Research Quarterly for Exercise and Sport
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/urqe20

Measurement Agreement Between Estimates of Aerobic Fitness in Youth: The Impact of Body Mass Index

Pedro F. Saint-Maurice [a] , Gregory J. Welk [a] , Kelly R. Laurson [b] & Dale D. Brown [b]
[a] Iowa State University
[b] Illinois State University
Published online: 21 Feb 2014.

To cite this article: Pedro F. Saint-Maurice , Gregory J. Welk , Kelly R. Laurson & Dale D. Brown (2014) Measurement Agreement Between Estimates of Aerobic Fitness in Youth: The Impact of Body Mass Index, Research Quarterly for Exercise and Sport, 85:1, 59-67, DOI: 10.1080/02701367.2013.872217

To link to this article: http://dx.doi.org/10.1080/02701367.2013.872217

Research Quarterly for Exercise and Sport, *85*, 59–67, 2014

ISSN 0270-1367 print/ISSN 2168-3824 online
DOI: 10.1080/02701367.2013.872217

Routledge
Taylor & Francis Group

Measurement Agreement Between Estimates of Aerobic Fitness in Youth: The Impact of Body Mass Index

Pedro F. Saint-Maurice and Gregory J. Welk
Iowa State University

Kelly R. Laurson and Dale D. Brown
Illinois State University

Purpose: The purpose of this study was to examine the impact of body mass index (BMI) on the agreement between aerobic capacity estimates from different Progressive Aerobic Cardiorespiratory Endurance Run (PACER) equations and the Mile Run Test. **Method:** The agreement between 2 different tests of aerobic capacity was examined on a large data set from 2 suburban school districts ($n = 1{,}686$ youth in Grades 3–10). Difference estimates between the Mile Run Test and several PACER equations were computed, and residuals were examined using cluster analysis. The implication of the discrepancy between these tests was also examined using FITNESSGRAM® health-related standards for BMI. Comparisons were made against corresponding estimates of peak oxygen consumption from the Mile run because this equation is more established. **Results:** Results supported a 2-cluster solution. The discrepancy between tests was higher in participants with higher BMI scores (*Z* scores for residuals in this group ranged from −0.07 to 1.57). BMI was able to explain 30% to 34% of the disagreement between the Mile and different PACER equations of aerobic fitness. Classification analyses revealed that kappa scores were lower among PACER equations that do not include a BMI term (kappa = .12–.34 vs. .59–.81). Overall, the test-equating approach used in the Fitnessgram program to process PACER data had better agreement than alternative PACER equations that included BMI. **Conclusion:** The results support the inclusion of BMI in prediction equations used to estimate aerobic capacity from the PACER.

Keywords: children, FITNESSGRAM®, Mile Run Test, PACER

The FITNESSGRAM® program provides teachers with a variety of methods to assess cardiovascular fitness (e.g., the Progressive Aerobic Cardiorespiratory Endurance Run [PACER] test, Mile run, Mile walk test), and each has been shown to have acceptable reliability and validity (Cureton & Plowman, 2008). A key advantage of the Fitnessgram program is that the tests are scored using age- and gender-specific, criterion-referenced standards (e.g., healthy fitness zone [HFZ]) so that the results can be compared regardless of what test is used.

Although the previous standards were empirically sound, new health-related Fitnessgram standards were recently developed to improve utility for school assessments (Welk, Going, Morrow, & Meredith, 2011). The new standards were developed using nationally representative health and fitness data from the National Health and Nutrition Examination Survey and have documented predictive utility for detecting risk for metabolic syndrome (Welk, Laurson, Eisenmann, & Cureton, 2011). Another major change with the new standards is that the aerobic capacity estimates from the PACER are now calculated based on a test-equating

Submitted February 12, 2012; accepted February 7, 2013.
Correspondence should be addressed to Pedro F. Saint-Maurice, Department of Kinesiology, Iowa State University, Room 164J, Ames, IA 50011. E-mail: pedrosm@iastate.edu

methodology developed by Zhu (Zhu, Plowman, & Park, 2010). This approach was selected, in part, because it offered a solution to correct for the observed differences in the classification of aerobic fitness by the PACER and the Mile run (Welk, Meredith, Lhmels, & Seeger, 2010). Scores on the PACER (in laps) are now transformed to estimated Mile run times (Zhu et al., 2010), and these times are then processed using the same prediction equation developed by Cureton et al. (Cureton, Sloniger, O'Bannon, Black, & McCormack, 1995) for the Mile Run Test.

A recent study (Welk, Saint-Maurice, Laurson, & Brown, 2011) revealed that the new method yielded improved classification agreement between the Mile run and PACER scores compared with the previous standards/method. However, it is not clear if the improvement is the result of the new standard or the test-equating methodology used to process the PACER data. It is also not clear if the improvement is due to the inclusion of body mass index (BMI) in the calculation of aerobic capacity from the PACER test. The Leger equation, previously used to process the PACER data (Leger, Mercier, Gadoury, & Lambert, 1988), was based solely on laps and age, while the Mile run Equation (Cureton et al., 1995) includes a BMI term. It is possible that the differential use of BMI contributed to some of the previously described (Welk et al., 2010) differences in aerobic fitness classification and the improved outcomes with the new method.

In general, the potential impact of BMI/body weight on aerobic fitness performance has largely been ignored in the exercise science field, and it clearly deserves more systematic evaluation. It is well established that body mass or/and body composition have a negative impact on endurance test performance. This impact is attributed to a higher-energy demand at submaximal speeds, despite no differences in maximal absolute oxygen consumption (VO_2; Cureton et al., 1978). The strong association between body weight and performance on these tests is a disadvantage for heavier individuals while favoring lighter individuals. Therefore, the inclusion/exclusion of correlates of body fat measures, such as BMI, on aerobic fitness prediction equations might lead to discrepancies between field tests.

It is important to better understand the contributions that weight (and BMI) makes to the agreement between field tests of aerobic fitness. Therefore, the purpose of this study was to specifically examine the impact of BMI on the agreement between the Mile Run Test and different PACER-derived peak oxygen consumption (VO_2peak) estimates. The study will evaluate outcomes from the previously used PACER scoring method (Leger et al., 1988), the new test -equating methodology (Zhu et al., 2010), as well as a number of alternative PACER equations, including recently released versions with and without BMI (Mahar, Guerieri, Hanna, & Kemble, 2011).

METHODS

Participants

The study utilized previously analyzed data collected through an ongoing university–K–12 partnership that facilitated the collection of fitness data through school physical education programming (Welk, Saint-Maurice, et al., 2011). This data set included fitness results from 1,711 youth (3rd–12th grade) from four schools (two district and two private schools in a small town in the Midwestern United States.

Data Collection

Participants completed both the Mile and the PACER in a counterbalanced design to facilitate evaluation of classification agreement. Trained graduate research assistants collected data during normal physical education sessions during a 2-week time span using standard Fitnessgram protocols. The project was approved by the local school board and by the university's institutional review board. All participants provided signed parent consent forms and completed youth assent forms.

Data Processing

Data from the PACER test and Mile were converted into estimates of aerobic capacity using the available equations and conversions. The Cureton Equation (Cureton et al., 1995) was used to estimate peak aerobic capacity from the Mile run (VO_2peak; Cureton et al., 1995).

A variety of PACER equations were also tested to evaluate relative agreement with the values from the Mile run. Some of the equations included a BMI term while others did not. The Leger equation previously used in Fitnessgram (Leger et al., 1988) includes terms for maximal speed, age, and a Speed × Age interaction term (Leger et al., 1988). An equation published by Barnett et al. (Barnett, Chan, & Bruce, 1993) includes gender, age, and speed. Several equations developed by Mahar and colleagues (2011) were also evaluated: a quadratic equation that included BMI (MaharQ_BMI), a linear equation without BMI (Mahar), and another linear equation with a BMI term (Mahar_BMI). The specific equations for predicting aerobic capacity (VO_2peak in mL/kg/min) are listed below to facilitate comparison:

$$\text{Cureton}: \ VO_2\text{peak} = (-8.41 \times min) + (0.34 \times min^2) + (0.21 \times (age \times gender)) - (0.84 \times BMI) + 108.94$$

$$\text{Leger}: \ VO_2\text{peak} = 31.025 + (3.238 \times speed) - (3.248 \times age) + (0.1536 \times (speed \times age)$$

$$\text{Mahar}: \ VO_2\text{peak} = 32.5694 + (0.2730 \times PACERlaps) + (3.2522 \times sex) + (0.0296 \times age)$$

$$\text{Barnett}: \ VO_2\text{peak} = 24.2 - 5.0(gender) - 0.8(age) + 3.4(maximal\ speed)$$

$$\text{Zhu\_BMI}: \ VO_2\text{peak} = \text{based on a test–equating approach (converts laps to Mile run time)}$$

$$\text{Mahar\_BMI}: \ VO_2\text{peak} = 40.3453 + (0.2143 \times PACERlaps) + (4.2729 \times sex) + (0.7944 \times age) - (0.7947 \times BMI)$$

$$\text{MaharQ\_BMI}: \ VO_2\text{peak} = 41.76799 + (0.49261 \times PACERlaps) - (0.00290 \times PACERlaps^2) - (0.61613 \times BMI) + (0.34787 \times sex \times age)$$

Note: The Cureton equation includes: age in years, gender as "0" if female and "1" if male, BMI in kg/m^2, and Mile run time in minutes. The Leger equation includes: speed as the maximal speed corresponding to the last stage completed in the PACER test and age in years. The Mahar equations include: sex coded as 1 if male and 0 if female, age in years, BMI expressed in kg/m^2, and PACERlaps corresponds to the maximum number of laps completed during the PACER test. The Barnett equation includes: gender coded 0 if male and 1 if female, age in years, and maximal speed as the maximal speed corresponding to the last PACER stage. The Zhu_BMI equation includes: the number of total laps completed during the PACER test converted to Mile run time in minutes and scored with the Mile run equation.

Data Analyses

Descriptive analyses were first conducted to examine the distribution of fitness classification among different age groups. Age is an important factor in the study because previous studies have indicated discrepancies in the validity of the PACER test in different age groups (Mahar et al., 2011; Mahar, Welk, Rowe, Crotts, & Mclver, 2006). Therefore, the sample was divided into a younger (11–14 years old) sample and an older (15–18 years old) sample. The age-specific breakpoints are intended to reflect the general distinctions between a middle school sample and a high school sample, respectively.

Analyses were then conducted to examine measurement agreement between different PACER equations and the Mile run. Pearson product–moment correlations were computed among the various measures to evaluate overall associations between the two assessments. Differences in estimated scores between the Mile and different PACER estimates (e.g., VO_2peak Mile – VO_2peak Leger; VO_2peak Mile – VO_2peak Mahar_BMI) were evaluated using paired t tests with alpha set at .05. Measures of effect size were computed using Cohen's d for standardized mean differences.

Cluster analyses were then performed on the standardized residuals to determine how the different Mile–PACER estimates varied with participants' BMI. An advantage of cluster analyses is that it is possible to examine the effect of BMI as a continuous measure instead of using existing BMI classifications. Residuals were first computed for each of the PACER estimates relative to the Mile run estimate (e.g., VO_2peak Mile – VO_2peak Leger). Separate residuals were then created for subgroups of participants (by age group), and the absolute values were then computed to be used in the cluster analysis technique. The use of absolute values provides a better evaluation for cluster analyses because it reflects total error. However, raw residual scores were later standardized to facilitate interpretation of differences between resultant clusters (a standardized residual for the PACER of 0.5, for example, means that the difference between the Mile test and the PACER estimate is on average 0.5 standard deviations greater from the overall distribution of residuals in the subgroup being considered). BMI values were also standardized (Z scores) to allow differences due to BMI to be further examined.

The cluster analyses were conducted using Ward's method based on squared Euclidian distance. This commonly used approach maximizes cases further apart and therefore increases the differences between clusters (Aldenderfer & Blashfield, 1984). Standard cluster analysis techniques were used to determine the final cluster solution—including visual inspection of the dendrogram, evaluation of the fusion coefficient, comparisons of indexes showing loss of information when two clusters/cases are merged, and interpretation of expected and actual R^2. The number of clusters was decided based on the maximization of the actual R^2 while minimizing the number of clusters. In other words, the analyses determined the minimal number of clusters that could be retained without affecting the total amount of variability to be explained in the data. This step is critical and requires a balance between the number of clusters (a low number of clusters is usually preferred to facilitate interpretation) and the maximization of the proportion of variability explained by the variable in the model (e.g., absolute residual scores associated with the Zhu procedure). The actual R^2 in cluster analysis is an overall indicator of the fit of the final cluster solution to the data. Therefore, this indicator was selected to determine the impact of selected variables on the disagreement among the Mile and several PACER estimates. Cluster validation was performed using logistic regression, although those results were not presented.

The final step in the analyses involved a systematic reporting of classification agreement based on the new Fitnessgram criterion-referenced standards: HFZ, needs improvement/some-risk zone, and needs improvement/high-risk zone (Welk, Laurson et al., 2011). Kappa statistics were first computed for each equation, resulting in 6 contingency tables (e.g., Zhu equation classifications vs. Mile run classifications). Similar agreement analyses were conducted by age group and gender resulting in 24 distinct

contingency tables (e.g., Zhu equation classifications vs. Mile run classifications for boys in the younger age group). Contingency tables were also computed separately by gender and BMI cluster (e.g., Zhu equation classification vs. Mile run classifications for younger participants in the "low BMI" cluster). Percent agreement and weighted kappa statistics (Cohen, 1960, 1968) were computed to understand the practical impact of using different equations. This resulted in six models. In other words, we quantified the extent to which field tests can provide similar estimates of peak aerobic capacity. Kappa was interpreted as either poor (less than .20), fair (.20–.40), moderate (.40–.60), good (.60–.80), and very good (.80–1.00; Altman, 1990).

RESULTS

Descriptive Results

The final sample included 911 male participants and 755 female participants with complete data on both the PACER and the Mile Run Test. Results were analyzed separately for 11- to 14-year-olds (Group 1: $M_{age} = 12.44 \pm 1.03$ year, $n = 1{,}170$) and 15- to 18-year-olds (Group 2: $M_{age} = 16.25 \pm 0.25$ years, $n = 516$). Based on the new Fitnessgram BMI standards, approximately 65% of the younger sample (Group 1) was classified into the HFZ, 12% were classified into the needs improvement/some-risk zone, and 23% were classified in the needs improvement/high-risk zone. The associated percentages for the older sample (Group 2) were 71%, 11%, and 18%, respectively.

Measurement Agreement: Comparisons of PACER Versus Mile

Table 1 illustrates average mean differences between the Mile Run Test (referent) and different estimates of VO_2peak using PACER-derived equations. The Barnett equation yielded significantly higher estimates of aerobic capacity than did the Mile Run Test, and this pattern was true for both age groups ($p < .05$; $d = -0.40$ and $d = -0.29$ for the younger and older age groups, respectively). In contrast, the Leger equation yielded significantly lower aerobic capacity estimates for both the younger and older age groups; however, this effect was small in the younger age group ($p < .05$; $d = 0.12$ and $d = 0.47$ for the younger and older age groups, respectively). The estimates from the Mahar, Mahar_BMI, and MaharQ_BMI equations yielded significantly lower estimates than did the Mile run in the younger group, but all of these differences, when standardized, were less than 0.2 standard deviations apart from the Mile run scores ($p < .05$; d ranged from 0.07 to 0.18) and values were all within 1 mL/kg/min. The same comparisons indicated that estimates from the Mahar, Mahar_BMI, and MaharQ_BMI equations for the older age group were significantly higher and at least −0.2 standard deviations from the Mile run estimates ($p < .05$; d ranged from −0.21 to −0.50). The Zhu_BMI equation yielded estimates within 1 mL/kg/min for both age groups, and the largest discrepancy was found in the younger age group even though the effect size was equal to 0.15 ($p < .05$). Correlations provide another indicator of agreement. The Mile test VO_2peak scores were strongly associated with Zhu_BMI ($r = .87$), Mahar_BMI ($r = .83$), and

TABLE 1
Predicted VO_2peak (ml/kg/min) for Both Younger and Older Adolescents Using Different Regression Equations

		Mean ± SD	*Mdifference ± SE*	*95% CI*	*t*	*d*
11–14 y						
	Mile	45.10 ± 6.28	referent			
	Leger	44.40 ± 5.40	0.69 ± 0.15*	[0.66, 1.34]	4.61	0.12
	Mahar	44.70 ± 5.98	0.40 ± 0.15*	[0.20, 0.85]	2.75	0.07
	Barnett	47.33 ± 4.79	−2.23 ± 0.14*	[−2.31, −1.67]	−15.81	−0.40
	Zhu_BMI[a]	44.18 ± 6.17	0.92 ± 0.09*	[0.92, 1.41]	9.69	0.15
	Mahar_BMI[a]	43.86 ± 7.38	1.23 ± 0.12*	[0.85, 1.39]	10.66	0.18
	MaharQ_BMI[a]	44.35 ± 7.17	0.74 ± 0.12*	[0.50, 0.98]	6.06	0.11
15–18 y						
	Mile	44.58 ± 7.26	referent			
	Leger	41.19 ± 7.03	3.39 ± 0.24*	[2.92, 3.86]	14.31	0.47
	Mahar	46.63 ± 5.01	−3.54 ± 0.23*	[−3.99, −3.09]	−14.11	−0.33
	Barnett	46.50 ± 5.80	−1.92 ± 0.22*	[−2.36, −1.48]	−8.59	−0.29
	Zhu_BMI[a]	44.65 ± 7.06	−0.08 ± 0.15	[−0.36, 0.21]	−0.52	−0.01
	Mahar_BMI[a]	48.20 ± 7.34	−3.62 ± 0.16*	[−3.94, −3.30]	−18.05	−0.50
	MaharQ_BMI[a]	46.17 ± 7.64	−1.58 ± 0.18*	[−1.95, −1.23]	−8.71	−0.21

Note. SD = standard deviation; SE = standard error of the mean difference; CI = confidence interval; d = Cohen's *d* for standardized mean differences.

*Significantly different from Mile test with $p < .05$.

[a]Include BMI in the prediction equation.

MaharQ_BMI ($r = .81$). Estimates derived from equations without a BMI term were moderately associated with the Mile estimates ($r_{Leger} = .65$, $r_{Mahar} = .67$, $r_{Barnett} = .68$).

Cluster Analyses

Cluster analyses provide a way to examine the impact of BMI on the differences between aerobic capacity estimates from the PACER and the Mile run. The essence of the technique is to maximize the amount of explained variance with the fewest number of clusters. The preliminary analyses of the dendrogram and r^2 distributions suggested that the optimum number of clusters was five in the younger group (Figure 1) and three in the older group (Figure 2). Figure 1 shows that the amount of explained variability in the younger group starts to plateau at a value of 5 and matches the expected R^2. Figure 2 provides a similar plot for the older sample, but in this case, the plateau in explained variability starts at around three or four clusters. The various solutions (five, four, three, and two clusters in the younger age group; four, three, and two clusters in the older age group) were examined in more detail to determine the most optimal solution. The inclusion of a larger number of clusters increased the overall explained variance (r^2), as shown in Figures 1 and 2, but the two-cluster solution was retained because three-, four-, and five-cluster solutions did not yield clear distinctions among clusters regarding BMI scores. The proposed two-cluster solution explained 34% (11- to 14-year-old age group) and 30% (15- to 18-year-old age group) of the total variance in each of the respective samples being considered. The average age was similar between the clusters (lower BMI vs. higher BMI) for both the younger age group (12.42 ± 1.03 vs. 12.55 ± 1.05) and the older age group (16.24 ± 0.95 vs. 16.28 ± 0.95). However, the two-cluster solution was clearly distinguishable by greater and lesser values of BMI in both the 11- to 14-year-old age group (28.24 ± 4.88 vs. 19.36 ± 2.94) and 15- to 18-year-old age group (29.15 ± 5.81 vs. 21.47 ± 2.70). The lower BMI group had, on average, individuals with lower BMI relative to the overall sample being considered for analysis ($Z_{11\text{-to}14\text{-year-olds}} = -0.29$ and $Z_{15\text{-to}18\text{-year-olds}} = -0.35$). In contrast, the higher BMI group included individuals with higher relative BMI scores ($Z_{11\text{-to}14\text{-year-olds}} = 1.62$ and $Z_{15\text{-to}18\text{-year-olds}} = 1.28$). Therefore, Cluster 1 was defined as the "lower BMI group" and Cluster 2 was defined as the "higher BMI group" (Table 2).

The distribution of error for the two-cluster solution is illustrated in Figure 3. The lower BMI group was characterized by lower-than-average and average absolute residuals in all different estimates (Z scores between −0.02 and −0.41). In contrast, the higher BMI group had higher-than-average scores in almost all different estimates (Z scores ranged from −0.07 to 1.57). The higher BMI group also had much larger absolute differences between the Mile and other estimates. In the younger group, the average absolute differences were 9.38 ± 3.50 mL/kg/min in the higher BMI group but 4.83 ± 3.34 mL/kg/min in the lower BMI group. In the older age group, the average absolute difference was 12.26 ± 4.02 mL/kg/min in the higher BMI group but 6.21 ± 3.43 mL/kg/min in the lower BMI group. The discrepancies between the two clusters were consistently larger in equations that did not include BMI (shown on left side of Figure 3) compared with those that did include BMI (shown on right side of Figure 3). An exception was the Leger equation in the 15- to 18-year-old age group. This finding can be possibly explained by age and anthropometric sample differences among the different calibration studies. Overall, the results suggest a greater overall amount of error in the cluster with higher BMI scores. The Zhu_BMI equation had the lowest error associated in both age groups (range = 1.76–3.54 mL/kg/min). Interestingly, the absolute error pattern was similar for

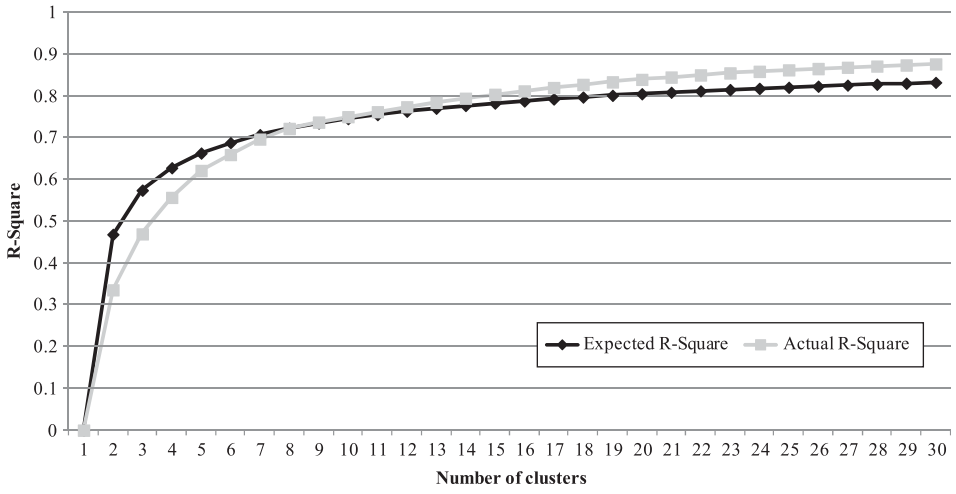

FIGURE 1 Distribution of the number of clusters (ranging from 0 to 30) according to the expected and actual R^2 for the younger age group.

both BMI cluster solutions (high vs. low BMI groups) and also in both age groups.

The direction of the residuals was further investigated by comparing the raw difference between the Mile and different PACER equations for different BMI subgroups. Figure 4 depicts the magnitude and direction of error for participants classified into the three new health-related fitness zones. The left side of the image shows residuals for equations that do not include BMI, while the right side shows residuals for several equations that do include BMI. Overall, the equations that do not include BMI overestimate aerobic capacity (relative to the Mile run) for youth in the *needs improvement*/some-risk zone or the *needs improvement*/high-risk zone. The Leger and Mahar equation underestimated aerobic capacity (relative to the Mile run) for those who achieve the HFZ for BMI. The distribution of aerobic capacity difference scores was more consistent (less variability) in equations that include BMI in their estimates. All the equations that include BMI underestimated aerobic capacity in all three BMI classification groups. The patterns were consistent for both the younger (top panel) and older (bottom panel) age groups with some minor exceptions. The two-cluster solution was further validated using logistic regression. Overall, higher absolute residuals derived from all equations were most likely (significantly) associated with the higher BMI group (data not shown).

Classification Analyses

Classification agreement based on Fitnessgram BMI standards provides an additional indicator of agreement between the Mile test and the various PACER equations. Classification agreement varied considerably among the different equations; however, all seemed to have reasonable agreement (66.1%–88.5%), with the highest agreement found for the Zhu_BMI approach and the lowest for the Leger equation. The Barnett equation absolute agreement with the Mile test was equal to 74.2%. Similar results were found for kappa scores (kappa ranged from .12 to .79) except for the Barnett equation. The Zhu equation had the highest index of agreement, while the Barnett equation had the lowest agreement with Mile classification scores. The Barnett equation did not classify any individual as "high risk"; therefore, to obtain kappa statistics for this algorithm, we simulated a perfect agreement between the Mile and Barnett for one individual score. The Mahar_BMI (kappa = .64) and MaharQ_BMI (kappa = .59) had an overall moderate agreement with the Mile test, and fair agreement was found for the Leger (kappa = .33) and Mahar (kappa = .31) equations. These results used combined scores from both age groups. Differences in agreement were noted when data were processed separately by age group, gender, and BMI cluster. All of those are described in the following paragraph; however, only age group differences are provided in Table 2.

Kappa scores were similar between age groups (kappa ranged from .12 to .78 and from .27 to .81 for the younger and older age groups, respectively). Moreover, in the younger age group, agreement was higher in male participants than in female participants, with greater discrepancies found in the two BMI Mahar equations: .74 for male participants vs. .57 for female participants, and .66 for male participants vs. .52 for female participants. This trend was reversed in the older age group. Kappa scores were higher in female participants than in male participants; however, there were no major discrepancies at the gender level.

DISCUSSION

The main goal of the study was to examine the impact of BMI on fitness estimation and classification agreement. This is an important issue for school-based fitness-testing

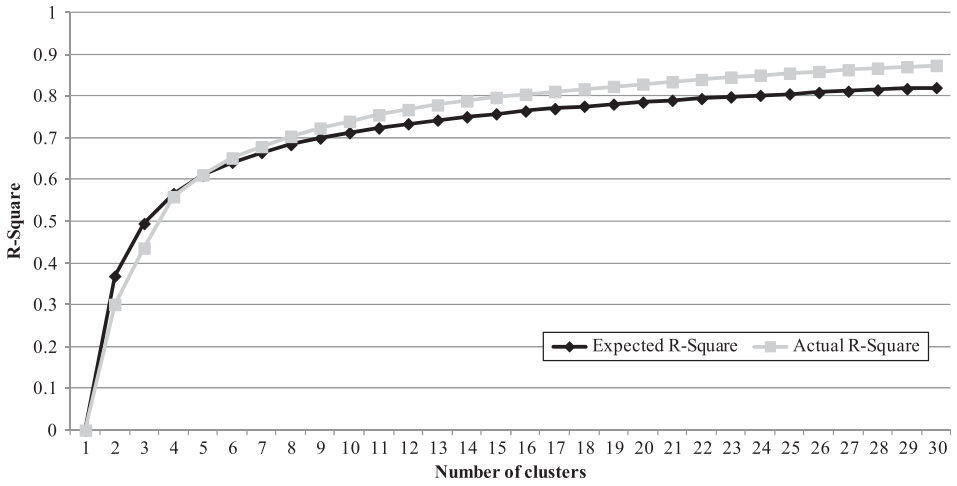

FIGURE 2 Distribution of the number of clusters (ranging from 0 to 30) according to the expected and actual R^2 for the older age group.

TABLE 2
Kappa Agreement Scores Between the Mile Test and Other PACER Estimates Using Fitnessgram Criterion-Referenced Standards

	Combined	*LBMI Group*	*HBMI Group*
11- to 14-year-olds			
Leger	.29	.22	.15
Mahar	.34	.36	.12
Barnett[b]	.12	.09	NA
Zhu_BMI[a]	.78	.72	.61
Mahar_BMI[a]	.65	.51	.59
MaharQ_BMI[a]	.59	.48	.49
15- to 18-year-olds			
Leger	.34	.22	.51
Mahar	.27	.36	.07
Barnett	NA	NA	NA
Zhu_BMI[a]	.81	.71	.74
Mahar_BMI[a]	.65	.55	.46
MaharQ_BMI[a]	.60	.56	.40

Note. NA = not available; LBMI = lower BMI group; HBMI = higher BMI group.
[a]Include BMI in the prediction equation.
[b]Adjusted Barnett score for Kappa computation.

programs such as Fitnessgram because the goal is to provide accurate indicators of health-related fitness for both individual- and school-level reports. As previously described, the Mile run equation currently used in the Fitnessgram program includes a BMI term to take into account body weight/body fat (Cureton & Plowman, 2008). The BMI term in the prediction equation is negative so youth with a higher BMI will have a lower prediction of aerobic capacity than will youth with a lower BMI—even if they have the same Mile time or PACER performance. This possible bias is a result of the negative association between BMI (or body fatness) and VO_2peak expressed relative to body weight (in mL/kg/min; Dencker et al., 2007). Based on field test scores, the aerobic fitness of individuals with higher BMI will systematically be underestimated when BMI is not accounted for (Cureton, 1982; Cureton et al., 1978; Rowland, Kline, Goff, Martel, & Ferrone, 1999). The present study was designed to determine the contribution of BMI to equivalent estimates of aerobic capacity to advance understanding of these issues.

The results confirm that BMI does have a major influence on aerobic fitness estimates from the PACER test. The cluster analyses revealed that BMI explains 34% (11- to 14-year-olds) and 30% (15- to 18-year-olds) of the variance associated with differences among estimates from the Mile run and the PACER equations. Evaluation of the residuals for different BMI groups (Figure 2) demonstrated that PACER equations that do not include a BMI term tend to

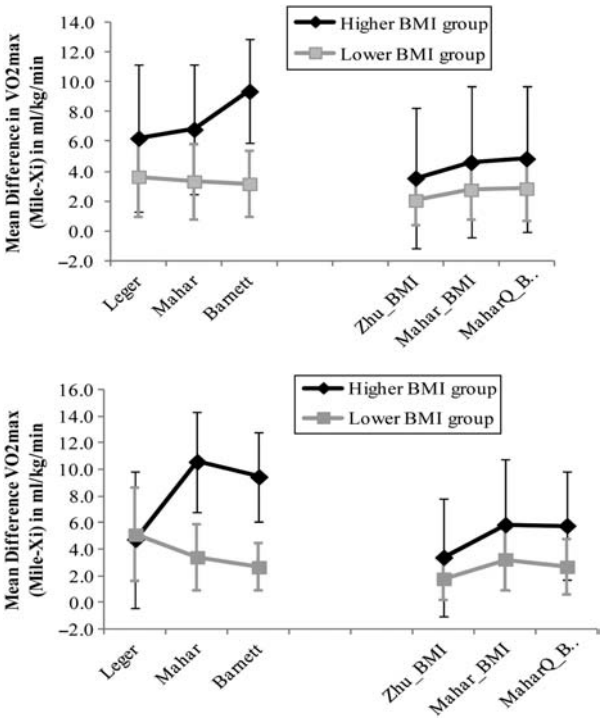

FIGURE 3 Distribution of error (Mile – Xi) by BMI subgroups (two-cluster solution) for both younger (top) and older (bottom) age groups. Equations that do not include BMI in the prediction are on the left side of the plots, while those that include BMI are on the right side of the plots.

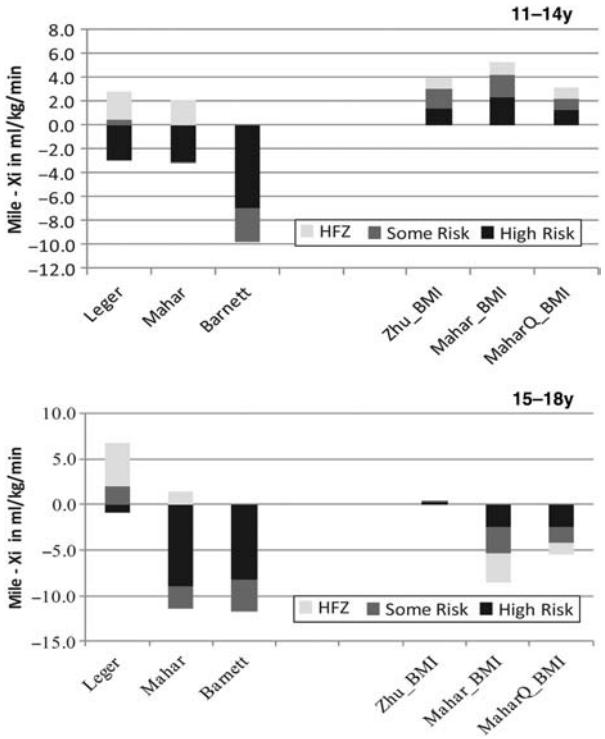

FIGURE 4 Magnitude and direction of error for PACER fitness estimates (relative to the mile run) for youth in three different BMI groups. The top figure illustrates the pattern of disagreement for the 11- to 14-year-old age group, while the bottom figure illustrates the pattern for the 15- to 18-year-old age group. Equations that do not include BMI in the prediction are on the left side of the plots, while those that include BMI are on the right side of the plots.

produce higher estimates of fitness than the Mile run. This was true for the Leger equation that has been used in Fitnessgram as well as for other alternative Equations (e.g., Mahar, Barnett). The higher estimation of fitness in these Equations (e.g., Leger equation) helps to explain why a higher percentage of youth have been found to achieve the health-related standard with the PACER than with the Mile Run Test (Welk, Laurson, et al., 2011).

The inclusion of BMI in PACER prediction equations appears to resolve this problem as error rates were considerably smaller in equations that included a BMI term. In most equations, the magnitude of error, associated with the high BMI group, was considerably higher in older children than in younger children, but this is due to fact that some of the equations were not specifically developed for use in older age groups. The Mahar equations, for example, were developed with a younger sample (11- to 14-year-olds) and were not expected to work well with the older age group. In the analyses, we segmented results by age to make it possible to examine the accuracy for these two distinct age groups. The Zhu approach yielded the lowest error in both age groups. This approach offers a number of advantages for use within the Fitnessgram program because it standardizes the way in which the PACER assessment is performed. The test-equating approach was designed to improve classification equivalency (Boiarskaia, Boscolo, Zhu, & Mahar, 2011; Zhu et al., 2010), and the present results suggest that the procedure accomplishes this goal.

The subsequent analyses confirmed that classification agreement can be substantially improved by using PACER equations that include BMI. Agreement was only fair when PACER was scored without including BMI (kappa ranged from .12 to .34), but agreement ranged from moderate to very good (kappa ranged from .59 to .81) in equations that include BMI. The Zhu approach and the Mahar BMI equations both yielded good classification agreement with the younger age group, but the Zhu method yielded good agreement for both age groups. The examination of kappa scores by BMI cluster revealed that individuals with high BMI tend to have lower classification agreement between the two tests. Agreement scores by BMI cluster were higher when using the Zhu approach. It is important to note that the kappa statistic provides a better indicator of agreement than does a simple percent agreement calculation. This is because the kappa statistic takes into account the potential of achieving agreement by chance (Hunt, 1986). We noted that the Barnett equation yielded reasonably good percent agreement with the Mile (74.2%), but the kappa values of .09 to .12 indicated poor overall agreement. This equation had consistently higher VO_2peak scores than did the Mile, and therefore, the rate of misclassification on the lowest category of fitness level seems to be high. Even though we did not use a criterion measure of VO_2peak, these results are not fully consistent with other studies that were able to examine the validity of this equation. Ruiz and colleagues (2009) used indirect calorimetry to test the Barnett equation against the Leger, Matsuzaka, and Ruiz equations. They found that the Barnett equation had the second highest standard error of estimate (and % error) but the lowest mean difference from measured VO_2peak values. The Barnett estimates correlated similarly with measured VO_2peak when compared with other equations but were still substantially higher than the Leger correlation value ($r = .73$ vs. .59). Overall, the authors concluded that this equation was shown to be more accurate than the Leger and the two other Equations (Ruiz et al., 2009).

Overall, the results of the study demonstrate the important effect that BMI has on fitness estimation and classification agreement. Absolute error rates for PACER equations that employed BMI were consistently lower than they were for equations that did not. The results suggest that agreement between the PACER and the Mile are improved when BMI is included in the PACER. However, as with any large-scale fitness assessment using field tests, there are some concerns regarding the psychometric properties of the assessments conducted. We were aware that this was critical and therefore provided standardized training to minimize error. Additionally, the cross-sectional design of this study does not allow precluding any definite conclusions. However, previous work from Cureton and colleagues (Cureton, 1982; Cureton et al., 1978) and the consistent pattern we found linking BMI scores to differences between alternative tests of aerobic fitness support our main findings. Nevertheless, the relation between aerobic capacity and body weight or measures of body composition requires further understanding and, therefore, additional research using measures of body composition and more accurate estimates of aerobic capacity.

WHAT DOES THIS ARTICLE ADD?

There are many alternative equations and algorithms available for estimating aerobic capacity in youth, and it has proven difficult to determine the most valid approach. The Fitnessgram youth fitness program recently instituted a new scoring approach for the PACER that enables estimates from the PACER to be processed with the same equation as the Mile run. Because this equation includes BMI, schools must now measure BMI to obtain estimates of aerobic capacity in Fitnessgram. This method has been shown to improve classification agreement, but it has been difficult to explain or justify to many teachers why BMI should be included in the prediction algorithm. The present study directly evaluated the impact of including or excluding BMI when processing PACER data. Algorithms that do not include BMI (or any other measure of body size) had larger discrepancies with the Mile run than did algorithms that did

take body size into account. Errors were larger for individuals with high BMI. The study also confirmed that the new test-equating approach used in the Fitnessgram program had better agreement with Mile run compared with alternative PACER equations that included BMI. Overall, the results support the importance of including BMI to ensure good agreement with estimates from the Mile run. Both tests can provide reasonable estimates of aerobic capacity; however, the relation between body size and aerobic fitness should be further examined using more accurate measures. Maximal VO_2 obtained through indirect calorimetry may overcome this limitation.

ACKNOWLEDGMENTS

The authors would like to thank Marilu Meredith, Mathew Mahar, and Kirk Cureton for their valuable insights during the revision of this manuscript.

FUNDING

Supported by a Cooper Institute/FITNESSGRAM grant.

REFERENCES

Aldenderfer, M. S., & Blashfield, R. K. (1984). A review of clustering methods. In Sage University Paper Series on Quantitative Applications in the Social Sciences (Ed.), *Cluster analysis* (No. 07-001, pp. 33–61). Beverly Hills, CA: Sage.

Altman, D. G. (1990). *Practical statistics for medical research*. London, England: Chapman and Hall/CRC.

Barnett, A., Chan, L. Y. S., & Bruce, I. C. (1993). A preliminary study of the 20-m multistage shuttle run as a predictor of peak VO_2 in Hong Kong Chinese students. *Pediatric Exercise Science*, *5*(1), 42–50.

Boiarskaia, E. A., Boscolo, M. S., Zhu, W., & Mahar, M. T. (2011). Cross-validation of an equating method linking aerobic FITNESSGRAM field tests. *American Journal of Preventive Medicine*, *41*(Suppl. 2), S124–S130.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

Cureton, K. J. (1982). Distance running performance tests in children: What do they mean? *Journal of Physical Education, Recreation and Dance*, *53*(8), 64–66.

Cureton, K., & Plowman, S. (2008). Aerobic capacity assessments. In G. J. Welk & M. D. Meredith (Eds.), *Fitnessgram/Activitygram reference guide* (3rd ed., pp. 9.1–9.25). Dallas, TX: Cooper Institute.

Cureton, K. J., Sloniger, M. A., O'Bannon, J. P., Black, D. N., & McCormack, W. P. (1995). A generalized equation for prediction of VO_2 peak from one-mile run/walk performance in youth. *Medicine and Science in Sports and Exercise*, *27*, 445–451.

Cureton, K. J., Sparling, P. B., Evans, B. W., Johnson, S. M., Kong, U. D., & Purvis, J. W. (1978). Effect of experimental alterations in excess weight on aerobic capacity and distance running performance. *Medicine and Science in Sports*, *10*, 194–199.

Dencker, M., Thorsson, O., Karlsson, M. K., Linden, C., Eiberg, S., Wollmer, P., & Andersen, L. B. (2007). Gender differences and determinants of aerobic fitness in children aged 8–11 years. *European Journal of Applied Physiology*, *99*, 19–26.

Hunt, R. J. (1986). Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. *Journal of Dental Research*, *65*, 128–130.

Leger, L. A., Mercier, D., Gadoury, C., & Lambert, J. (1988). The multistage 20 metre shuttle run test for aerobic fitness. *Journal of Sports Science*, *6*, 93–101.

Mahar, M. T., Guerieri, A. M., Hanna, M. S., & Kemble, D. (2011). Estimation of aerobic fitness from 20-m multistage shuttle run test performance. *American Journal of Preventive Medicine*, *41*(Suppl. 2), S117–S123.

Mahar, M. T., Welk, G. J., Rowe, D. A., Crotts, D. J., & Mclver, K. J. (2006). Development and validation of a regression model to estimate VO_2 peak from PACER 20-m shuttle run performance. *Journal of Physical Activity and Health*, *3*(2, Suppl.), S34–S46.

Rowland, T., Kline, G., Goff, D., Martel, L., & Ferrone, L. (1999). One-mile run performance and cardiovascular fitness in children. *Archives of Pediatrics & Adolescent Medicine*, *153*, 845–849.

Ruiz, J. R., Silva, G., Oliveira, N., Ribeiro, J. C., Oliveira, J. F., & Mota, J. (2009). Criterion-related validity of the 20-m shuttle run test in youths aged 13–19 years. *Journal of Sports Science*, *27*, 899–906.

Welk, G. J., Going, S. B., Morrow, J. R., Jr. & Meredith, M. D. (2011). Development of new criterion-referenced fitness standards in the FITNESSGRAM program: Rationale and conceptual overview. *American Journal of Preventive Medicine*, *41*(Suppl. 2), S63–S67.

Welk, G., Laurson, K., Eisenmann, J., & Cureton, K. (2011). Development of youth aerobic-capacity standards using receiver operating characteristic curves. *American Journal of Preventive Medicine*, *41*(Suppl. 2), S111–S116.

Welk, G. J., Meredith, M. D., Lhmels, M., & Seeger, C. (2010). Distribution of health-related physical fitness in Texas youth: A demographic and geographic analysis. *Research Quarterly for Exercise and Sport*, *81*(Suppl. 3), S6–S15.

Welk, G. J., Saint-Maurice, P. F., Laurson, K. R., & Brown, D. D. (2011). Field evaluation of the new FITNESSGRAM® criterion-referenced standards. *American Journal of Preventive Medicine*, *41*(Suppl. 2), S131–S142.

Zhu, W., Plowman, S. A., & Park, Y. (2010). A primer-test centered equating method for setting cut-off scores. *Research Quarterly for Exercise and Sport*, *81*, 400–409.