

Criterion-Referenced Reliability and Equivalency Between the PACER and 1-Mile Run/Walk for High School Students

Michael W. Beets and Kenneth H. Pitetti

Background: To examine the Healthy Fitness Zone (pass/fail) criterion-referenced reliability (CRR) and equivalency (CRE) of the 1-mile run/walk (MRW) and Progressive Aerobic Cardiovascular Endurance Run (PACER) in adolescents (13 to 18 years). *Methods:* Seventy-six girls and 165 boys were randomly assigned to complete 2 trials of each test. *Results:* CRR for the boys on the MRW ($Pa = 77\%$, $\kappa_q = 0.53$) was lower than on the PACER ($Pa = 81\%$, $\kappa_q = 0.63$); girls were classified more similarly on the MRW ($Pa = 83\%$, $\kappa_q = 0.67$) than on the PACER ($Pa = 79\%$, $\kappa_q = 0.58$). The CRE between the MRW and PACER indicated boys ($Pa = 77\%$, $\kappa_q = 0.55$) were classified more consistently on both tests than girls ($Pa = 73\%$, $\kappa_q = 0.46$). *Conclusions:* No test provided greater consistency. Practitioners may consider other features, such as ease of administration, environmental conditions, and comparative use in the literature.

Key Words: cardiovascular fitness, adolescents, *FITNESSGRAM*, field test

School-based physical fitness testing of children and adolescents is becoming increasingly recognized as an important component to the public health surveillance of the levels of fitness of our nation's youth. As of the year 2000, approximately 76.5% of schools nationwide assessed the physical fitness of their students.¹ Increased levels of fitness, specifically cardiovascular fitness, are associated with a lower risk of the metabolic syndrome, characterized by hyperinsulinemia, low glucose tolerance, hyperlipidemia, and overweight in youth.^{2,3} The etiology of cardiovascular disease begins during childhood and tracks into adulthood,⁴ with several longitudinal studies associating high levels of physical fitness during adolescence with lower risk profiles for developing cardiovascular disease later in life.^{5,6,7} The results of the aforementioned studies signify that cardiovascular fitness should be monitored on a continuous basis throughout childhood and adolescence.

The ability to track cardiovascular fitness (VO_{2peak} , $ml \cdot kg^{-1} \cdot min^{-1}$), and therefore, cardiovascular health status of large populations of school age youth using standard laboratory protocols (i.e., treadmill, metabolic cart, staff expertise) is not

Beets is with the Dept. of Public Health, Oregon State University, 256 Waldo Hall, Corvallis, OR 97331-6406. Pitetti is with the Dept. of Physical Therapy, Wichita State University, Wichita, KS 67208.

feasible. Thus, alternative valid and reliable field measures have been developed to indirectly assess cardiovascular health status. The 1-mile run/walk (MRW) and the Progressive Aerobic Cardiovascular Endurance Run (PACER or 20-meter shuttle run) have been established as the most valid and reliable field test when assessing the cardiovascular fitness of school age youth.^{8–12} Based on student achievement on the MRW (in seconds) and PACER (in laps completed), criterion-referenced performance standards have been developed to provide individual, diagnostic information regarding the attainment of performance directly associated with good health.¹³ Specifically, criterion-referenced standards represent minimal levels of cardiovascular fitness that offer protection against disease¹⁴ and are used in national fitness test batteries (e.g., *FITNESSGRAM*[®]) to identify youth who are either *in need of improvement* or are within a zone of optimal health (i.e., *healthy fitness zone*) for cardiovascular fitness.

The *FITNESSGRAM* includes both the MRW and PACER, leaving the decision on which to administer up to the teacher. And while the MRW and the PACER measure the same outcome (i.e., cardiovascular fitness), a question remains as to which assessment provides the most accurate and consistent results of current cardiovascular fitness across gender and age. That is, due to age, gender, or additional influences, certain populations of students may do better on one test when compared to the other.

Four studies have addressed one or both of these issues with children,¹⁵ early adolescents,^{16, 17} and college age students.¹⁸ Rikli et al.¹⁵ examined the criterion-referenced reliability—the consistency of being classified the same (i.e., *in need of improvement* or in the *healthy fitness zone*) on two trials of the same test—of the MRW in kindergarten to 4th grade students. Results indicated that anywhere from as low as 45% up to 94% of the students were correctly classified after completing two trials. Examinations of criterion-referenced reliability of the PACER have indicated high levels of classification consistency for 10- to 11-year-old boys (82 and 83%) and girls (97 and 94%),^{16, 17} with similar results for 18- to 30-year-olds (95%).¹⁸ The criterion-referenced equivalency, the consistency of being classified the same on two parallel tests (i.e., MRW and PACER), has been found to be adequate with boys (83%) yet unacceptable for girls (65%),¹⁷ with one study indicating moderate levels (79 and 76%) of consistency.¹⁶ For college age individuals,¹⁸ criterion-referenced equivalency indicates a high level of consistency in classification between the two tests (95 and 86%).

Unfortunately, no such investigations have been conducted in the high school age population. Thus, any implication as to the criterion-referenced reliability and equivalency for these two tests in this age group is not possible. Therefore the purpose of this study was twofold: (1) to examine the criterion-referenced reliability for the MRW and PACER; and (2) to examine the criterion-referenced equivalency between the MRW and PACER in a sample of high school age students.

Methods

Participants

All students ($N = 241$, 96% White non-Hispanic) enrolled in physical education classes at one local high school in the Midwest were administered two trials on both the MRW and PACER as part of their regular physical education curriculum on the

topic of cardiovascular health. The testing took place over the course of 3 weeks during September 2004. The number of students (boys and girls) reported for each test varies slightly due to absence or sport related injury on any given assessment day. The small sample of girls was due to the limited number of upper-level students (10th to 12th graders) voluntarily enrolled in PE. Nonetheless, the majority of the sample were freshmen (50%, $n = 38$), suggesting they were representative of the girls in this high school. Approval for the study was obtained by the university's institutional review board, the high school administrators and teachers, and the school district. Because the tests were part of the regular physical education curriculum and were conducted on a routine basis, no informed consent was required from the students.

Procedure

The sample consisted of 165 boys and 76 girls 13 to 18 years of age. Height (to the nearest 0.2 cm) and weight (to the nearest 1.0 lb), without shoes, were measured for each student during physical education class using a portable stadiometer (Siber Hegner & Co. anthropometric kit, Baltimore, MD) and physician's scale (Detecto balance beam scale, Daugherty Webb City, MO). Units were converted and body mass index (BMI) was calculated using the formula: weight (kg) divided by height (m) square. Participants' age (in decimal) was determined by subtracting date of birth from date of assessment. The descriptive measures of the sample can be found in Table 1.

Table 1 Descriptive Characteristics for Boys and Girls Separately

Sex	Variable	Mean	$\pm SD$
Boys	Age (years)	16.0	1.29
	Height (cm)	174.8	7.8
	Weight (kg)	73.2	16.7
	BMI	24.4	5.1
Girls	Age (years)	15.7	1.2
	Height (cm)	163.1	7.2
	Weight (kg)	61.0	11.2
	BMI	23.3	3.6

The physical education classes were scheduled in blocks with weekly attendance alternating on a 2- and 3-days-per-week schedule. Each day consisted of four separate classes for a total of eight different classes across 2 days. Test administration was randomized, with some classes during the same day running the MRW and the other classes running the PACER. Random administration of the tests was performed until each class completed two trials on both the MRW and PACER, with a minimum of 2 days elapsing between tests.

1-Mile Run/Walk (MRW). The MRW was administered on an outdoor 400-meter athletic track (4 laps = 1 mile) according to the *FITNESSGRAM* testing procedures.¹⁴ Students were run in groups of no more than 15. One research staff member timed the entire group using a digital hand-held stopwatch (Model 226, Sportline; Yonkers, NY) and read the elapsed time for each completed lap. The final fourth lap time (in minutes and seconds) was recorded by the physical education teacher for each student. Time was converted to seconds for the final analysis. The temperature during the outdoor testing averaged 70.2 °F during the morning classes (8 to 11 a.m.) and 78.7 °F during the afternoon classes (1 to 3 p.m.) (source: <http://www.wunderground.com>).

Progressive Aerobic Cardiovascular Endurance Run (PACER). The distance for the PACER was marked by painted stripes (baseline of indoor basketball court) with orange cones set at each end. Participants were instructed to run the distance between cones in the allotted time. The PACER protocol outlined by the *FITNESSGRAM* Test User's Manual¹⁴ was used for all tests. One lap represented jogging or running from one set of cones to the other. The test was terminated either due to volitional exhaustion or because the participant could not keep up the required speed for two laps. The number of laps completed was recorded for data analysis by the physical education teacher.

Data Analysis

Criterion-Referenced Reliability and Equivalency. Students were classified as being in either one of two conditions based on their MRW and PACER performance for each trial: (a) in need of improvement, or (b) in the healthy fitness zone. The criterion-referenced standards were based on age- and sex-specific performance standards outlined in the *FITNESSGRAM* Test User's Manual.¹⁴ Criterion-referenced reliability was estimated using the proportion of agreement (Pa) and modified kappa (κ_q , correcting for chance agreement) statistics. For criterion-referenced reliability, the Pa is defined as the proportion of students classified the same on both trials of the PACER or MRW. For criterion-referenced equivalency, Pa is defined as the proportion of students classified the same on both the MRW and PACER. To provide the most representative score, students' average MRW and PACER performances for the two trials were used to calculate the criterion-referenced equivalency between the two tests.

Means and standard deviations for all measures were calculated for boys and girls separately. Single and average measure intraclass correlation coefficients (ICC) and 95% confidence intervals (95 CI) were calculated to determine the reliability of the PACER and MRW for the entire sample, and for boys and girls separately. Bland-Altman¹⁹ plots were constructed to illustrate the difference in students' performance between Trial 1 and Trial 2 for both the MRW and PACER tests. For the PACER and MRW, Trial 1 was subtracted from Trial 2 so that positive numbers indicated better (i.e., greater number of laps completed) performance on the PACER and worse (i.e., greater time to complete the 1-mile distance) performance on the MRW. Bivariate correlations were calculated to assess the relation between the students' average MRW and PACER performance for the entire sample and for boys and girls separately. All analyses were conducted using Statistical Package for the Social Sciences (v 12.0).

Table 2 Performance on Both 1-Mile Run/Walk and Progressive Aerobic Cardiovascular Endurance Run Trials and Average MRW and PACER Trials

Test	<i>n</i>	Boys		<i>n</i>	Girls	
		Mean	$\pm SD$		Mean	$\pm SD$
MRW-1 (sec)	135	527.1	192.7	70	620.8	140.1
MRW-2 (sec)	125	617.3	245.5	70	646.0	166.0
PACER-1 (laps)	140	47.9	25.8	71	33.5	17.4
PACER-2 (laps)	140	46.7	32.2	65	31.7	18.9
Average MRW time (sec)	114	561.3	97.3	66	638.1	144.2
Average PACER (laps)	123	46.1	25.4	62	32.8	16.4

Table 3 Comparison Among Estimates of Criterion-Referenced Reliability and Equivalency for the PACER and MRW

Criterion	Test	Statistic	Total sample		
			Boys	Girls	
Reliability	PACER	Proportion of agreement	0.81	0.81	0.79
		Modified kappa	0.61	0.63	0.58
	MRW	Proportion of agreement	0.79	0.77	0.83
		Modified kappa	0.58	0.53	0.67
Equivalency	PACER-MRW ^a	Proportion of agreement	0.76	0.77	0.73
		Modified kappa	0.52	0.55	0.46

^a Average PACER and MRW times

Results

The results of the MRW and PACER can be found in Table 2. Not unexpectedly, boys outperformed girls on each test. The criterion-referenced reliability P_a and κ_q for the MRW and PACER can be found in Table 3. Results indicated that 81% of the boys and 79% of the girls were classified the same on both trials of the PACER, and 77% of the boys and 83% of the girls were classified the same on both trials of the MRW. The κ_q indicated low to moderate classification agreement for both the MRW and PACER. For boys, κ_q was greater for the PACER ($\kappa_q = 0.63$) than for the MRW ($\kappa_q = 0.53$), while the classification agreement was greater for the MRW ($\kappa_q = 0.67$) compared to the PACER ($\kappa_q = 0.58$) for the girls. The criterion-referenced equivalency resulted in 77% of the boys and 73% of the girls being classified the same on their average MRW and PACER performance (see Table 3). The modified kappa indicated low results for the tests for both boys ($\kappa_q = 0.55$) and girls ($\kappa_q = 0.46$).

Table 4 Reliability (intraclass correlation coefficients, 95% CI) Estimates for the PACER and MRW

Sex	<i>n</i>	Test	Statistic	ICC	(95 CI)
Boys	123	PACER	Single measure	0.68	(0.57, 0.76)
			Two-trial average	0.81	(0.72, 0.86)
	114	MRW	Single measure	0.66	(0.54, 0.75)
			Two-trial average	0.80	(0.70, 0.86)
Girls	62	PACER	Single measure	0.64	(0.46, 0.77)
			Two-trial average	0.78	(0.63, 0.87)
	66	MRW	Single measure	0.77	(0.64, 0.85)
			Two-trial average	0.87	(0.78, 0.92)

Reliability estimates revealed low single measure ICC for both tests (see Table 4). Similar reliability estimates were obtained on the PACER (ICC = 0.68) and MRW (ICC = 0.66) for boys, while greater reliability was observed on the MRW (ICC = 0.77) in comparison to the PACER (ICC = 0.64) for girls. The results of the Bland-Altman plots for the MRW and PACER are presented in Figures 1 and 2. The number of laps completed from Trial 1 to Trial 2 on the PACER shows a slight trend in increasing performance, specifically for the girls. The reverse can be seen for the boys, with many reducing the amount of laps they completed from Trial 1 to Trial 2, some of these being considerable. Greater clustering of boys and girls can be seen for the MRW, with the majority of the students either maintaining their time or performing slightly better (i.e., a decrease in the amount of time to complete the 1-mile distance) on Trial 2. Bivariate correlations indicated moderate association between the MRW and PACER for the total group ($r = -0.650$) and for boys ($r = -0.633$), whereas a slightly higher correlation was observed for the girls ($r = -0.700$).

Discussion

The use of the 1-mile run/walk (MRW) and Progressive Aerobic Cardiovascular Endurance Run (PACER) for evaluating the cardiovascular fitness of youth is widespread, with over two-thirds of the schools in the U.S. requiring physical fitness testing as an evaluative component to physical education curriculum. These tests have also been instrumental in determining cross-sectional differences^{20, 21} as well as examining secular trends in cardiovascular fitness.^{22, 23} Given the importance of cardiovascular fitness in childhood and adolescence and its link to cardiovascular disease risk during adulthood,^{5, 6, 7} the ability to indirectly appraise a student's level of cardiovascular fitness allows for the identification of and the ability to take preventive action toward those youth who may be at risk for low levels of cardiovascular fitness later in life. However, prior to using such diagnostic criteria (i.e., minimal level of health that offers protection against disease), student performance on multiple trials of the same test or on parallel tests should provide the same feedback. Investigations of this sort have yet to be conducted with adolescents. Therefore,

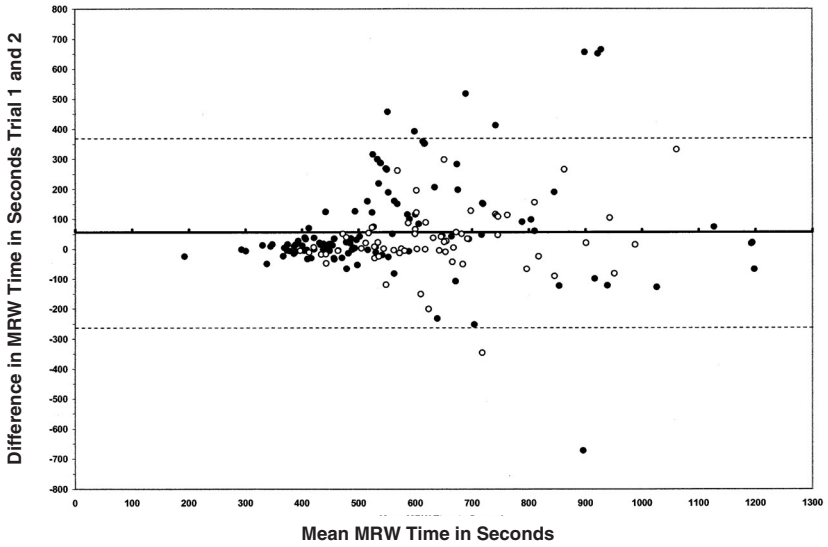


Figure 1 — Bland-Altman plot of the mean difference in time between Trials 1 and 2 on the MRW for boys (black circles) and girls (white circles). Dashed lines indicate the 95% confidence interval.

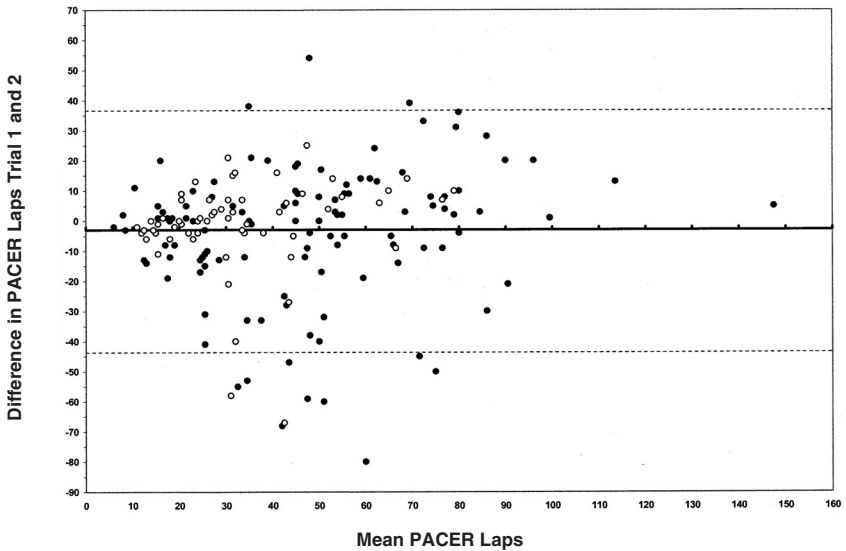


Figure 2 — Bland-Altman plot of the mean difference in laps between Trials 1 and 2 on the PACER for boys (black circles) and girls (white circles). Dashed lines indicate the 95% confidence interval.

the purpose of this study was to examine the criterion-referenced reliability and equivalency between the MRW and PACER tests of cardiovascular fitness in a sample of high school students.

There were minor differences between the criterion-referenced reliability estimates between boys and girls. Boys were classified more consistently on the PACER (81%) as opposed to the MRW (77%). The reverse was observed for girls, with greater classification consistency observed on the MRW (83%) in comparison to the PACER (77%). The corresponding modified kappa values, ranging from 0.53 to 0.67, suggest moderate levels of consistency in classification for both tests. Thus, no one test proved to have greater criterion-referenced reliability over the other.

The results for the PACER are mixed in comparison to the findings from previous studies (see Table 5). Mahar et al.¹⁷ and Dinschel¹⁶ reported the *Pa* for two trials on the PACER for boys (ages 10 to 11 years) to be 82 and 83%, respectively, which is similar to the 81% for the boys in the current study. Conversely, the girls in those studies achieved a considerably higher classification agreement for the two PACER trials, with values of 97 and 94%, respectively. This is in contrast to the 79% found in this sample of girls. Several possibilities may account for this discrepancy. First, the passing criteria (i.e., minimal number of laps required to achieve the *healthy fitness zone* criteria) for the PACER was based on 1992 FITNESSGRAM Healthy Fitness Zone standards and have since been changed (from a minimum of 7 and 9 laps to 15 laps for 10- and 11-year-olds). This was noted as a reason for the unusually high classification rate for the girls, particularly in the Mahar et al. sample.

The current criterion standards set the minimal laps at 23 for 12- to 15-year-olds, at 32 for 16-year-olds, and at 41 for 17-year-olds and above for girls. Therefore, with changes in criterion scores, direct comparisons between studies may not be appropriate. Second, one cannot rule out the possibility that a decrease in performance occurred between the administrations of the two PACER tests. A decrease in performance across administrations would cause girls who were classified in the *healthy fitness zone* during the first administration to be classified as *in need of improvement* for the second trial, thereby causing the *Pa* to be artificially low. This is supported by the lower than expected single measure ICC for the PACER (ICC = 0.64). On the other hand, girls may have performed poorly on the first trial and improved on the second, also causing the low ICC and leading to the low *Pa*.

While both scenarios are plausible, the latter is supported from the results of the Bland-Altman plot (see Figure 2). The plot shows that a majority of the girls improved their performance from Trial 1 to Trial 2. This learning effect may have originated from the fact that the students, while having run the PACER during previous physical education classes, had not run the test since the spring and thus the summer vacation may have caused some of the students to become less familiar with the protocol, thereby leading to a poorer performance on the first trial as opposed to the second. Low reliability was also observed for the boys (ICC = 0.68) on the PACER, yet the classification for the two trials was consistent with previous findings. This low ICC can be explained by the Bland-Altman plot, which shows the boys having greater variance in their performance between trials.

Research into the criterion-referenced reliability of the MRW has received little attention. Rikli et al.¹⁵ examined the criterion-referenced reliability of the MRW in kindergarten through 4th grade. Estimates of classification consistency for two trials ranged from 75 to 91% and 69 to 94% during the fall for boys and girls,

Table 5 Comparison Among Estimates of Criterion-Referenced Reliability and Equivalency for PACER and MRW

Criterion	Test	Study	Age group	Statistic	Tot sample	Boys	Girls
Reliability	PACER	Mahar et al.	10–11 yrs	Proportion of agreement	0.89	0.82	0.97
		Dinschel et al.	10–11 yrs	Modified kappa	0.78	0.65	0.94
	This study	Plowman et al. ^a	13–18 yrs	Proportion of agreement	0.88	0.83	0.94
			18–30 yrs	Modified kappa	0.88	0.66	0.76
		Rikli et al. ^b	5 yrs	Proportion of agreement	0.81	0.81	0.79
			6 yrs	Modified kappa	0.61	0.63	0.58
	MRW	Rikli et al. ^b	7 yrs	Proportion of agreement	0.95	–	–
			8 yrs	Modified kappa	0.90	–	–
		This study	13–18 yrs	Proportion of agreement	–	0.75 [0.70]	0.69 [0.51]
			18–30 yrs	Modified kappa	–	0.76 [0.66]	0.71 [0.45]
Equivalency	PACER	Mahar et al. ^c	10–11 yrs	Proportion of agreement	–	0.85 [0.77]	0.81 [0.85]
		Dinschel et al.	10–11 yrs	Proportion of agreement	–	0.91 [0.85]	0.90 [0.84]
	This study	Plowman et al. ^a	13–18 yrs	Proportion of agreement	–	0.86 [0.83]	0.83 [0.94]
			18–30 yrs	Modified kappa	0.79	0.77	0.83
		Mahar et al. ^c	10–11 yrs	Proportion of agreement	0.58	0.53	0.67
			13–18 yrs	Modified kappa	0.76	0.83	0.66
	& MRW	Dinschel et al.	10–11 yrs	Proportion of agreement	0.51	0.65	0.33
			13–18 yrs	Modified kappa	0.71	0.76	0.65
		This study ^d	10–11 yrs	Proportion of agreement	0.43	0.52	0.30
			13–18 yrs	Modified kappa	0.78	0.79	0.76
Plowman et al.	18–30 yrs	Proportion of agreement	0.56	0.58	0.52		
	18–30 yrs	Modified kappa	0.76	0.77	0.73		
Plowman et al.	18–30 yrs	Proportion of agreement	0.52	0.55	0.46		
	18–30 yrs	Modified kappa	0.90	0.95	0.86		
					0.80	0.90	0.72

^a Proportion of agreement and modified kappa not reported separately for sexes; ^b Proportion of agree. estimated for two times: fall [spring]; ^c Proportion of agree. between PACER Trial 1 & MRW, and PACER Trial 2 & MRW; ^d avg performance from two trials for both PACER and MRW.

respectively, and from 66 to 85% and 45 to 94% during the spring, respectively (see Table 5). The findings from this study are similar, with 77 and 83% of the boys and girls classified the same on the two trials, suggesting that these values may be typical of classification consistency on the MRW in these age groups. Nonetheless, despite the similarities between studies, 23 and 17% of the boys and girls in the current sample were not classified similarly across trials.

Possible reasons for this discrepancy are similar to those mentioned previously for the PACER tests (see above). In examining the Bland-Altman plot for the MRW trials, a number of students decreased their performance from Trial 1 to Trial 2, with some decreases approaching 30 seconds—the time increments the criterion-referenced scores decrease/increase for each age and sex group. Thus, students meeting their age/sex criterion score, yet by only 30 seconds or less on Trial 1, would be classified as in need of improvement on Trial 2 if their time was 30 seconds slower. The single measure ICCs also support this supposition, with boys exhibiting low reliability on the MRW (ICC = 0.66), while girls performed with slightly greater consistency (ICC = 0.77) (see Table 4).

Although the students were familiar with both the PACER and MRW protocol from previous testing occasions, our findings underscore several important considerations when utilizing these tests with adolescents. The allocation of sufficient practice to become refamiliarized with the test may be necessary to ensure that students perform at peak levels. This can be seen in the systematic improvement (see Figures 1 and 2) in time and lap performance from Trial 1 to Trial 2. Better performance on Trial 2 also suggests that, when given an additional chance to perform the test, students choose to improve on their initial performance. It may be necessary, therefore, to perform at minimum three trials of the tests, excluding the initial trial, and using the second and third to calculate reliability coefficients. However, not all students increased their performance from Trial 1 to Trial 2. A number of students decreased their performance, which may indicate one of two possibilities: (a) once a “maximum” effort is provided, motivation to perform well on a subsequent trial is diminished; or (b) a minimum of 48 hours between trials is not enough to ensure adequate recovery. These scenarios are likely to have affected the results of the current study and therefore should be viewed as limitations when interpreting our findings.

The *FITNESSGRAM* offers two choices of cardiovascular tests for physical educators to evaluate their students’ cardiovascular fitness. Although there are fundamental differences in the administration of the tests, with the MRW primarily conducted outdoors on an athletic track and being self-paced while the PACER is conducted indoors, covers a distance of 20 meters, and is externally paced using an audio signal, the two tests are designed to measure the same construct: cardiovascular fitness. Thus, despite the apparent differences in the tests, if both are designed to measure the same construct, then performance feedback on the two tests should be similar. That is, a student meeting the *healthy fitness zone* criteria on one test should also be within the *health fitness zone* on the other. This is referred to as criterion-referenced equivalency and has been explored in three prior studies (see Table 5).

Both Mahar et al.¹⁷ and Dinschel¹⁶ observed higher equivalency rates (proportion of agreement) between the MRW and PACER for boys than for girls. As mentioned previously, Mahar et al.¹⁷ cited the low minimal criterion scores for PACER laps for girls which were most likely not equivalent to the criterion stan-

dards for the MRW. This caused most of the girls to achieve the *healthy fitness zone* standard for the PACER, yet fail to achieve it for the MRW, thereby contributing to the low equivalency rates observed. In the current study, boys were classified with greater consistency than girls for the two tests, albeit only minimally (77 vs. 73%). Our results are close to Dinschel's with equivalency rates of 79 and 76% for boys and girls, respectively.

In their investigation of test equivalency in college age individuals, Plowman and Liu¹⁸ reported equivalency rates of 95 and 86% for males and females, respectively. However, it should be noted that the Plowman and Liu sample was composed of volunteers, and therefore preferential self-selection of individuals who are already cardiovascularly fit cannot be ruled out and possibly contributes to the high rates of classification consistency on the two tests. The current sample was not self-selected, however, but rather was composed of all students enrolled in physical education class within one high school. As a result, the findings from this sample are deemed to be representative of the levels of cardiovascular fitness of the high school students in this school.

Our results suggest that both the MRW and PACER are providing similar misclassification rates for both boys and girls, and neither test emerged as a more reliable indicator of criterion score achievement. Further, the criterion-referenced equivalency suggests that inconsistent feedback on a student's cardiovascular fitness level may occur depending on which test is administered. Cureton and Warren¹³ indicated that providing misclassification status (i.e., false master, false non-master) on fitness test may not be severe. They suggest that false non-mastery status of a student may result in increased physical activity, whereby the student is motivated to achieve the needed fitness levels to meet the criterion requirements.

Nevertheless, negative consequences may also arise. Whitehead and Corbin²⁴ examined the effect of false feedback on perceived competence during a test of motor performance (i.e., Illinois Agility Run) in 7th and 8th grade students. The experimental procedures entailed providing false test performance information to the students in order to determine the effect on intrinsic motivation to perform well on a subsequent trial. Those students who received false information indicating poor performance reported lower intrinsic motivation, while students receiving false information indicating good performance increased their intrinsic motivation. These findings suggest that providing false information or false classification (i.e., false-negative), specifically feedback that indicates worse performance, can detract from a student's interest to achieve a health-enhancing level of physical fitness.

It appears that some youths perform at a higher level on one test in comparison to another. Possible contributing factors may stem from social and physical environment influences. Peer dynamics in the form of encouragement may account for increased performance, while it is likely that decreased performance may result from comparing one's performance with that of others (ego orientation). For instance, students who perform well on the MRW finish the test first, thereby creating a situation where the less physically fit students are subjected to peer scrutiny of their performance. Conversely, those who perform well on the PACER are the last to finish the test. On the PACER all the students ran in unison, whereas on the MRW the students ran separately (self-paced). Environmental variables may also account for performance differences between the tests. The MRW is generally performed outdoors. This introduces uncontrollable factors, such as temperature and wind, which have a known effect on run performance. On the other hand, the PACER

is performed indoors under more controlled conditions, essentially removing the possibility of external physical conditions influencing test performance.

In conclusion, both the MRW and PACER tests exhibited similar criterion-referenced reliability and equivalency. No one test, therefore, can be considered as providing more stable diagnostic feedback to the students over the other. These findings preclude promoting the use of a specific test based on criterion-referenced reliability and equivalency. Given this, we suggest that practitioners choose a cardiovascular fitness test based on other criteria, such as the ability to test outdoors versus indoors, ease of administration, and comparative use in the literature.

References

1. Centers for Disease Control and Prevention. *School Health Policies and Programs Study (SHPPS) 2000 Fact Sheet: Physical Education and Activity*. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion: Division of Adolescent and School Health; 2000.
2. Brage S, Wedderkopp N, Ekelund U, et al. Features of the metabolic syndrome are associated with objectively measured physical activity and fitness in Danish children: The European Youth Heart Study (EYHS). *Diabetes Care*. 2004; 27:2141-2148.
3. Rodriguez-Moran M, Salazar-Vazquez B, Violante R, Guerrero-Romero F. Metabolic syndrome among children and adolescents aged 10–18 years. *Diabetes Care*. 2004; 27:2516-2517.
4. Nicklas TA, von Duvillard SP, Berenson GS. Tracking of serum lipids and lipoproteins from childhood to dyslipidemia in adults: The Bogalusa Heart Study. *Int J Sports Med*. 2002; 23:S39-S43.
5. Boreham CAG, Twisk J, Neville C, Savage MJ, Murray L, Gallagher A. Associations between physical fitness and activity patterns during adolescents and cardiovascular risk factors in young adulthood: The Northern Ireland Young Hearts Project. *Int J Sports Med*. 2002; 23:S22-S26.
6. Janz KF, Dawson JD, Mahoney LT. Increases in physical fitness during childhood improve cardiovascular health during adolescence: The Muscatine Study. *Int J Sports Med*. 2002; 23:S15-S21.
7. Twisk JWR, Kemper HCG, van Mechlen W. The relationship between physical fitness and physical activity during adolescence and cardiovascular disease risk factors at adult age: The Amsterdam Growth and Health Longitudinal Study. *Int J Sports Med*. 2002; 23:S8-S14.
8. Boreham CAG, Paliczka VJ, Nichols AK. A comparison of the PWC₁₇₀ and 20-MST test of aerobic fitness in adolescent schoolchildren. *J Sports Med Phys Fitness*. 1990; 30(1):19-23.
9. Cureton KJ, Baumgartner TA, McManis BG. Adjustment of 1-mile run/walk test scores for skinfold thickness. *Ped Exerc Sci*. 1991; 3:152-167.
10. Cureton KJ, Sloniger MA, O'Bannon JP, Black DM, McCormack WP. A generalized equation for the prediction of VO_{2peak} from 1-mile run/walk performance. *Med Sci Sports Exerc*. 1995; 27:445-451.
11. Leger LA, Mercier D, Gadoury C, Lambert J. The multistage 20 metre shuttle run test for aerobic fitness. *J Sports Sci*. 1988; 6:93-101.
12. Liu NYS, Plowman SA, Looney MA. The reliability and validity of the 20-meter shuttle test in American students 12 to 15 years old. *Res Q Exerc Sport*. 1992; 63(4): 360-365.
13. Cureton KJ, Warren GL. Criterion-referenced standards for youth health-related fitness tests: A tutorial. *Res Q Exerc Sport*. 1990; 61(6):7-19.

14. Welk GJ, Morrow JRJ, Falls HB. (Eds.) *Fitnessgram Reference Guide*. Dallas, TX: The Cooper Institute; 2002.
15. Rikli RE, Petray C, Baumgartner TA. The reliability of distance run tests for children in grades K–4. *Res Q Exerc Sport* 1992; 63:270-277.
16. Dinschel KM. *The influence of agility on the mile run and PACER tests of aerobic endurance in fourth and fifth grade school children* [M.S. in Education Thesis]. Dekalb, IL: Northern Illinois University; 1994.
17. Mahar MT, Rowe DA, Parker CR, Mahar FJ, Dawson M, Holt JE. Criterion-referenced and norm-referenced agreement between the Mile Run/Walk and PACER. *Measurement in Phys Educ Exerc Sci*. 1997; 1(4):245-258.
18. Plowman S.A., Liu NY-S. Norm-referenced and criterion-referenced validity of the one-mile run and PACER in college age individuals. *Measurement in Phys Educ Exerc Sci*. 1999; 3(2):63-84.
19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measure. *The Lancet*. 1986; 1:301-310.
20. Beets MW, Pitetti KH. 1-Mile run/walk and body mass index of an ethnically diverse sample of youth. *Med Sci Sports Exerc*. 2004; 10:1796-1803.
21. Beets MW, Pitetti KH. A comparison of 20 meter shuttle run performance of Mid-western youth to national and international counterparts. *Ped Exerc Sci*. 2004; 16: 94-112.
22. Tomkinson GR, Leger LA, Olds TS, Carzorra G. Secular trends in the performance of children and adolescents (1980–2000): An analysis of 55 studies of the 20m shuttle run test in 11 countries. *Sports Med*. 2003; 33:285-300.
23. Tomkinson GR, Olds TS, Gulbin J. Secular trends in physical performance of Australian children: Evidence from the Talent Search program. *J Sports Med Phys Fitness*. 2003; 43:90-98.
24. Whitehead JR, Corbin CB. Youth fitness testing: The effect of percentile-based evaluative feedback on intrinsic motivation. *Res Q Exerc Sport*. 1991; 62:225-231.