Development of New Criterion-Referenced Fitness Standards in the FITNESSGRAM® Program

Rationale and Conceptual Overview

Gregory J. Welk, PhD, Scott B. Going, PhD,
James R. Morrow Jr, PhD, Marilu D. Meredith, EdD

Fitness testing is a common, if not characteristic, component of most physical education (PE) programs.[1,2] The FITNESSGRAM® youth fitness program has been widely used in school-based physical education programming to facilitate the collection and processing of youth fitness and physical activity data.[3] The FITNESSGRAM program provides teachers with a battery of validated field-based fitness and activity assessments to facilitate effective physical education programming.[4] Appropriate uses of fitness and activity assessments include teaching self-monitoring skills, promoting educational outcomes, providing personalized information about levels of health-related fitness/activity, and assisting in evaluating school-level outcomes over time (for tracking or curricular assessment).[5]

A recent supplement in *Measurement in Exercise Science and Physical Education* outlined the many advantages of school-based fitness testing when conducted in an educationally, pedagogically, and motivationally sound manner.[6] Mahar and Rowe[7] described the importance of using valid and reliable assessments and how the context and purpose of testing influence the way it is perceived by students and parents. Silverman and colleagues[8] described the pedagogic value of fitness testing and how it can be conducted most effectively to promote physical education learning outcomes. Wiersma and Sherman[9] emphasized the psychological aspects of fitness testing and how it can be set up to emphasize positive experiences such as maximizing effort, enjoyment, and motivation. Welk[10] discussed the value of coordinated fitness and activity assessments and how the information can help youth establish lifelong patterns of physical activity. The importance and value of standardized testing was also summarized more recently by Morrow and Ede.[11] Although there has been some debate in the field, the consensus is that youth fitness testing can provide valuable information if used properly within a quality physical education program.

A unique advantage of the FITNESSGRAM program for coordinated fitness testing is the use of criterion-referenced (CR) standards that reflect how fit children need to be to receive health benefits.[3] Another unique advantage of FITNESSGRAM is that it enables teachers to produce personalized reports that provide information about the child's level of health-related fitness and suggestions to improve his/her fitness profile. The distribution of fitness reports can help educate both children and parents about health-related fitness and promote involvement in physical activity. Thus, FITNESSGRAM is positioned to both assess and promote physical fitness in youth.

In recent years, many large districts and several states have adopted requirements for coordinated youth fitness testing with FITNESSGRAM.[11] California has had legislation since 2003, Texas passed legislation, and Georgia recently announced a plan for coordinated state-level testing. Many other states have also implemented plans for more-coordinated fitness testing. These mandates have been driven, in part, by the increased public health attention on childhood obesity. Schools have not been implicated as a "cause" of the epidemic; they are clearly viewed as being a critical part of the solution. This is due primarily to the ability to reach and influence large numbers of children in a comprehensive and systematic way. At a broader level, the coordinated tracking of data with FITNESSGRAM provides a way for districts and states to evaluate curricular changes designed to promote physical activity and prevent obesity. It also enables tracking of overall trends in the population. These broader applications are appropriate uses of FITNESSGRAM data[5]; however, they place emphasis on different attributes or characteristics of youth fitness testing.

From the Department of Kinesiology, Iowa State University (Welk), Ames, Iowa; the Department of Nutritional Sciences, University of Arizona (Going), Tucson, Arizona; the Department of Kinesiology Health Promotion and Recreation, University of North Texas (Morrow), Denton; and the Cooper Institute (Meredith), Dallas, Texas

Address correspondence to: Gregory J. Welk, PhD, Department of Kinesiology, Iowa State University, 257 Forker Building, Ames IA 50011. E-mail: gwelk@iastate.edu.

0749-3797/$17.00
doi: 10.1016/j.amepre.2011.07.012

The transition to larger adoptions has shifted emphasis from awareness/educational applications toward aggregate school-, district-, and state-level reporting. The visibility of these aggregate reports (and the implications of the findings) place greater emphasis on the validity of the standards. One key issue is whether age- and gender-related trends in fitness are indicative of actual differences in health-related risks. The FITNESSGRAM standards are age- and gender-specific, but if they are used for population surveillance, it is important to ensure that standards reflect actual changes in health-related risk. Another key issue is with classification agreement since the FITNESSGRAM program provides teachers with several options for assessing both cardiovascular fitness and body composition. The availability of multiple options for evaluating fitness is important for teachers and schools, but it can present challenges when data are used to evaluate patterns or trends in the population. Achievement levels need to be matched to ensure comparability across schools regardless of the tests used. Some illustrations of these issues are summarized in the following text to demonstrate the challenges and issues associated with the interpretation of aggregated fitness data.

In Texas, teachers were required by Senate Bill 530 to complete an assessment of health-related fitness but had choices about which test item to use. Schools were required to submit their results to the state, and these data were used to report levels of health-related fitness in the population. A comprehensive research supplement was recently released summarizing results from this project.[12] A main outcome of interest to the state was the percentage of students who could meet the various health-related standards. Clear gender- and age-related patterns were evident in the cardiovascular fitness results, with girls having higher levels of achievement than boys for most ages.[13] The patterns indicate that health-related fitness declines with age but also suggest that girls have higher levels of health-related fitness than boys. Although this may be true, it may also be an artifact of the age- and gender-specific standards used to evaluate the data. The psychometric properties of the FITNESSGRAM tests have been well established,[14] but few studies have directly evaluated the utility of the health-related standards related to risks for chronic disease.

Another observation from the Texas Youth Fitness study was that there were differences in achievement depending on the fitness test item selected.[13] For cardiovascular fitness, teachers can choose from the mile run or the Progressive Aerobic Cardiovascular Endurance Run (PACER) test. The PACER test is the recommended test due to the more standardized and objective methodology, pedagogic utility, psychological advantages, and the built-in pacing; however, many schools use the mile run due to familiarity with the assessment or to provide a unique (performance-based) challenge for the students. The mile run and the PACER test are processed using the same health-related standard, but a comparison of results demonstrated clear age and gender differences in achievement depending on what test items were used. Achievement levels were considerably higher for young girls than young boys on the PACER, but this pattern was not evident with the mile run. Although it is possible that this could be due to testing effects or motivation issues, it may also reflect differences in the ways that the tests are scored and processed. Both tests have been shown to have good reliability and validity,[14] but it is possible that classification agreement can still be limited.

The examples just described illustrate issues that needed to be examined and/or resolved with the FITNESSGRAM aerobic fitness assessments. Other evidence revealed potential discrepancies in the body composition assessments. The body fat standards were established based on associated health risks with excess fatness,[15] and BMI standards were established to correspond to these values. However, the values yielded some unique age-related patterns. For some ages, the FITNESSGRAM standards were higher than the CDC percentile norms, but for other ages, they were lower. This created confusion when schools compared FITNESSGRAM results to data processed using the widely used CDC body composition standards. The FITNESSGRAM standards were not developed to correspond to CDC percentile standards, but it was important to re-evaluate the predictive utility of the body fat standards and corresponding risks associated with BMI.

The FITNESSGRAM Scientific Advisory Board discussed these issues and initiated a process to re-evaluate and redevelop the FITNESSGRAM standards for aerobic fitness and body composition. The papers in this supplement[16–25] summarize the process and evidence used to establish new FITNESSGRAM standards. A brief history of fitness testing and fitness standards provides valuable context for the transition to (and importance of) health-related standards. This is followed by brief descriptions of the individual papers.

Background on Fitness Standards and Overview of Supplement

Early fitness tests grew out of a concern for military preparedness.[2] Not surprisingly, they emphasized performance-related traits such as speed, agility, and muscular strength. Although of some interest, some tests of athletic fitness do not necessarily relate to health. Recognition of this fact, coupled with concern that many apparently healthy children could not pass some tests, led to interests in tests

focused on health.[3] By the 1980s, obesity was on the rise, and there was increasing evidence that excess adiposity and low levels of aerobic fitness significantly predicted chronic disease risk. It was about this time that health-related tests began to be adopted with items that assessed body composition, aerobic endurance, and lower-back flexibility, all considered important public health concerns.[2,26]

Tests require evaluative standards, and two main approaches have been used for setting youth fitness standards (norm-referenced and CR). Tests of athletic/performance fitness are usually norm-referenced, with standards based on population distributions of scores on the item of interest. A limitation is that passage is dependent on one's performance relative to the group (e.g., a score above a predetermined percentile, such as the 85th percentile of the population). Percentile-based standards predetermine passing levels, which remain constant, even when the population distribution shifts. Defining child and adolescent obesity, for example, as the age- and gender-specific BMI ≥95th percentile sets obesity prevalence at 5%. Reports of a pediatric obesity epidemic, with prevalence at 15%–20%, are confusing, unless it is made clear that the standards are based on BMI distributions from past surveys conducted before the current assessment.[27] Unlike a norm-referenced standard, a CR standard is set based on how the score relates to an appropriate reference value or criterion. In the case of fitness tests, the standards are typically referenced to an appropriate health indicator.

Health-related (CR) standards were introduced by FITNESSGRAM in 1987. These standards established a single standard for each test item. Scores above the cutoff were classified as acceptable; no label was associated with scores below the cutoff. The cut-off points were based on empirical data, normative data, and the professional judgment of an advisory council.[28] They were intended to set a minimum level of performance on each test item that was consistent with good health (minimal disease risk) and adequate function (the ability to conduct tasks of daily life) independent of the population tested or the proportion of the population that meets the standard. The FITNESSGRAM CR standards were the first for youth fitness that were put into widespread national and international use. In 1992, the concept of a fitness zone replaced the notion of a single cut-off score. Results since 1992 have been evaluated as either in the needs improvement zone (NIZ) or in or above the healthy fitness zone (HFZ). The goal for all participants was achievement of HFZ, but it was recognized that scores higher than the upper limit of HFZ were attainable and healthy, with the possible exception of excessive leanness.

Criterion-referenced standards for health-related fitness require criterion and field test items that relate to health status and function. They also require scores that are responsive to health status and physical activity. The reliability and validity must be established for both the field tests and the CR standards. Available physiologic and psychometric research on each item in the FITNESSGRAM battery was first presented in 1994 and updated in 2003 and 2008.[14] These standards have served the program well, but advances in new statistical methods and the availability of new data sets provide opportunities to re-examine some of the FITNESSGRAM youth fitness standards and their utility as appropriate field-based indicators of health risk.

Although the FITNESSGRAM program provides assessments for a variety of dimensions of health-related fitness, the research in this supplement addressed only body composition and aerobic fitness since these two dimensions of fitness have more established links with health and are of greater public health interest. The article in the supplement by Zhu et al.[16] reviews issues associated with setting standards and the various statistical methodologies that have been used to establish standards. The paper in this supplement by Going and colleagues[17] illustrates the associations of chronic disease risk factors with excess body fat and provides a foundation for establishing an appropriate indicator of health risk for youth. Associations among various risk factors are complex, so a common approach has been to examine health status using composite indicators of metabolic syndrome. Age- and gender-specific levels of risk have been established using nationally representative data, and standardized definitions of metabolic syndrome have been established for determining the presence of metabolic syndrome in adolescents.[29] Although other indicators and coding strategies are possible the presence of metabolic syndrome was selected as the criterion against which standards for body composition and aerobic fitness were derived.

When the original FITNESSGRAM standards were developed, clinical data linking fitness with health were lacking. Data sets with clinical outcomes are now more common. The analyses presented in this supplement were conducted using data from the National Health and Nutrition Examination Survey (NHANES), a nationally representative sample with assessments of body composition, fitness, and risk factors. A strength of the analyses presented herein is the use of the same nationally representative data set to derive the new body composition and aerobic fitness standards.

Another unique aspect of the standards is that they were developed using a well-refined empirical methodology that included the use of LMS (L=skewness, M=median, and S=coefficient of variation) curves and re-

ceiver operating characteristic (ROC) curves. The use of LMS curves provides a way to account for growth and maturation. In this supplement, the article by Laurson et al.[18] established growth centiles for body fatness, and the article by Eisenmann and colleagues[21] establishes growth centiles for cardiovascular fitness. The LMS parameters provided a way to standardize fitness levels across the developmental transition from ages 12 to 18 years. The ROC methodology was used to determine levels of fitness that are indicative of increased risk of metabolic syndrome in this age group. The paper by Laurson et al.[19]reports on the process used to establish standards for body fatness, and the paper by Welk et al.[22]provides a similar report on the development of standards for aerobic fitness. These methods provide an empirically sound approach for establishing health-related standards since the sensitivity and specificity of the standards can be directly examined.

As described in the previous text, it is also important to ensure good classification agreement for alternative field-based assessments. Body composition can be assessed using estimates of body fatness or BMI, but it is important for youth to be classified similarly on both assessments. The paper by Laurson et al.[20] describes the use of ROC curves to create BMI standards that correspond to the standards established for body fatness. Although these BMI standards are not intended to match the existing percentile standards used by the CDC, they are not dissimilar. The empirical linkage to body fatness is methodologically more defensible than percentile standards, given the inherent limitations of BMI.

The two primary assessments of aerobic fitness are the PACER test and the mile run, and several approaches were considered for improving classification agreement between these assessments. The paper by Mahar et al.[23] reports on new PACER prediction equations that were developed with larger samples and more robust methods than the prediction equation previously used in the program.[30] A consideration in the process was the inclusion of a BMI term to improve predictive accuracy. The Cureton equation[31] used to process the mile run includes a BMI term to improve the predictive accuracy of the test, and Mahar et al.[23] demonstrated that the inclusion of BMI also improves accuracy with the PACER. Another approach that was considered to improve classification agreement was the test-equating methodology recently proposed by Zhu et al.[32] to link PACER scores to estimated mile-run time. The paper in the supplement by Boiarskaia et al.[24] directly compared the test-equating methodology with the various PACER equations. The paper provided good support for the test-equating approach, and this was ultimately selected for use in the FITNESSGRAM program to improve classification agreement. There are many issues involved in the refinement of the standards, but the broader use of FITNESSGRAM in school-, district-, and state-level reporting placed a premium on ensuring good classification agreement.

The various papers in the supplement each contributed unique information, and collectively, this information provides empirical justification for the new FITNESSGRAM standards.[33] The new standards provide a clear resolution to some lingering inconsistencies in past standards while providing a stronger scientific basis for evaluating health-related fitness in youth. For example, the past body composition standards were static; that is, the same gender-specific cut off was used across ages 6–18 years. The use of LMS curves enabled the creation of standards that reflect gender differences as well as normal changes in growth and maturation.

A unique aspect of the new ROC-derived standards is that two separate thresholds were established. This created three distinct zones, a new HFZ, and two different NIZs (one labeled "some risk" and one labeled "higher risk"). The past dichotomous categorization (HFZ and NIZ) was limited, since there is not much difference among youth who happen to have scores that lie close to the standard on either side of the threshold. The use of three zones enables more effective and prescriptive messages to youth and their parents since the zones are based on clear differences in sensitivity and specificity. Children in the NIZ– higher risk receive messages warning them of potential risk if they continue tracking at that level. This is defensible since the strong specificity values reduce the risk of misclassifying students. Children in the HFZ would receive messages indicating that they likely have sufficient fitness for health, and this is justified by the high sensitivity for this cut point. Children in the NIZ–some risk receive a message that they are close to the higher-risk zone and that they should strive to move into the HFZ.

The transition to new standards has implications for schools, teachers, parents, and children, as well as public health and pediatric researchers. It is important to fully understand the effects of the new standards on fitness classification and on classification agreement. The final paper in the supplement[25] used data from a large number of students in elementary, middle, and high school to directly compare the old standards to the newly developed ones. This concluding paper provides a way to examine the impact of the changes for school-based fitness reporting.

Setting CR standards for fitness is an extremely difficult task, and it is especially challenging in youth. This is because health risks are not easily detected in this age group and because of the inherent complexities resulting

from growth and maturation. It is truly a bit of science and a bit of art. The availability of national data and the increasing sophistication of new analytic techniques warranted a systematic evaluation of the previous FITNESSGRAM standards and the development of these more refined ones. The articles in this supplement provide documentation of the steps taken and the resulting decisions about the new FITNESSGRAM health-related, CR standards for body composition and aerobic fitness.

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. Lee SM, Burgeson CR, Fulton JE, Spain CG. Physical education and physical activity: results from the School Health Policies and Programs Study 2006. J School Health 2007;77(8):435–63.
2. Morrow JR Jr., Zhu W, Franks BD, Meredith MD, Spain C. 1958–2008: 50 years of youth fitness tests in the United States. Res Q for Exerc Sport 2009;80:1–11.
3. Plowman SA, Sterling CL, Corbin CB, Meredith MD, Welk GJ, Morrow JR Jr. The history of FITNESSGRAM®. J Phys Act Health 2006;3(S2):S5–20.
4. Meredith M, Welk GJ, eds. FITNESSGRAM® ACTIVITYGRAM: test administration manual. 4th ed. Developed by the Cooper Institute (Dallas TX). Champaign IL: Human Kinetics, 2007.
5. Ernst MP, Corbin CB, Beighle A, Pangrazi RP. Appropriate and inappropriate uses of FITNESSGRAM®: a commentary. J Phys Act Health 2006;3(S):S90–100.
6. Liu Y. Youth fitness testing: if the "horse" is not dead, what should we do? Meas Phys Educ Exerc Sci 2008;12:123–5.
7. Mahar MT, Rowe DA. Practical guidelines for valid and reliable youth fitness testing. Meas Phys Educ Exerc Sci 2008;12:126–45.
8. Silverman S, Keating XD, Phillips SR. A lasting impression: a pedagogical perspective on youth fitness testing. Meas Phys Educ Exerc Sci 2008;12:146–66.
9. Wiersma LD, Sherman, CP. The responsible use of youth fitness testing to enhance student motivation, enjoyment, and performance. Meas Phys Educ Exerc Sci 2008;12:167–83, 2008.
10. Welk GJ. The role of physical activity assessments for school-based physical activity promotion. Meas Phys Educ Exerc Sci 2008; 12:184–206.
11. Morrow JR Jr., Ede A. Statewide physical fitness testing: a BIG waist or a BIG waste? Res Q Exerc Sport 2009;80:696–701.
12. Cooper KH. Reflections on the Texas youth evaluation project and implications for the future. Res Q for Exerc Sport 2010;81(S3):S79–83.
13. Welk GJ, Meredith MD, Lhmels M, Seeger C. Distribution of health-related physical fitness in Texas youth: a demographic and geographic analysis. Res Q for Exerc Sport 2010;81(3):S6–15.
14. Welk GJ, Morrow JR Jr. FITNESSGRAM® reference guide. Dallas TX: The Cooper Institute, 2008.
15. 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.
16. Zhu W, Mahar MT, Welk GJ, Going SB, Cureton KJ. Approaches for development of criterion-referenced standards in health-related youth fitness tests. Am J Prev Med 2011;41(4S2):S68–76.
17. 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.
18. 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.
19. Laurson KR, Eisenmann JC, Welk GJ. Development of youth percent body fat standards using receiver operating characteristic curves. Am J Prev Med 2011;41(4S2):S93–9.
20. Laurson KR, Eisenmann JC, Welk GJ. Body mass index standards based on agreement with health-related body fat. Am J Prev Med 2011;41(4S2):S100–5.
21. Eisenmann JC, Laurson KR, Welk GJ. Aerobic fitness percentiles for U.S. adolescents. Am J Prev Med 2011;41(4S2):S106–10.
22. 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.
23. Mahar MT, Guerieri AM, Hanna MS, Kemble CD. Estimation of aerobic fitness from 20-m multistage shuttle run test performance. Am J Prev Med 2011;41(4S2):S117–23.
24. 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.
25. Welk GJ, De Saint-Maurice Maduro PF, Laurson KR, Brown DD. Field evaluation of the new FITNESSGRAM® criterion-referenced standards. Am J Prev Med 2011;41(4S2):S131–42.
26. Jackson SA. The evolution and validity of health-related fitness. Quest 2006;58:160–75.
27. Kuczmarski RJ, Ogden CL, Guo SS, et al. 2000 CDC growth charts for the U.S.: methods and development. Vital Health Stat 11 2002;(246): 1–190.
28. Cureton KJ, Warren GL. Criterion-referenced standards for youth health-related fitness tests: a tutorial. Res Q Exerc Sport 1990; 61(1):7–19.
29. Jolliffe CJ, Janssen I. Development of age-specific adolescent metabolic syndrome criteria that are linked to the Adult Treatment Panel III and International Diabetes Federation criteria. J Am Coll Cardiol 2007;49:891–8.
30. Leger LA, Mercier D, Gadoury C, Lambert J. The multistage 20 metre shuttle run test for aerobic fitness. J Sports Sci 1988;6(2):93–101.
31. Cureton KJ, Sloniger MA, O'Bannon JP, Black DM, McCormack WP. A generalized equation for prediction of VO2peak for 1-mile run/walk performance. Med Sci Sports Exerc 1995;27(3):445–51.
32. Zhu W, Plowman SA, Park Y. A primary field test centered equating method for cut-off score setting. Res Q Exerc Sport 2010;81(4):400–9.
33. Meredith M, Welk GJ, eds. FITNESSGRAM-ACTIVITYGRAM: test administration manual. Updated 4th ed. Developed by the Cooper Institute (Dallas TX). Champaign IL: Human Kinetics, 2010.