of the National Children and Youth Fitness Study II (NCYFS) MPU test, he concluded the alternative test item was effective due to the elimination of zero scores (Cotten, 1990). In 2004, Clemons et al. (2004) investigated the relationship among selected measures of muscular fitness and the FAH. They suggested that the FAH had a stronger correlation with relative strength as opposed to absolute strength and muscular endurance.

Although the reliability and validity of test items have been investigated in past studies, researchers have shifted their focus to investigating the equivalence reliability among the test items of muscular strength and endurance. In 2001, Romain and Mahar investigated the equivalence reliability of the 90° PSU–MPU comparison. They found that reliability estimates of the 90° PSU–MPU comparisons were unacceptable. Similar to Romain and Mahar, Sherman and Barfield (2006) looked at the equivalence reliability of the muscular strength and endurance test items. However, as opposed to comparing only the 90° PSU and MPU, they compared all three alternative test items (i.e., FAH, 90° PSU, and MPU). They indicated that although data for males resulted in acceptable estimates, data for females resulted in unacceptable reliability estimates.

To obtain a true measurement of muscular strength and endurance, future research needs to be geared toward investigating the reliability and validity of alternative upper body tests among different age groups. For physical educators, it is imperative for children to learn the importance of muscular strength and endurance, and it is equally critical for adolescents to develop strategies they may use throughout life (Welk & Meredith, 2008). The Fitnessgram protocols and literature indicate that the alternative testing procedures are reliable and provide students with the same result (i.e., HFZ). However, recent research supports the growing belief that implementing alternative test items of upper body strength may alter the HFZ classification of a student (Sherman & Barfield, 2006). Therefore, the purpose of this study was to investigate the validity of alternative Fitnessgram upper body tests of muscular strength and endurance among seventh and eighth grade males and females.

Methods

Participants

The participants (N = 123) were derived from one public (33 females, 28 males) and one private (25 females, 37 males) school in Southern California. Participants were aged 12 to 14 and were enrolled in seventh and eighth grade physical education classes. The researchers obtained permission from the principals, school administration, physical educators, and the institutional review board prior to testing. Parents and participants signed consent and assent forms to participate in the study.

Instruments and Procedures

Participants engaged in the Fitnessgram assessment measures. The procedures and guidelines of each test item were outlined in the Fitnessgram manual (Welk & Meredith, 2008). The principal investigator, who remained the same for each assessment, strictly followed the protocols. The upper body tests of muscular strength and endurance included the 90° PSU, MPU, and FAH.

The principal investigator collected the data within the study. Although the principal investigator had prior experience with the Fitnessgram protocols, the principal investigator reviewed the Fitnessgram manual (Welk & Meredith, 2008) for administrative procedures. The principal investigator engaged in practice assessments 6 months (i.e., pre-Fitnessgram testing) prior to data collection. In addition to assessments that targeted students' flexibility and aerobic capacity, data was collected for each test item of muscular strength and endurance (i.e., 90° PSU, MPU, and FAH). By engaging within practice trials, the principal investigator was able to modify factors that could improve the data collection process.

Due to the Fitnessgram test items being a component of the physical education curriculum at both schools, participants were experienced with the test items and protocols. Likewise, throughout the first school semester, participants were given additional practice and instruction regarding the alternative test items. Testing was conducted during the 50-min class session (2 days during the first week; 1 day during the second week) for 2 weeks. During the first week of testing, data collection occurred on Monday and Friday. The second week of data collection occurred on the following Tuesday. The testing format not only enabled participants proper recovery but also eliminated the possibility of delayed onset muscle soreness having a

significant impact. During the data collection process, the principal investigator counterbalanced the order of testing at the sites. On the days of testing and data collection, students were given extra time at the fitness stations to ensure that the data collection process was completed. The principal investigator returned the following week to collect make-up tests/results.

Following the Fitnessgram protocols, the principal investigator conducted the 90° PSU during the data collection process. Participants were divided into four groups of eight. Each group started at the FAH, MPU, 90° PSU, or other activity stations. Participants rotated every 5 to 7 min. While participants were engaged at the 90° PSU station, the principal investigator had each participant assume the push-up position. The principal investigator reviewed with participants the critical elements of the 90° PSU that had been established since the beginning of the year. Participants placed their hands shoulder width apart and their legs were straight and slightly apart. In the downward position, participants' elbows bent at a 90° angle and extended straight in the upward position (Welk & Meredith, 2008). While the assessment was being executed, the participants maintained a straight back. After questions or concerns were answered, the compact disc (from the Fitnessgram testing materials) with the recorded push-up cadence was played while participants engaged in the test. The test ended when the individual experienced discomfort due to fatigue or when a second form correction was made (Welk & Meredith, 2008).

The MPU test battery was incorporated into the second test item. Using the same dimensions (44 in. × 48 in. × 24 in.) and schematics of the Clovis Manufacturing Company, LLC (Clovis, California), who manufactures a Modified Pull-up Bar, the principal investigator used a replicated version of the Modified Pull-Up Bar measuring device. The traditional pull-up targets a student's entire body weight, whereas the MPU equipment allows students to be positioned at a 45° angle. This modification results in lesser body weight being pulled up, which ultimately leads to greater success. Participants were accustomed to the MPU instrument device from prior engagement. As participants approached the MPU station, they were given instruction regarding the format of the assessment and a review of its critical elements. With an overhand grip, participants pulled themselves up until their chest touched the elastic strap (hanging 7.5 in. from the bar) using only their arms and keeping their body straight (Welk & Meredith, 2008). The skill was repeated at a con-

trolled pace and no time limit was imposed. The principal investigator stopped the test when the participant experienced discomfort or when a second form correction was determined by the principal investigator (Welk & Meredith, 2008).

The FAH test battery was incorporated into the third assessment. The pull-up bars on the school sites (approximately 7 ft high and 3 ft wide) were used. The principal investigator reviewed the critical elements of the test item. While grasping the bar with an overhand grip, participants were required to hang with the chin above the bar as long as possible (Welk & Meredith, 2008). The stopwatch was started once the participant's chin was above the bar and the position was held as long as possible. The principal investigator stopped the test if the participant broke any of the critical elements.

Analysis

The 2011–2012 HFZ classifications of the Fitnessgram (12- to 14-year-old males and females) were used to classify participants as passing or failing the 90° PSU, MPU, and FAH. A passing classification was obtained when the participant met or exceeded the criterion-referenced standards for age and gender (Romain & Mahar, 2001). A failing classification occurred when the participant did not meet the criterion-referenced standards. The validity was determined by equivalence reliability estimates for the 90° PSU–MPU and 90° PSU–FAH comparisons. Similar to Sherman and Barfield (2006), the principal investigator in this study used percentage agreement (Pa) and modified kappa (Kq) to determine the relationships between variables. According to Looney (1989), percentage agreement and modified kappa should be included within reliability studies, particularly with nominal variables. Percentage agreement reflects the participant obtaining the same fitness zone classification on two tests and is influenced by chance agreement; the modified kappa reflects the participant obtaining the same fitness zone classification after controlling for chance agreement (Sherman & Barfield, 2006). These estimates of equivalence reliability were used to determine whether the fitness tests of the same construct (i.e., muscular strength and endurance) were consistent (Romain & Mahar, 2001). All statistical output regarding agreement statistics were calculated using the Online Kappa Calculator commissioned by Randolph (2008). Passing rate percentages among the muscular strength and endurance test items were reported with SPSS (version 19.0).

Results

The 90° PSU is the recommended test item for upper body strength and endurance in the Fitnessgram test. Thus, statistical output regarding percentage agreement and modified kappa were computed between the 90° PSU and other alternative test items (MPU and FAH). As Looney (1989) mentioned, percentage agreement and modified kappa estimates need to be identified within studies of reliability and validity. Percentage agreement values between .50 and 1.0 were deemed acceptable estimates; however, to establish equivalence reliability, percentage agreement values should be closer to 1.0 (Looney, 1989; Sherman & Barfield, 2006). Modified kappa values $\geq .75$ have been deemed excellent. Modified kappa values between .60 and .75 have been reported good, and estimates between .40 and .60 have been deemed acceptable (Morrow, Jackson, Disch, & Mood, 2000). Any modified kappa value lower than .40 has been termed an unacceptable estimate.

Percentage agreement and modified kappa estimates for the sample along with each age group (male and female) are presented in Table 1. Based on the criteria and statistical output recorded, the percentage agreement and modified kappa estimates for the 90° PSU–MPU comparison was unacceptable for the sample of the study (Pa = .68, Kq = .37) due to classification consistency not being demonstrated. The sample also produced unacceptable reliability estimates (Pa = .62, Kq= .24) for the 90° PSU–FAH comparison. Males and females aged 12 obtained acceptable estimates (Pa= .70, Kq = .40) for the 90° PSU–MPU comparison. On the contrary, the same age group acquired unacceptable reliability estimates (Pa = .54, Kq = .07) for the 90° PSU–FAH comparison due to lack of classification agreement. For males and females aged 13, reliability estimates were acceptable for the 90° PSU–FAH comparison (Pa = .71, Kq = .43) but unacceptable for the 90° PSU–MPU comparison (Pa = .63, Kq = .25). Last, 14-year-old males and females reported unacceptable reliability estimates (Pa = .62 to .67, Kq = .25 to .33) for the 90° PSU–MPU and 90° PSU–FAH comparisons.

Percentage agreement and modified kappa estimates were reported for males and females (see Table 2). Due to sample sizes being on the smaller end of the scale ($N \leq 30$), modified kappa estimates were reported for each gender (aged 12 to 14) collectively to enhance the general findings of the study. The validity for the 90° PSU–MPU comparison reported acceptable estimates for males aged 12 (Pa = .72, Kq = .44), with good agreement for males aged

14 (Pa = .81, Kq = .63). However, reliability estimates were not acceptable for males aged 13 (Pa = .69, Kq = .38), and males aged 12, 13, and 14 were combined (Pa = .66, Kq = .35) due to classification agreement not being demonstrated. Regarding the comparison between the 90° PSU and FAH, males aged 14 produced acceptable comparisons with good agreement (Pa = .81, Kq = .63). However, reliability estimates for males aged 12 and 13 were deemed unacceptable for the same comparisons (Pa = .62 to .63, Kq = .23 to .24). Last, males aged 12, 13, and 14 combined generated unacceptable reliability estimates due to a lack of classification consistency (Pa = .65, Kq = .29).

Table 1
Percentage Agreement and Modified Kappa Values Between the 90° Push-Up Test and the Alternative Tests (Modified Pull-Up, Flexed Arm Hang) of Upper Body Strength and Endurance

Age	Statistic	PSU–MPU	PSU–FAH
Total sample (N = 123)	Pa	.68	.62
	Kq	.37	.24
12 (n = 54)	Pa	.70	.54
	Kq	.40	.07
13 (n = 48)	Pa	.63	.71
	Kq	.25	.43
14 (n = 21)	Pa	.67	.62
	Kq	.33	.25

Note. MPU = modified pull-up; PSU = 90° push-up; FAH = flexed-arm hang. Pa= percentage agreement; Kq = modified kappa. Age is represented in years. Kq > .75 = Excellent; .60 ≤ Kq ≤ .75 = Good; .40 ≤ Kq ≤ .60 = Acceptable.

Table 2
Males and Females Percentage Agreement and Modified Kappa Values Between the 90° Push-Up Test and the Alternative Tests (Modified Pull-Up, Flexed Arm Hang) of Upper Body Strength and Endurance

		PSU–MPU		PSU–FAH	
Age	Statistic	M	F	M	F
12[a]	Pa	.72	.80	.62	.48
	Kq	.44	.60	.24	–.04
13[b]	Pa	.69	.68	.62	.63

Table 2 (cont.)

		PSU–MPU		PSU–FAH	
Age	**Statistic**	**M**	**F**	**M**	**F**
	Kq	.38	.36	.23	.27
14[c]	Pa	.81	.30	.81	.50
	Kq	.63	−0.4	.63	0.0
12, 13, and 14[d]	Pa	.66	.66	.65	.57
	Kq	.35	.31	.29	.14

Note. MPU = modified pull-up; PSU = 90° push-up; FAH = flexed-arm hang. M = males; F = females; Pa = percentage agreement; Kq = modified kappa. Age is represented in years. Kq > .75 = Excellent; .60 ≤ Kq ≤ .75 = Good; .40 ≤ Kq ≤ .60 = Acceptable.

[a]n_{males} = 29, $n_{females}$ = 25. [b]n_{males} = 25, $n_{females}$ = 22. [c]n_{males} = 11, $n_{females}$ = 10. [d]n_{males} = 65, $n_{females}$ = 58.

Reliability estimates for the 90° PSU–MPU comparisons were acceptable with good agreement for females aged 12 (Pa = .80, Kq = .60). In regard to females aged 13 and all females combined (aged 12, 13, and 14), reliability estimates were unacceptable (Pa = .66 to .68, Kq = .31 to .36). Unlike the other categories, females aged 14 obtained unacceptable estimates and lacked classification agreement (Pa = .30, Kq = −0.4) for the same comparison. Last, validity for the 90° PSU–FAH comparison was deemed unacceptable across all ages (Pa = .48 to .63, Kq = −0.04 to .27). Table 3 shows the 90° PSU–MPU and 90° PSU–FAH as an unacceptable or acceptable comparison for both genders.

Table 3

Summary of Reliability Estimates Between the 90° Push-Up Test and the Alternative Tests (Modified Pull-Up, Flexed Arm Hang) of Upper Body Strength and Endurance

Age	Gender	PSU–MPU	PSU–FAH
Total Sample		Unacceptable	Unacceptable
12	Male	Acceptable	Unacceptable
	Female	Acceptable	Unacceptable
13	Male	Unacceptable	Unacceptable
	Female	Unacceptable	Unacceptable

Table 3 (cont.)

Age	Gender	PSU–MPU	PSU–FAH
14	Male	Acceptable	Acceptable
	Female	Unacceptable	Unacceptable
12, 13, and 14	Male	Unacceptable	Unacceptable
	Female	Unacceptable	Unacceptable

Note. MPU = modified pull-up; PSU = 90° push-up; FAH = flexed-arm hang. Pa = percentage agreement; Kq = modified kappa. Age is represented in years. Kq > .75 = Excellent; .60 ≤ Kq ≤ .75 = Good; .40 ≤ Kq ≤ .60 = Acceptable.

Discussion

A lifestyle embedded with physical fitness and healthy eating habits is imperative for promoting lifelong physical activity. The physical education setting may be a learning environment in which students engage in activities that target muscular strength and endurance, body composition, aerobic capacity, and flexibility. Thus, students engage in fitness assessments that provide physical educators a tool to assess the overall level of students' health-related physical fitness and students an awareness of their own fitness levels. Following the Fitnessgram protocols and testing procedures, physical educators should promote an environment in which the guidelines are upheld and remain consistent throughout the assessment. Likewise, due to familiarization and understanding of the Fitnessgram protocols and procedures, physical educators are able to obtain results from students that are applicable and valid. When assessing for muscular endurance and strength in students, proctors typically use the traditional test item known as the 90° PSU. However, the Fitnessgram provides alternative assessments such as the MPU and FAH, which have been deemed as test items that also measure students' muscular strength and endurance (Sherman & Barfield, 2006).

Males

As the current study revealed glimpses of acceptable reliability estimates between the 90° PSU and MPU for males, estimates of classification agreement indicated that both test items do not effectively categorize males within the same HFZ classification. For physical educators to adopt the alternative assessments of the Fitnessgram, the test items need to obtain high levels of classification agreement (Romain & Mahar, 2001). Thus, the notion of using the

test items interchangeably is hindered due to its lack of classification consistency. When Romain and Mahar (2001) investigated the equivalence reliability of the 90° PSU and MPU among fifth and sixth grade subjects, the data indicated that 70% of males were similarly classified between the 90° PSU and MPU. In addition, they suggested the test items were unable to classify subjects similarly and deemed the reliability estimates unacceptable. Similar to Romain and Mahar, Sherman and Barfield (2006) concluded that although reliability estimates between the 90° PSU and MPU were acceptable for boys, estimates of classification agreement indicated the test items did not effectively classify subjects within the correct HFZ classification. Sherman and Barfield reported that 20% of boys were categorized differently between tests. In addition, regarding the classification agreement statistics, only 48% to 72% of males would ultimately receive the same HFZ classification on both tests.

As in these previous studies, acceptable estimates were reported among the 90° PSU–MPU comparisons for males in this study. Classification agreement estimates, however, hinder the conclusion that the test items could generate similar classifications of upper body strength and endurance. Similar to Romain and Mahar (2001), the the principal investigators in this study reported that approximately 74% of males were similarly classified between the 90° PSU and MPU. However, when influence of chance was removed, only 38% to 63% of males received the same classification. Thus, although the classification agreement estimates in the current study are slightly lower than those in Sherman and Barfield's (2006) study, the estimates suggest that test items cannot be used to classify students within the same HFZ classification effectively. Therefore, the validity cannot be concluded between the 90° PSU and MPU.

Similar to the reliability estimates of the 90° PSU–MPU comparison, the reliability estimates of the 90° PSU–FAH comparison for males exhibited acceptability but with insufficient classification agreement. Based on the passing rates of the muscular strength and endurance test items, the FAH was the least passed test item (25%) among males aged 12, 13, and 14. Statistical output showed that approximately 32% of males were classified under different HFZ classifications. This percentage is slightly greater than the 25% Sherman and Barfield (2006) reported. Similar to the reliability estimates, the continuum of classification agreement estimates were slightly larger than those Sherman and Barfield reported. As a result of reliability estimates being deemed unacceptable due to classification consis-

tency not being demonstrated, the validity of the FAH test item cannot be deemed equivalent.

Females

Equivalence reliability estimates for the 90° PSU–MPU comparison were unacceptable across all ages except females aged 12 (Pa = .80, Kq = .60). Excluding 12-year-old females, the percentage agreement ranged from .30 to .68, a continuum that is slightly larger than the .48 to .59 that Sherman and Barfield (2006) reported. Regarding the equivalence reliability estimates in further detail, approximate average of 40% of females (as a whole group) were classified differently between test items. Although a greater sample size ($N = 383$) was involved in Sherman and Barfield's study (2006), the current estimate is lesser than the 40% of girls in each age group that was reported. Sherman and Barfield and Romain and Mahar (2001) also reported unacceptable reliability estimates for the 90° PSU–MPU comparison among females (Pa = .69, Kq = .38). Although estimates are slightly lower in the current study for females aged 12, 13, and 14 (Pa = .66, Kq = .31), both pairs of researchers suggested that criterion-referenced classification needs to be addressed. Thus, to assess muscular strength and endurance for females effectively, physical educators and other practitioners are advised not to use the test items interchangeably due to the different HFZ classifications that result from the tests.

The 90° PSU–MPU comparison was only acceptable for females aged 12, whereas reliability estimates for the 90° PSU–FAH comparisons were unacceptable across all ages. Similar to the passing rate of males for the FAH, the passing rate of females aged 12, 13, and 14 for the FAH was the least passed test item (20%). Percentage agreement statistics ranged from .50 to .63, estimates similar to those Sherman and Barfield (2006) reported (Pa = .56 to .64). Classification agreement statistics revealed that the equivalence reliability of the comparison is unacceptable. As a result of the unacceptable estimates of reliability, physical educators should caution the usage of the 90° PSU and FAH interchangeably among females.

Conclusion

This study revealed that the validity of the 90° PSU and alternative test items were unacceptable in most cases for middle school adolescents. The perception is that whatever test is being implemented, extending testing from the laboratories to the field should not have a negative impact on the reliability and validity of the data

collected (Morrow, Martin, & Jackson, 2010). Although some reliability estimates within the study revealed signs of statistical acceptability, classification agreement estimates indicate that test items of muscular strength and endurance were unacceptable comparisons. A delimiting factor associated within the study was the size of the sample. The researchers of this study incorporated a sample of 123 students, a relatively smaller sample size than Sherman and Barfield (2006) used (N = 383). However, the current study was geared toward applying supportive knowledge to the foundational research established by Romain and Mahar (2001) and Sherman and Barfield.

For the Fitnessgram to be used to measure muscular fitness and health-related fitness, the classification across future tests and criteria needs to be consistent (Sherman & Barfield, 2006). Due to the limited amount of research and background information regarding the validity of the Fitnessgram upper body tests, many physical educators are posed with confusion whether to use the muscular strength and endurance test items interchangeably. Romain and Mahar (2001) established the foundation of the initial equivalence reliability study among subjects in Grades 5 and 6. They suggested that the reliability of the 90° PSU–MPU unacceptable and that criterion-referenced classifications need to improve for test items (90° PSU and MPU) to be deemed valid. Sherman and Barfield (2006) implemented all three alternative tests of muscular strength and endurance (FAH, PU, and MPU). They also used a large sample size of subjects within Grades 3 to 6 (N = 383). Like Romain and Mahar, Sherman and Barfield found consistent themes, particularly with the need to adjust and modify alternative test items classifications. The current study adds another pillar to the equivalence reliability of the Fitnessgram test. Using a different population as the sample (Grades 7 and 8), the researchers in this study did not find concrete comparisons and acceptability between the test items.

Further studies need to continue the ongoing development of established HFZ classifications (i.e., criterion-referenced standards). Cuerton and Warren (1990) stated that criterion-referenced assessments have predetermined standards that represent a desired and specified level of performance. Although The Cooper Institute has provided desired standards, Sherman and Barfield (2006) mentioned that no research has been attempted to validate the criterion-referenced standards incorporated in the Fitnessgram test. They noted a method of equating test items of muscular strength and endurance. By equating test items, practitioners will be allowed to determine

whether a score from one test items relates to that of the "gold standard" (Sherman & Barfield, 2006). This method could eliminate the possibility of subjects being misclassified under different HFZ classifications. The current HFZ (2011–2012) classifications of the Fitnessgram exemplify a transition into the adaptation and modification of merging test item classifications, particularly upper body strength and endurance. Although some age group classifications have remained consistent, the upper body strength and endurance classifications coincide more effectively. Modifying standards that are consistent will ultimately promote a higher level of confidence among practitioners when using the test items interchangeably.

Although many suggestions and considerations have been implied toward the Fitnessgram tests battery, The Cooper Institute has continued its efforts not only to merge classifications but also to adjust test items to suit present-day society. Although the 90° PSU, MPU, and FAH tests were measured and assessed among subjects, the pull-up (PU) assessment was eliminated from the current Fitnessgram HFZ classifications. One significant problem with the PU assessment was the inability to use the test to differentiate among individuals at the lower end of the scale (Rutherford & Corbin, 1994). Due to the limited success and subjects' inability to execute one pull-up repetition effectively, officials at The Cooper Institute have directed physical educators to incorporate current assessments. Just as the PU was eliminated from the Fitnessgram criterion-referenced standards, the FAH test item should be considered for removal based on the results of this study. Although passing rates of the FAH were not as significantly low as the PU assessment, this study revealed that the FAH is not a valid measure of upper body strength and endurance. Again, future editions of the Fitnessgram should be written with the effectiveness of the FAH test item in mind.

If the 90° PSU remains as the recommended test item to assess muscular strength and endurance, classification consistency needs to continue to improve within the Fitnessgram. Adapting and modifying criterion-referenced standards to suit present-day society will increase the level of acceptability of the Fitnessgram among physical educators and other practitioners. As opposed to identifying alternative test items as equivalent measures of muscular strength and endurance, the Fitnessgram should contain suggestions that the alternative assessments support and strengthen the recommended test item. Likewise, practitioners need to be aware of the value of implementing alternative test items to obtain an additional measurement

of subjects' upper body strength. Further research related to the validity of the test items is needed to develop a greater understanding of the effectiveness of Fitnessgram among society.

References

Clemons, J. M., Duncan, C. A., Blanchard, O. E., Wendel, H. G., Hollander, D. B., & Doucet, J. L. (2004). Relationships between the flex-arm hang and select measures of muscular fitness. *Journal of Strength and Conditioning Research, 18*(3), 630–636.

Cotten, D. J. (1990). An analysis of the NCYFS II modified pull-up test. *Research Quarterly for Exercise and Sport, 61*, 272–274.

Cuerton, K. J., & Warren, G. L. (1990). Criterion-referenced standards for youth health-related tests: A tutorial. *Research Quarterly for Exercise and Sport, 61*, 7–19.

Looney, M. (1989). Criterion-referenced measurement: Reliability. In M. Safrit & T. Wood (Eds.), *Measurement concepts in physical education and exercise science* (pp. 137–152). St. Louis, MO: Mosby.

Martin, S. B., Ede, A., Morrow, J. R., & Jackson, A. W. (2010). Statewide physical fitness testing: Perspectives from the gym. *Research Quarterly for Exercise and Sport, 81*(Suppl. 2), S31–S41.

McManis, B. G., Baumgartner, T. A., & Wuest, D. A. (2000). Objectivity and reliability of the 90 degree push-up test. *Measurement in Physical Education and Exercise Science, 4*, 57–67.

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., Jackson, A. W., Disch, J. G., & Mood, D. P. (2000). *Measurement and evaluation in human performance* (2nd ed.). Champaign, IL: Human Kinetics.

Morrow, J. R., 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*(Suppl. 2), S24–S30.

Pate, R., Burgess, M., Woods, J., Ross, J., & Baumgartner, T. (1993). Validity of field tests of upper body muscular strength. *Research Quarterly for Exercise and Sport, 64*, 17–24.

Randolph, J. J. (2008). Online kappa calculator. Retrieved February 26, 2012, from http://justus.randolph.name/kappa

Romain, B. S., & Mahar, M. T. (2001). Norm-referenced and criterion-referenced reliability of the 90 degree push-up and modified pull-up. *Measurement in Physical Education and Exercise Science, 5*, 67–80.

Rutherford, W. J., & Corbin, C. B. (1994). Validation of criterion-referenced standards for tests of arm and shoulder girdle strength and endurance. *Research Quarterly for Exercise and Sport, 65*, 110–119.

Sherman, T., & Barfield, J. P. (2006). Equivalence reliability among the FITNESSGRAM® upper body tests of muscular strength and endurance. *Measurement in Physical Education and Exercise Science, 10*(4), 241–254.

Welk, G. J., & Meredith, M. D. (2008). *Fitnessgram / Activitygram reference guide*. Dallas, TX: The Cooper Institute.