Approaches for Development of Criterion-Referenced Standards in Health-Related Youth Fitness Tests

Weimo Zhu, PhD, Matthew T. Mahar, EdD, Gregory J. Welk, PhD, Scott B. Going, PhD, Kirk J. Cureton, PhD

Introduction

Youth fitness testing in the U.S. has a rich history of over 50 years.[1–4] Key developments and changes include the development of the American Alliance for Health, Physical Education and Recreation (AAHPER) Youth Fitness Test, the birth of the health-related fitness construct, and changes in evaluation and awards.[1] The transitions from performance-related fitness to health-related fitness and from norm-referenced standards to criterion-referenced (CR) standards are noteworthy since they influenced how fitness is assessed and interpreted. The current paper reviews historical trends in fitness testing and explains the advantages of a CR framework. Methods used for establishing CR standards are described, providing a background for the subsequent articles in this supplement to the *American Journal of Preventive Medicine*.

Historical Background on Youth Fitness Testing and Standards

Early interest in youth fitness testing in the U.S. has been attributed to Kraus and Hirschland's comparative study in the 1950s,[5,6] in which they found that American youth were far less fit than their European counterparts. President Dwight D. Eisenhower, former Allied Commander in the European Theater of WWII, learned of the study and worried about the impact of fitness levels on the readiness of American youth for military service. Under his leadership, the President's Council on Youth Fitness was established in 1956, and the first AAHPER Youth Fitness Test was published in 1958. Interest in the possible link between fitness and preparedness for military service continued into the 1960s.[1] In his then well-known article "The Soft American" in *Sports Illustrated*, President-Elect John F. Kennedy stated:

> We face in the Soviet Union a powerful and implacable adversary determined to show the world that only the Communist system possesses the vigor and determination necessary to satisfy awakening aspirations for progress and the elimination of poverty and want. To meet the challenge of this enemy will require determination and will and effort on the part of all Americans. Only if our citizens are physically fit will they be fully capable of such an effort.[7]

Consistent with this vision, fitness testing protocols evolved to focus on the importance of performance. The original AAHPER Youth Fitness Test was the only national test for many years, until several states, such as California, Illinois, Indiana, New York, Oregon, South Carolina, Texas, Vermont, and Washington, started developing their own state tests during the 1950s and through the 1970s. Performance-related fitness was also consistent with the growing emphasis on sports, both in school and in society. Together, the drive for military preparedness and society's interest in sport led to performance-related fitness being the predominant paradigm during that time.

The concept and practice of health-related fitness emerged in the 1970s.[8–10] Many factors are believed to have contributed to this change: the impending end of the Cold War, better understanding of the relationship between physical fitness and health, the publication of *Aerobics* by Dr. Kenneth H. Cooper in 1968[11] and its subsequent popularity, and the development and maturation of exercise physiology, physical activity epidemiology, and measurement,[8] to name just a few of the important influences. Health-related physical fitness was defined in 1980 as "... a multifaceted continuum extending from birth to death. Affected by physical activity, it ranges from optimal abilities in all aspects of life through high and low levels of different fitness, to severely limiting diseases and dysfunction."[12] Four key traditional components of

From the Department of Kinesiology and Community Health, University of Illinois at Urbana-Champaign (Zhu), Urbana, Illinois; the Department of Exercise and Sport Science, East Carolina University (Mahar), Greenville, North Carolina; the Department of Kinesiology, Iowa State University (Welk), Ames, Iowa; the Department of Nutritional Sciences, University of Arizona (Going), Tucson, Arizona; and the Department of Kinesiology, University of Georgia (Cureton), Athens, Georgia

Address correspondence to: Weimo Zhu, PhD, Professor, Department of Kinesiology and Community Health, University of Illinois at Urbana-Champaign, 205 Freer Hall, MC-052, Urbana IL 61801. E-mail: weimozhu@illinois.edu.

0749-3797/$17.00
doi: 10.1016/j.amepre.2011.07.001

health-related physical fitness are cardiorespiratory function, body composition, muscular strength, and endurance and flexibility. The latter two are now sometimes integrated into the component defined as musculoskeletal function, reducing the number of components to three.[13] The scientific validity and measurement milestones of these key components are well described in the literature.[8]

The second noticeable change in fitness testing contributing to the shift from a norm-referenced to a CR perspective is directly related to the evolving definition and operationalization of fitness. When the interest was on performance, the focus in testing reflected the view that "the more (e.g., number of pull-ups a student can do) or less (e.g., how fast a student can finish a 1-mile run/walk test), the better," depending on the fitness measure. The norm-referenced evaluation framework, in which a student's performance is compared with his/her peers, is appropriate in this case since the emphasis is on peak performance or high-level achievement. The Presidential Physical Fitness Award Program (PCPFS) is a good example of norm-referenced evaluation, in which students must score at or above the 85th percentile on all five test items to qualify for the award.[14] Many similar examples in fitness, sports performance, and health can be found in a recent collection of norms.[15]

Technically, constructing a norm-referenced test is relatively easy as long as a nationally representative sample can be obtained and regularly updated. With such a sample, norms (e.g., percentiles and percentile ranks) can be computed and derived. There are, however, three major limitations associated with the norm-referenced evaluation framework. First, it is difficult to update norms regularly due to cost, time, and manpower constraints. As an example, the PCPFS's norms were based on the 1985 National School Population Fitness Survey,[16] and there have been no major national fitness studies in the U.S. since the 1980s (note: the other major national fitness studies in the 1980s included National Children and Youth Fitness Study I [NCYFS I], 1985; and NCYFS II, 1987).[17,18] As a result, these outdated values likely do not reflect current norms (e.g., an 85th percentile from the 1980s may now be equivalent to the 95th percentile), but rather how the values compare to the previous norms, making them inaccurate in its original evaluation framework.

The second related limitation of the norm-referenced evaluation framework is that the interpretation depends on the fitness of the reference population. The designations of average and above average have limited meaning if the majority of a population is unfit or unhealthy. The CDC obesity-evaluation criterion is a good example of this limitation. According to CDC's current standard, a child is defined as overweight with a BMI at or above the age- and gender-specific 85th percentile, and obese if the child's BMI is at or above the 95th percentile of their peers. The percentile is defined as the score value for a specific percentage of cases in a distribution of scores. If the CDC norm is current and true, it would define 15% of American children as overweight and 5% as obese. Clearly, this is not reflective of the childhood obesity epidemic that we hear about almost daily wherein one third (33%) of children and adolescents are identified as overweight or obese.[19] The difference in prevalence estimates is explained by the fact that the CDC's norms were derived from 1970s and 1980s data when American children were relatively healthy.[20] If the 85th/95th percentile standards based on today's norms are used, a large of percentage of overweight and obese children would be misclassified as having normal weight.

The third limitation of the norm-referenced evaluation framework is that it tends to reward children and youth who are already fit while potentially discouraging those who are not fit. If rewards are based on achieving the 85th percentile (as with the Presidential Fitness Award in the President's Challenge program), only highly fit youth may be motivated to try to achieve it. Less-fit youth may be less motivated because they know their chances of achieving the standard are low. If unfit students are less motivated during physical fitness testing, they may come to perceive physical education classes as a punishment/ordeal, rather than an enjoyable experience. Although other award systems are available in the President's Challenge program for students with lower levels of fitness, these limitations can be better overcome by employing the CR evaluation framework.

The concept of CR evaluation and testing was introduced in the field of education in the 1960s by Glaser.[21] However, real development and applications of CR assessment were not done until in the late 1970s and early 1980s.[22,23] The field of physical education and fitness testing embraced the new concept[24] and started to apply it in assessment practice from the late 1980s.[25–28] In contrast to the norm-referenced framework in which the evaluation of a test-taker's competency is judged relative to the performance of other students, the CR evaluation compares the test-taker's performance with an absolute criterion. In educational assessment, the "absolute criterion behavior" could be if a student has mastered the information taught in a specific subject or grade; in the context of youth health-related fitness, the interest could be if a child meets a minimal needed physical fitness level based on a criterion. Thus, the norm-referenced evaluation can be considered a relative evaluation, whereas the CR evaluation is an absolute one.

Because the criterion behavior is defined independently from the behavior of others, it is not affected by changes in a population. Therefore, the limitation of population dependence in the norm-referenced evaluation will likely have no impact on the CR-based evaluation. Similarly, although there are always some students classified as below average, average, and above average in a norm-referenced evaluation framework, there is a possibility that all students will be classified as fit or not fit based on a criterion (i.e., it is possible for everyone to either meet or not meet the CR standards) in a CR evaluation framework. As a result, the limitation of needing a fit population in order for the evaluation to be useful in the norm-referenced evaluation is eliminated in the CR evaluation framework.

Finally, since the focus is on the minimal needed fitness for a child, the evaluation standard established is often attainable by any child as long as an effort is made. Thus, the limitation of discouraging unfit participants associated with the norm-referenced approach is minimized in the CR evaluation approach. However, CR evaluation is not without its own challenges. Setting an appropriate standard, known as the cut-off score, is one of the most important challenges.

Methods Used in Setting Criterion-Referenced Standards

The fundamental interest in setting a CR standard is to determine whether a test-taker is "good enough" on the construct being measured, which could be the test-taker's reading comprehension, math problem-solving skill, or language proficiency. For health-related fitness testing, the key interest is in whether a test-taker is fit enough to be free of potential health risks. For children's fitness testing, the interest could be further extended to represent whether a child is fit enough for the future (i.e., fit enough to likely grow up to be a healthy adult). Because the key interest and outcome of the CR test/evaluation is the classification (e.g., pass versus fail, fit versus not fit, or at-risk versus needs improvement versus in the healthy fitness zone [HFZ]), the accuracy of the classification is key.

Many methods have been developed to set performance standards or simply determine CR standards. In general, these methods can be classified as either test centered or examinee centered. In the test-centered methods, a panel of experts is asked to examine each item on a competency test and set the cut-off score accordingly. In the Angoff method,[29] for example, the panel is asked to examine each item and estimate the probability that the "minimally acceptable" person would answer each item correctly. The sum of these probabilities would then represent the minimally acceptable score.

In the examinee-centered methods, the focus is on identifying examinees with/without defined minimum competency, from which the cut-off score is established. Two procedures in this category are the borderline-group and the contrasting-groups procedures,[30] and the latter has been applied to setting CR standards for a number of motor-skill tests. The contrasting group method is based on evaluating the relative distributions of a trained and an untrained group on a specific test. Standards are set to try to minimize the number of false positives (passing the standard if untrained) while also minimizing the number of false negatives (not achieving the standard if trained). Meanwhile, the health outcome–centered method has been the predominant approach in setting CR standards for health-related fitness tests.

The key steps of the health outcome–centered method include:

- determine the components of health-related fitness, which often include cardiorespiratory fitness or aerobic capacity, body composition, and muscular fitness (i.e., muscular strength, endurance, and flexibility);
- select a criterion measure, as well as field tests, of the fitness component (e.g., VO_2max as the criterion measure and 1-mile run/walk and Progressive Aerobic Cardiovascular Endurance Run [PACER] as the field tests for cardiorespiratory fitness);
- determine the relationships between the criterion measure/field tests and health-outcome measures, which could be mortality, an individual factor (e.g., if a person's blood pressure is high), or a group of health-risk measures (e.g., if a person has metabolic syndrome);
- set the standards or cut-off scores according to the relationship determined (i.e., determine the point or level on which a fitness parameter is associated with an increased risk of a disease outcome or risk factors of the disease);
- validate or cross-validate using additional measures and samples.

The procedures used to set up the original CR standards for body composition in FITNESSGRAM® provide a good example of these steps. The original cut-off scores for body composition were based on the relationship between body fatness and cardiorespiratory disease risk factors, including blood pressure, total cholesterol, and serum lipoprotein ratios in children and adolescents[31] (Going et al.[32] in this supplement has a detailed review of this procedure). The original cut-off scores for aerobic capacity were developed in a slightly different way by Cureton[33] in 1994. Based on an extensive literature review, morbidity and mortality in adults were chosen as the health outcomes. Because morbidity (caused mainly by unwanted pregnancy, substance abuse, physical/sexual abuse, and stress) and mortality (caused mainly by accidents, suicide, and homicide) in children and youth is not directly related to physical fitness, cut-off scores cannot be directly related to children's morbidity and mortality data. Instead, Cureton[25,33] derived the cut-off scores based on the information of both adult morbidity and mortality and age-/growth-related changes in VO_2max. The assumptions and decisions used in setting these standards have been supported by subsequent studies based on related health-risk factors in other children (Welk et al.[34] in this supplement contains additional discussion). However, as described in the preface to this supplement, several unresolved issues with the standards necessitated a re-evaluation.

Critical Issues and Challenges in Setting Criterion-Referenced Standards

Although CR evaluation is able to address the shortcomings of the norm-referenced evaluation and fits the needs of health-related fitness assessment very well, it has its own issues and challenges, including the selection of health outcome measures, equivalence of cut-off scores across field tests, consequence of misclassification, and cross-group and cultural differences.

Selecting a Health-Outcome Measure

Although the theoretic relationships among physical activity, fitness, and health[35] and health-related fitness and health[8] have been well described in the literature, limited information is available on which health outcome should be employed when validating heath-related fitness assessments. Like fitness, health is a construct. In the past, it was simply defined as "freedom from physical disease or pain." A more accepted definition of health now is the definition set by the WHO in 1948: "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity."[36]

In theory, there are endless ways to measure health. A natural question then is: Which health measure/outcome should be used in validating health-related fitness? There is no absolute correct answer to this question, and "select the most appropriate one" (i.e., select the most appropriate measure/outcome(s) based on the existing theoretic and empirical knowledge base and evidence) may be the best answer. As described in the previous text, the health outcomes in determining body composition standards included total cholesterol, serum lipoprotein ratios, and blood pressure,[31] whereas morbidity and mortality were the measures when setting aerobic-capacity standards.[33]

Another related question in selecting health outcome measures is: How many outcome measures should be selected? Again, there is no absolute correct answer to this question, but the recommendation of the authors of this paper is to consider and examine all available outcome measures although there is no need to use all of them when making the final decision. As described in this supplement, metabolic syndrome was selected as the most appropriate outcome measure for establishing new standards for both body fatness and aerobic capacity. Finally, another related selection question is which age group should be the focus: children, youth, adults, or older adults. As illustrated in both body composition and aerobic-capacity standard setting, the decision depends on the assessment of interest (i.e., to determine the current fitness status, to predict future fitness status, or both), along with other information availability. The authors' recommendation, once again, is to try to use all available information and make a decision accordingly.

Equivalence of Cut-Off Scores

As with the health outcome measures, a number of field tests are often used simultaneously to measure the same construct. For example, the 1-mile run/walk, PACER, and 1-mile walk tests are used to measure aerobic capacity in FITNESSGRAM. Usually, when a new field test is developed, the cut-off scores often will be set based on a new, small-sample study or simply derived from the normative data or the existing literature by an expert panel.[37] Because of sample variations and other factors, the standard equivalencies among field tests are often not consistent. For example, Mahar et al.[38] reported that 34% of 4th- and 5th-grade girls who achieved PACER standards failed to pass the 1-mile run/walk standards (see also Beets and Pitetti[39]). Although it is expected that there will be a difference in achievement levels among tests, such a large difference is not acceptable.

As another example, several field tests are frequently used to measure upper-body muscular strength: pull-ups, flexed arm hang, push-ups, modified pull-ups, and modified push-ups. The scoring formats range from the number of repetitions to time in seconds performing a test. According to a validity study of five such field tests,[40] only moderate correlations (r ranged from 0.50 to 0.70) were found among these tests. Therefore, classification systems developed for these tests will likely be inconsistent.

A simple solution for this inconsistency problem is to adopt a standardized single-test approach, (i.e., use a single test for a fitness component). Although theoretically sound, this single-test approach is unlikely to be adopted in reality due to many historical (e.g., one country/area has already used a specific test for many years) and practical (e.g., limitations in space and facilities) reasons. Fortunately, this problem can be addressed by employing a new "primary test centered equating method,"[41] described briefly in the following text (and in Boiarskaia et al.[42] in this supplement).

Consequences of Misclassification

There will be misclassification when an assessment serves a classification role no matter how well the related cut-off score is set up. There are usually two kinds of misclassifications: false-positive classification (e.g., an unfit test-taker misclassified as fit in the context of fitness testing) and false-negative classification (a fit test-taker misclassified as unfit). As well described by Cureton and Warren,[25] the false-positive classification may be a more serious error in this case since the misclassified test-takers may get the wrong impression that they are fit enough already, and therefore not exercise at a desirable level and consequently fail to reduce or even increase their risk of disease. Although a call was made 20 years ago by Cureton and Warren[25] for more research to understand the consequences of these misclassifications, little progress has been made in this area.

Cross-Group and Culture Differences

Finally, whether a cut-off score should be set up differently for various subpopulations must be empirically examined and determined. Although age and gender have often been taken into consideration in setting cut-off scores, many other factors, such as ethnicity and disability, have not been considered. It is noted that to address cross-cultural differences, WHO developed and published an international BMI standard in 2006.[43] The WHO's standard is norm-referenced as is the CDC's standard, which was discussed earlier as being a reference population issue. This is an area that needs more research.

New Measurement and Statistical Methods and Applications

Some new measurement and statistical methods have been developed to facilitate establishment of standards. In particular, the use of test-equating procedures and approaches based on receiver operating characteristic (ROC) curves offer considerable potential for addressing some of the CR evaluation–related issues and challenges noted in the previous text.

Test Equating

Equating is a set of statistical procedures that puts two or more tests that measure the same construct in different ways onto the same scale so they can be directly compared.[44,45] To address the issue of inconsistency in setting a standard for cross-test classification when measuring aerobic capacity, Zhu et al.[41] proposed the primary test centered equating method. The primary field test refers to a field test whose validity related to the criterion test has been well documented (e.g., 1-mile run/walk for estimating VO_2max and skinfold measurements for predicting body fat percentage). The key steps in the method for setting a standard for a new field test, whose validity has been confirmed by other studies, are as follows:

- select a validated field test (e.g., validity and reliability coefficients ≥0.80) as the primary field test;
- administer both the primary field test and new field test to a large sample (say $n = 200$) from the targeted population using a counterbalanced order; make sure there is adequate rest time between tests to avoid carryover effect;
- set the field test onto the scale of the primary field test using an equating procedure;
- use the cut-off scores already set for the primary test or set them based on the equivalent relationship developed.

Using aerobic assessment as an example, the primary field test is the 1-mile run/walk, and the "new" field test is the PACER. After the PACER is equated to the scale of the 1-mile run/walk, the equivalent 1-mile run/walk score can be used to estimate VO_2max and determine HFZ classification using the cut-off score set for the 1-mile run/walk or VO_2max. The concept of this new cut-off score setting method is illustrated in Figure 1. The method's validity has been confirmed by Zhu et al.[41] and further cross-validated in the study by Boiarskaia et al.[42] reported in this supplement.

Receiver Operating Characteristic Curves

Many statistical procedures have been developed to evaluate accuracy and consistency of classifications. Percentage agreement and kappa statistics are among the most popular.[46] A contingency table can best illustrate these statistics (Figure 2). When determining the classification accuracy of a field test, the focus is on the agreement between the criterion measure, which is used to represent true classification status, and the field test. Cases classified positively by both the field test and the criterion measure are categorized as true positives (TP), whereas cases classified negatively by both tests are categorized as true negatives (TN). A false-negative (FN) error occurs when a field test erroneously indicates that a person does not achieve the standard on the criterion. Alternately, a false-positive (FP) error occurs when a field test incorrectly identifies a person as achieving the standard on the criterion.

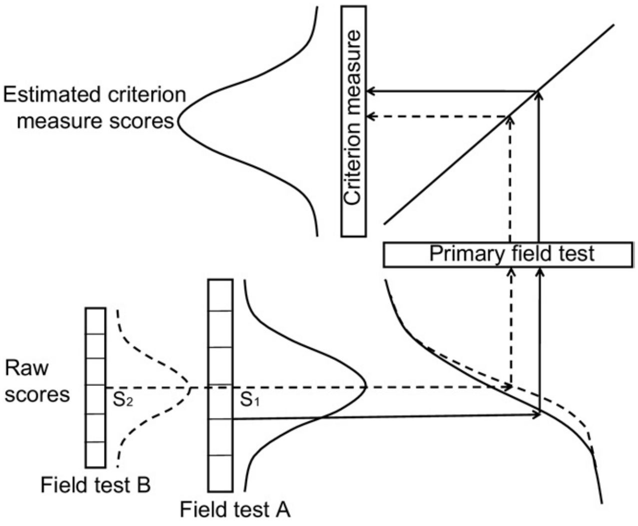

Figure 1. Conceptual illustration of the primary field test centered equating method for cut-off score setting

Note: Using aerobic assessment as an example, the criterion measure is VO_2max, and the primary field test is 1-mile run. Field Tests A and B are PACER and 1-mile walk, respectively. After Tests A and B are equated to the scale of the primary field test, the raw testing scores of S_1, who performed the PACER test, and S_2, who performed the 1-mile walk test, can be transferred onto the scale of 1-mile run and used to estimate their VO_2max. Now their performance can be evaluated and compared on the same scale.

Reprinted with permission from © American Alliance for Health, Physical Education, Recreation and Dance[41]

PACER, Progressive Aerobic Cardiovascular Endurance Run; S_1, Subject 1; S_2, Subject 2

Note that in the context of setting health-related fitness standards, one, or a set of, health measure(s)/outcome(s) is used as the criterion measure, and fitness tests as the field test. Using a similar analogy, the health measure/outcome can be classified as healthy (H) and unhealthy (U), and fitness measure can be classified as fit (i.e., health-risk free [F]) and not fit (i.e., having some health risks [N]). Accordingly, HF (being classified both as healthy and fit) = TP, UN (unhealthy/not Fit) = TN, HN (healthy/not fit) = FN, and UF (unhealthy/fit) = FP (Figure 3).

Two commonly used statistical indexes for classification accuracy are the Proportion of Agreement [$P = (TP + TN)/(TP + TN + FP + FN)$] and the kappa statistic, which removes the chance factor from P.[46] To determine the optimal cut-off score, one can vary the cut-off scores of the field test and calculate the corresponding agreement statistics, as well as FP and FN rates. The optimal cut-off score is the one with the highest agreement and fewest classification errors.

The development of ROC curves provides a graphical procedure that enables errors to be systematically evaluated across all possible scores.[47] The ROC curve displays the sensitivity (probability of correctly detecting TP results) and specificity (probability of correctly detecting TN results) of a particular field test for a range of cut-off points or thresholds. Ideally, a diagnostic cut-off point value should result in low FP and low FN rates across a reasonable range of cut-off values. The primary indicators of ROC analyses can be calculated from the contingency table in Figure 2:

- accuracy (i.e., P) = (TP + TN)/(TP + TN + FP + FN);
- sensitivity = TP/(TP + FN);
- specificity = TN/(FP + TN).

The unique value of ROC curves is that cut-off points can be selected based on the relative importance of sensitivity or specificity (i.e., the ROC approach makes it possible to weigh the relative costs of one type of error over another). Although ROC has been widely used in clinical medicine and was introduced to kinesiology a few years ago,[48,49] it has not been widely employed in setting cut-off scores in health-related fitness testing. Studies reported by Laurson et al.[50] and Welk et al.[34] in this supplement represent the first wave of ROC applications in this area.

Remaining Issues and Future Research Needs

There are still a number of unresolved issues in setting cut-off scores in health-related fitness measurement and evaluation, namely standards for muscular fitness (strength, endurance, and flexibility), understanding CR-based fitness growth assessment and evaluation, and related matters of motivation.

		Criterion measure	
		Positive	Negative
Field test	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Figure 2. Contingency tables for classification accuracy and errors in the context of criterion-referenced fitness testing

		Health measure/outcome	
		Healthy (H)	Unhealthy (U)
Fitness test	Fit (F)	HF	UF
	Not fit (N)	HN	UN

Figure 3. Contingency tables for classification accuracy and errors in the context of health-related fitness testing

Standards for Muscular Fitness

The cut-off scores of aerobic capacity and body composition have been well studied and established, as illustrated in this supplement. The well-described relationship between health measures and these two variables, as well as available rich data and information, are perhaps the reasons.[3,8] In contrast, although the validity and reliability of commonly used tests of muscular strength, endurance, and flexibility are generally well supported,[51] the relationships between these tests and health have not been well established.

For instance, sit-up and sit-and-reach tests were included in health-related fitness testing because they were believed to be good indicators of lower-back health.[12,52] Others, however, showed that there is little, if any, relationship between physical fitness and lower-back pain, a symptom of bad lower-back health.[37,53,54] Plowman,[37] based on a comprehensive review, stated almost 20 years ago: "While items of trunk strength/endurance and lower-back and hamstring flexibility can be marginally accepted as predictor tests, what absolute values on these tests might prove to be protective is a total unknown due to the wide overlap of scores between those who eventually had lower-back problems and those who did not," which is still true today. This is clearly an area requiring more research.

Criterion Referenced–Based Fitness Growth

The focus on health-related fitness and CR evaluation has been concurrent with the relationship between fitness status and health, and little effort has been made to understand criterion-related fitness growth in children. When studying CR fitness growth, the focus shifts to whether a test-taker is on track to being fit, known also as growth to standard. There are several reasons for this understudied research area. An assumption in youth fitness testing is that fit children grow up to become fit adults, but evidence to support this fit child = fit adult hypothesis is limited. It is likely that fitness needs may change along with normal growth and maturation changes, and this needs confirmation. The application of LMS (L = skewness, M = median, and S = coefficient of variation) growth curves provides a way to model growth-related changes over time, and new curves reported in this supplement were developed specifically for this purpose.[50,55]

Another consideration related to growth is that due to many factors (e.g., parent's education and SES, and local preschool sport program availability), children enter school at different fitness levels. Children's improvement over time (relative to their initial status) should be the basis of education so these data can be used for the evaluation of the effectiveness of a school, program, and teacher. When linked with a predetermined evaluation standard, this type of evaluation is referred to as criterion-related growth, a critical part of standard-based assessments and evaluations. The concepts of criterion-related growth, value-added assessment, and modeling are being introduced and used in educational research and standard-based assessments.[56–58] Physical education and fitness researchers and practitioners need to catch up with the progress already being made in these areas.

Standards and Students' Motivation

It is generally believed that a norm-referenced evaluation will discourage students whose fitness levels might be moderate or low since only a small percentage of students will be able to meet the standards under such an evaluation framework. For example, less than 5% of students could actually qualify for the President's Challenge Award (i.e., scored at the 85th percentiles or higher for all five tests).[59] In contrast, it is believed that in a CR-evaluation framework, such as FITNESSGRAM, children are encouraged to focus on their own health status rather than their level compared with others.[59] As a result, students are able to enhance their motivation and self-confidence.

A recent study provided some support for such beliefs:[60] A majority of students studied (86%) believed fitness tests enhanced their knowledge of the importance of being healthy, and motivated them to be more physically active. Meanwhile, according to the report from the latest Texas Youth Fitness Study,[61] many teachers still reported negative experiences when using FITNESSGRAM, such as apathy/unwillingness, self-consciousness, frustration, and teasing. More studies are needed to understand the impacts, especially long-term ones, of norm- and criterion-referenced fitness testing on evaluations, and on subsequent behavior of the youth evaluated.

Conclusion

In summary, two of the most notable changes in youth fitness testing are the shift from performance-centered assessment to health-related fitness testing, and from norm-referenced evaluation to CR evaluation. Setting the standards, or cut-off scores, is one of the most important issues in the design of a CR test. Many methods have been developed to set cut-off scores in CR tests, and the health outcome– centered method is the most popular and effective one for setting standards for health-related fitness tests.

Critical issues related to this method include selecting appropriate health outcomes, equivalence of cut-off scores, consequences of misclassification, and cross-group and cultural differences. Recent developments and applications in statistical techniques, such as test equating and ROC, have proven to be helpful in addressing some of these issues. Several of these techniques were specifically employed in the development of the new body composition and aerobic-capacity standards for FITNESSGRAM. Although progress has been made in these areas, many issues remain; including the need for setting standards for muscular components and determining CR-based fitness growth.

Publication of this article was supported by The Cooper Institute through a philanthropic gift from Lyda Hill.

No financial disclosures were reported by the authors of this paper.

References

1. Morrow JR Jr., Zhu W, Franks BD, Meredith MD, Spain C. 1958–2008: 50 years of youth fitness tests in the U.S. Res Q Exerc Sport 2009;80(1):1–11.
2. Plowman SA, Sterling CL, Corbin CB, Meredith MD, Welk GJ, Morrow JR Jr. The history of FITNESSGRAM. J Phys Act Hlth 2006;3(S2):S5–20.
3. Safrit MJ. The validity and reliability of fitness tests for children: a review. Pediatr Exerc Sci 1990;2(1):9–28.
4. Seefeldt V, Vogel P. Physical fitness testing of children: a 30-year history of misguided efforts? Pediatr Exerc Sci 1989;1:295–302.
5. Kraus H, Hirschland RP. Muscular fitness and health. JOPERD 1953;24(10):17–9.
6. Kraus H, Hirschland RP. Minimum muscular fitness tests in school children. Res Q 1954;25:178–88.
7. Kennedy JF. The soft American. Sports Illustrated 1960;Dec;13(26): 14–7.
8. Jackson AS. The evolution and validity of health-related fitness. Quest 2006;58(1):160–75.
9. Pate RR. The evolving definition of physical fitness. Quest 1988;40(3): 174–9.
10. Pate RR. A new definition of youth fitness. Phys Sports Med 1983; 11(4):77–83.
11. Cooper KH. Aerobics. New York NY: Bantam Books, 1968.
12. American Alliance for Health, Physical Education, Recreation and Dance. Health related fitness test. Reston VA: American Alliance for Health, Physical Education, Recreation and Dance, 1980.
13. Corbin CB, Pangrazi RP. FITNESSGRAM and ACTIVITYGRAM: an introduction. In: Welk GJ, Meredith MD, eds. FITNESSGRAM/ ACTIVITYGRAM reference guide. Dallas TX: The Cooper Institute, 2008.
14. The President's Council on Fitness, Sports & Nutrition. The President's Challenge. Choose a Challenge. Physical Fitness Test. Award Benchmarks. www.presidentschallenge.org/challenge/physical/benchmarks.shtml.
15. Hoffman J. Norms for fitness performance, and health. Champaign IL: Human Kinetics, 2006.
16. Reiff G, Dixon W, Jacoby D, Ye GX, Spain C, Hunsicker P. The President's Council on Physical Fitness and Sports 1985: national school population fitness survey. Washington DC: U.S. Government Printing Office, 1986.
17. Ross JG, Gilbert GG. The National Children and Youth Fitness Study: a summary of findings. JOPERD 1985;56(1):45–50.
18. Ross J, Pate R, Relpy L, Gold R, Svilar M. The National Children and Youth Fitness Study II: new health-related fitness norms. JOPERD 1987;58(9):66–70.
19. Ogden CL, Carroll MD, Curtin LR, Lamb MM, Flegal KM. Prevalence of high body mass index in U.S. children and adolescents, 2007–2008. JAMA 2010;303(3):242–9.
20. Kuczmarski RJ, Ogden CL, Grummer-Strawn LM, et al. CDC growth charts: U.S. Advance data from vital and health statistics, no. 314. Hyattsville MD: National Center for Health Statistics, 2000.
21. Glaser R. Instructional technology and the measurement of learning outcomes: some questions. Am Psychol 1963;18:519–21.
22. Popham WJ. Criterion referenced measurement. Englewood Cliffs NJ: Prentice Hall, 1978.
23. Berk RA, ed. Criterion-referenced measurement: the state of the art. Baltimore MD: Johns Hopkins University Press, 1980.
24. Safrit MJ, Baumgartner TA, Jackson AS, Stamm CL. Issues in setting motor performance standards. Quest 1980;32(2):152–62.
25. Cureton KJ, Warren GL. Criterion-referenced standards for youth health-related fitness tests: a tutorial. Res Q Exerc Sport 1990;61(1): 7–19.
26. Kalohn JC, Wagoner K, Gao LG, Safrit MJ, Getchell N. A comparison of two criterion-referenced standard setting procedures for sports skills testing. Res Q Exerc Sport 1992;63(1):1–10.
27. Safrit MJ. Criterion-referenced measurement: validity. In: Safrit MJ, Wood TM, eds. Measurement concepts in physical education and exercise science. 1st ed. Champaign IL: Human Kinetics, 1989.
28. Looney MA. Criterion-referenced measurement: reliability. In: Safrit MJ, Woods TM, eds. Measurement concepts in physical education and exercise science. 1st ed. Champaign IL: Human Kinetics, 1989.
29. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, ed. Educational measurement. 2nd ed. Washington DC: American Council on Education, 1971.
30. Zieky MJ, Livingston SA. Manual for setting standards on the basic skills assessment tests. Princeton NJ: Educational Testing Service, 1977.
31. Williams DP, Going SB, Lohman TG, et al. Body fatness and risk for elevated blood pressure, total cholesterol and serum lipoprotein ratios in children and adolescents. Am J Public Health 1992;82(3):358–63.
32. Going SB, Lohman TG, Cussler EC, Williams DP, Morrison JA, Horn PS. Percent body fat and chronic disease risk factors in U.S. children and youth. Am J Prev Med 2011;41(4S2):S77–86.
33. Cureton KJ. Aerobic capacity. In: Morrow JR Jr., Falls HB, Kohl HW III, eds. The Prudential FITNESSGRAM technical reference manual. Dallas TX: Cooper Institute for Aerobics Research, 1994.
34. Welk GJ, Laurson KR, Eisenmann JC, Cureton KJ. Development of youth aerobic-capacity standards using receiver operating characteristic curves. Am J Prev Med 2011;41(4S2):S111–6.
35. Bouchard C, Shephard RJ. Physical activity, fitness, and health: the model and key concepts. In: Bouchard C, Shephard RJ, Stephens T, eds. Physical activity, fitness, and health: international proceedings and consensus statement. Champaign IL: Human Kinetics, 1994.
36. World Health Organization. Preamble to the Constitution of the WHO as adopted by the International Health Conference, New York, 19–22 June 1946, and entered into force on 7 April 1948.
37. Plowman SA. Criterion referenced standards for neuromuscular physical fitness tests: an analysis. Pediatr Exerc Sci 1992;4(1):10–9.
38. Mahar MT, Rowe DA, Parker CR, Mahar FJ, Dawson DM, Holt JE. Criterion-referenced and norm-referenced agreement between the mile run/walk and PACER. Meas Phys Educ Exerc Sci 1997;1(4):245–58.
39. Beets MW, Pitetti KH. Criterion-referenced reliability and equivalency between the PACER and 1-mile run/walk for high school students. J Phys Act Health 2006;3(S May):S17–29.
40. Pate RR, Brugess ML, Woods JA, Ross JG, Baumgartner T. Validity of field tests of upper body muscular strength. Res Q Exerc Sport 1993;64(1):17–24.
41. Zhu W, Plowman SA, Park Y. A primer-test centered equating method for cut-off score setting. Res Q Exerc Sport 2010;81(4):400–9.

42. Boiarskaia EA, Boscolo MS, Zhu W, Mahar MT. Cross-validation of an equating method linking aerobic FITNESSGRAM® field tests. Am J Prev Med 2011;41(4S2):S124–30.
43. WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development. Geneva, Switzerland: World Health Organization, 2006.
44. Zhu W. Test equating: what, why, how? Res Q Exerc Sport 1998; 69(1):11–23.
45. Zhu W. Scales, norms, and score comparability. In: Wood T, Zhu W, eds. Measurement theory and practice in kinesiology. Champaign IL: Human Kinetics, 2006.
46. Safrit MJ, Wood TM. Introduction to measurement in physical education and exercise science. St. Louis MO: Mosby, 1995.
47. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39(4):561–77.
48. Looney MA. Measurement issues in the clinical setting. In: Wood T, Zhu W, eds. Measurement theory and practice in kinesiology. Champaign IL: Human Kinetics, 2006.
49. Jackson AS. Preemployment physical testing. In: Wood T, Zhu W, eds. Measurement theory and practice in kinesiology. Champaign IL: Human Kinetics, 2006.
50. Laurson KR, Eisenmann JC, Welk GJ. Body fat percentile curves for U.S. children and adolescents. Am J Prev Med 2011;41(4S2): S87–92.
51. Plowman SA. Muscular strength, endurance and flexibility assessments. In: Welk GJ, Meredith MD, eds. FITNESSGRAM/ ACTIVITYGRAM reference guide. Dallas TX: The Cooper Institute, 2008.
52. Payne N, Gledhill N, Katzmarzyk PT, Jamnik V. Health-related fitness, physical activity, and history of back pain. Can J Appl Physiol 2000;25(4):236–49.
53. Jackson AW, Morrow JR Jr., Brill PA, Kohl HW, Gordon NF, Blair SN. Relations of sit-up and sit-and-reach tests to low back pain in adults. J Orthop Sports Phys Ther 1998;27(1):22–6.
54. Nachemson AL. Exercise, fitness, and back pain. In: Bouchard C, Shephard RJ, Stephens T, Sutton JR, McPherson BD, eds. Exercise, fitness, and health: a consensus of current knowledge. Champaign IL: Human Kinetics, 1990.
55. Eisenmann JC, Laurson KR, Welk GJ. Aerobic fitness percentiles for U.S. adolescents. Am J Prev Med 2011;41(4S2):S106–10.
56. Amrein-Beardsley A. Methodological concerns about the education value-added assessment system. Educ Res 2008;37(2):65–75.
57. Braun HI. Using student progress to evaluate teachers: a primer on value-added models. Princeton NJ: Educational Testing Service, 2005.
58. Braun H, Chudowsky N, Koenig J, eds. Getting value out of value-added: report of a workshop. Washington DC: National Academy of Science, 2010.
59. Koebel CI, Swank AM, Shelburne L. Fitness testing in children: a comparison between PCPFS and AAHPERD standards. J Strength Cond Res 1992;6(2):107–14.
60. Sampson BB. Children's perceptions of the FITNESSGRAM fitness test [Master's thesis]. Salt Lake City UT: Brigham Young University, 2008.
61. Zhu W, Welk G, Meredith M, Boiarskaia E A survey of Texas schools' physical education programs and policies. Res Q Exerc Sport 2010; 81(S3):S42–52.