Research Quarterly for Exercise and Sport, *84*, 456–468, 2013

ISSN 0270-1367 print/ISSN 2168-3824 online
DOI: 10.1080/02701367.2013.845517

Routledge
Taylor & Francis Group

SPECIAL TOPICS: *Research Quarterly for Exercise and Sport* Lecture

Science and Art of Setting Performance Standards and Cutoff Scores in Kinesiology

Weimo Zhu
University of Illinois at Urbana-Champaign

Setting standards and cutoff scores is essential to any measurement and evaluation practice. Two evaluation frameworks, norm-referenced (NR) and criterion-referenced (CR), have often been used for setting standards. Although setting fitness standards based on the NR evaluation is relatively easy as long as a nationally representative sample can be obtained and regularly updated, it has several limitations—namely, time dependency, population dependence, discouraging low-level performers, and favoring advantaged or punishing disadvantaged individuals. Fortunately, these limitations can be significantly eliminated by employing the CR evaluation, which was introduced to kinesiology by Safrit and colleagues in the 1980s and has been successfully applied to some practical problems (e.g., set health-related fitness standards for FITNESSGRAM®). Yet, the CR evaluation has its own challenges, e.g., selecting an appropriate measure for a criterion behavior, when the expected relationship between the criterion behavior and a predictive measure is not clear, and when standards are not consistent among multiple field measures. Some of these challenges can be addressed by employing the latest statistical methods (e.g., test equating). This article provides a comprehensive review of the science and art of setting standards and cutoff scores in kinesiology. After a brief historical overview of the standard-setting practice in kinesiology is presented, a case analysis of a successful CR evaluation, along with related challenges, is described. Lessons learned from past and current practice as well as how to develop a defendable standard are described. Finally, future research needs and directions are outlined.

Keywords: classification, criterion-referenced, decision making, evaluation, testing

"Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted," is a quotation often attributed to Albert Einstein (Calaprice, 2010, p. 482). We face these kinds of "counted" versus "cannot be counted" questions all the time in our daily kinesiology practice. For example, how many sit-ups should a 9-year-old be able to do to be called "fit"? How many minutes of moderate-to-vigorous physical activity (MVPA) should an adult do daily to be qualified as "active" enough to maintain good health? At what point on a teacher effectiveness scale we can start to call a teacher "effective"?

To make a number accountable, a decision-making system that is scientifically sound must be available. Specifically, to be able to correctly address these questions, a test/scale that measures the constructs of the interest

Correspondence should be addressed to Weimo Zhu, Department of Kinesiology & Community Health, University of Illinois at Urbana-Champaign, 205 Freer Hall, MC-052, Urbana, IL 61801. E-mail: weimozhu@illinois.edu

(e.g., physical fitness, physical activity, and teaching effectiveness) with supporting validity and reliability evidences must be developed and a link between a set of cutoff scores, also known as cut scores, from a specific test and corresponding performance standards, or simply standards (e.g., fit vs. not fit; active vs. sedentary; effective vs. ineffective) must be established and verified. The process to develop such a decision-making system, establish the relationship between cutoff scores and performance standards, and provide related validity and reliability evidence is called "performance standard setting" or simply "standard setting."

This article provides a comprehensive review of the current practice of standard setting in kinesiology. After a brief review of standards, cutoff scores, and their relationship, commonly used testing theory frameworks, on which cutoff scores are set up, will be reviewed, and the major methods for standard setting as well as validation of the cutoff scores will be introduced. Thereafter, a historical review of standard-setting practice in the field of kinesiology will be provided and major challenges will be described. Using physical fitness testing as an example, the standard-setting practice and related challenges in kinesiology will be further illustrated. Finally, lessons learned will be summarized and future research directions will be outlined.

STANDARD VERSUS CUTOFF SCORES

Although there are various definitions and descriptions of "standard" and "cutoff/passing scores," Kane's (1994) definition seems most appropriate: "The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version of the desired level of competence" (p. 426). Kane later (2001) extended the description to:

> The performance standards provide qualitative descriptions of the intended distinctions between adjacent levels of performance (e.g., between acceptable performance and unacceptable performance). The cutscores are points on the score scale, with one cutscore associated with each performance standard. The cutscore provides an operation version of the corresponding performance standards. (p. 55)

The relationship between cutoff scores and a standard can be described as an inferential relationship (i.e., using a cutoff score to make a classification, interpretation, conclusion, or meaning about a test taker's underlying, unobserved level of knowledge, skill, or ability). Simply, cutoff scores create meaningful categories on a scale with original raw or scaled scores that distinguish between individuals who meet some performance standard and those who do not (Cizek & Bunch, 2007, p. 17). Thus, standard setting is to establish the inferential relationship between a cutoff score and standard and is a process "to evoke and synthesize reasoned human judgment in a rational and defensible way so as to create these categories and partition the score scale on which a real trait is measured into meaningful and useful intervals" (Cizek & Bunch, 2007, p. 18). Standard setting therefore is not to be considered as a process to discover a knowable or estimable parameter. This is because, most of the time, the standard-setting process itself does not seek to find some preexisting or "true" score that separates real unique categories on a continuous underlying trait or competency.

Note that, interestingly, through the inference or standard-setting process, scores from the original continuous scale are usually "reduced" onto a category scale with only two (e.g., "pass/fail") or a few (e.g., "below basic, basic, proficient, advanced") categories or intervals (see Figure 1). This practice fits standard judgment in decision making (e.g., although we can measure one's height using finer continuous units such as "centimeters" or "inch," we only judge height by a few categories, such as "short, average, tall") and reflects humans' limited ability to process information ("magical number seven, plus or minus two") recognized a long time ago by Miller (1956).

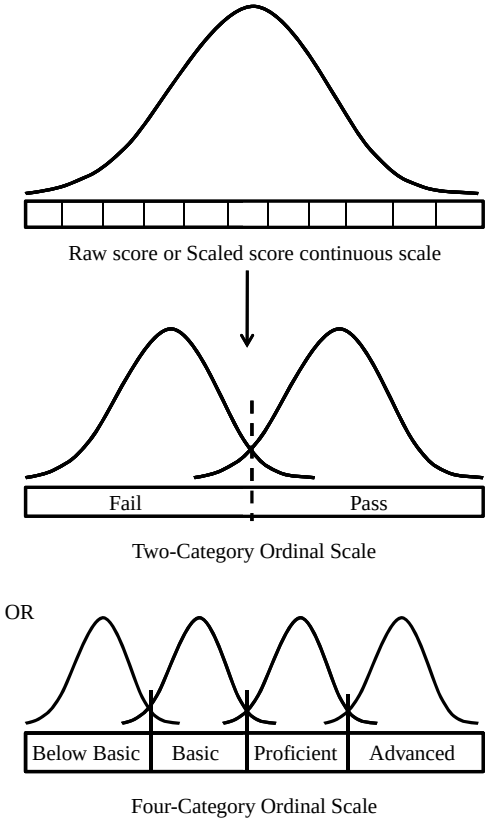

FIGURE 1 Illustration of standard setting.

THEORETICAL FRAMEWORKS FOR SETTING STANDARDS

Interpreting/evaluating the meaning of a test score, which is a piece of information about the construct measured, often depends on two theoretical frameworks—namely, norm-referenced (NR) and criterion-referenced (CR) evaluations. This is also true for standard setting.

NR Evaluation

Under the NR evaluation, a person's score is compared to their peers (e.g., same age and gender). In practice, the score is often compared to a previously developed norm, and either a specific percentile rank is derived or a specific ranking category related to others (e.g., "average, above average, or below average" or simply "pass/fail") is assigned for the score. Because the classification is based on one's performance relative to others, NR evaluation is a "relative" classification (see top portion of Figure 2). If the measurement interest is to rank or determine a person's relative position in the population, the NR evaluation is appropriate.

Technically, standard setting based on the NR evaluation is simple, and it can be established as long as a representative sample from the targeted reference population is available and the information is regularly updated. In practice, however, the NR evaluation has several challenges/limitations including: (a) It is difficult to update due to cost, time, and manpower constraints; (b) the referenced population has to be "normal" or "healthy"; otherwise, derived classifications become meaningless (e.g., a classification to "average" is meaningless if the entire population is unhealthy; see Zhu, 2012, for more details); and (c) the NR evaluation tends to reward the top group of test takers and discourage the lower group (Zhu, Mahar, et al., 2011). Fortunately, these limitations can be overcome by employing the CR evaluation.

CR Evaluation

The concept of CR evaluation and its related measurement practice, known as mastery testing, were introduced to education by Glaser in 1963, but real development and applications were not implemented until the late 1970s (see, e.g., Popham, 1978) and 1980s (see, e.g., Berk, 1980a, 1980b; Livingston & Zieky, 1982). In contrast to the NR framework, in which the evaluation of a test taker's competency is judged relative to other test takers' performance, the CR evaluation compares the test taker's

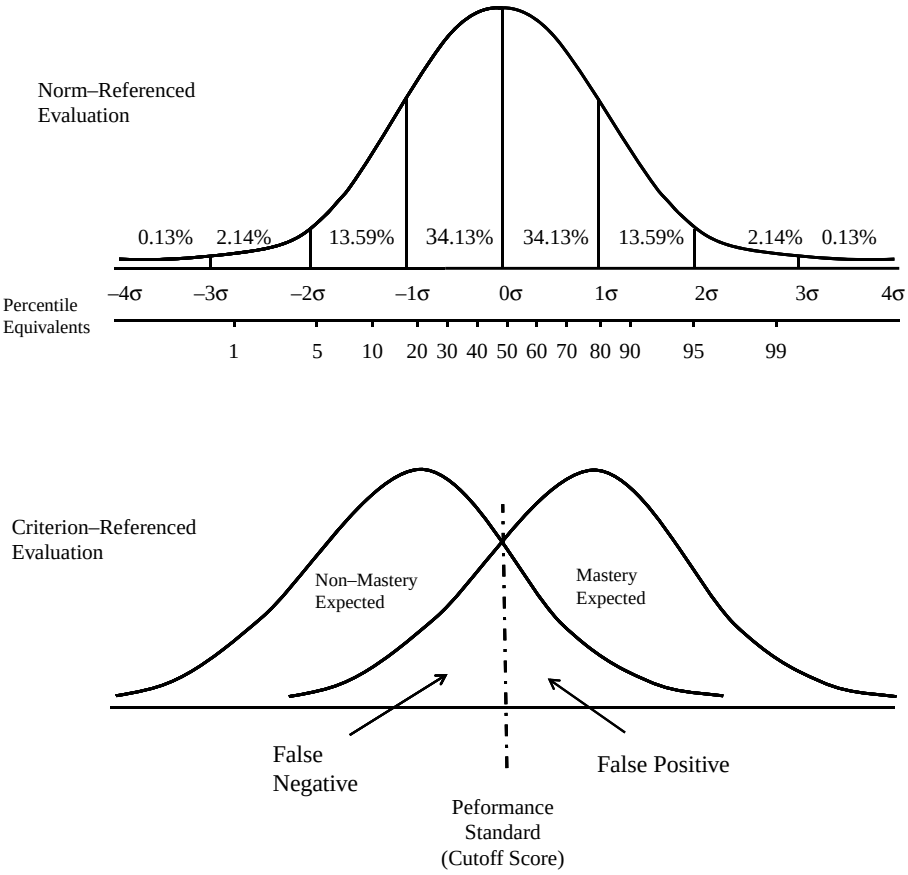

FIGURE 2 Nature of norm- and criterion-referenced evaluations.

performance with a predetermined absolute criterion or criterion behavior. As a result, the test taker is classified either as having mastery or nonmastery of the trait measured (see the bottom of Figure 2 for an illustration of the CR evaluation). In assessment practice, the "absolute criterion behavior" could be if a student has mastered the information taught in a specific subject or grade, or if a test taker is skilled enough to be certified as a personal trainer. Thus, the nature of the CR evaluation is "absolute" and will not be impacted by the performance or status of the test taker's peers. In addition, the standard set up for a CR evaluation focuses often on the minimal required competency of the trait measured. Because of the nature of the CR evaluation, the limitations "a – c" of the NR evaluation noted previously can be eliminated. See Zhu, Mahar, et al. (2011) for more details about the comparison of NR and CR evaluations.

Although CR evaluation has a number of measurement advantages over the NR evaluation, it has its own issues and challenges. Setting a valid "absolute" standard is perhaps the most challenging one. This is because the consequence of the decision based on such a standard could be very serious (e.g., if a test taker is qualified for a job, or if a child is fit). The standard set up has to be connected to the construct measured and support the classification made. In addition, the standard should be able to stand alone and be consistent with any individual from the targeted population (see more about the validity of the standard in a later section of this article).

It should be noted that depending on the measurement interest, a test taker's performance can be judged by either the NR or CR evaluation, or both. As an example, a student's sit-up test performance in the to-be-retired President's Challenge Physical Fitness Test was evaluated by norms while the sit-up test performance in FITNESSGRAM® is evaluated according to the "Fitness Zone," which is based on a CR evaluation. As an example of mixed usage, when a student takes the ACT to apply to colleges, the student will get a scaled score between 1 and 36, as well as the corresponding percentile rank, which is based on an NR evaluation; whether a student can be admitted to a specific college, however, is often dependent on the cutoff score set up by the college, which is an application of a CR evaluation. When a test is designed specifically for classification, it is called a "criterion-referenced test."

It should also be noted that when a test is calibrated using the item-response theory (IRT), a more advanced testing theory, cutoff scores are often set onto the scaled scores of a specific test. As a result, the cutoff scores set up are invariant to the test forms being used on a specific testing date. As an example, the Reading Skill Scale in the TOEFL iBT test (Test of English as a Foreign Language, Internet-based Test), an Internet-based English proficiency test that measures a test taker's ability to use and understand English at the university level, ranges from 0 to 30, with three proficiency levels: low (0–14), intermediate (15–21), and high (22–30). See Zhu (2006) for more about IRT-based test construction.

METHODS TO SET UP STANDARDS

Methods to set up an NR evaluation are rather straightforward. With a large, representative sample from the targeted population, one can develop norms by computing percentiles and percentile ranks (Safrit & Wood, 1995), which can be completed using any statistical software rather easily. Some smoothing techniques (e.g., LMS [skewness-median-coefficient of variation]; Cole & Green, 1992) are usually included in the computing to create smoothed percentile curves, which are more sensitive to changes by age.

Methods to set a CR evaluation, due to its "absolute" nature, are much more complex and have been the focus of many studies. Related methods to set up a CR evaluation have been well covered in the literature (see, e.g., Cizek, 2001; Cizek & Bunch, 2007; Zieky, Perie, & Livingston, 2008). In general, these methods can be classified as either "test- or item-centered" or "examinee-centered." In the item-centered methods, a panel of experts is asked to examine each item on a competency test and set the cutoff score accordingly. For example, in the Angoff method, one of the most popular methods, the panel is asked to examine each item and estimate the proportion of the "minimally acceptable" test takers who would answer each item correctly. The sum of these proportions would then represent the minimally acceptable score. Most of the methods in this category are subjective, although some of them start to combine with "objective" information in the judgment process. For example, in the bookmark method, another item-centered popular approach, items in a test (or a subset of them) are ordered by difficulty first, and each panel then places a "bookmark" in the sequence at the location of the cutoff score (Mitzel, Lewis, Patz, & Green, 2001).

In the "examinee-centered" methods, the focus is on identifying examinees with/without defined "minimum competency," from which the cutoff score is established. Two commonly used procedures in this category are the contrasting-groups and the borderline-group procedures, which could be based on evaluating the relative distributions of a skilled and unskilled group on a specific test.

VALIDATION OF THE STANDARDS

Although the standard is critical in making a decision (e.g., if a student can graduate from high school or if a test taker is fit enough for a job) and a well-described theoretical framework for standard validation was presented a long

time ago (Kane, 1994, 2001), limited efforts have been made to validate the developed standards. This is likely due to the fact that validating a cutoff score is much more difficult than developing one. To fully understand this challenge, the nature of standard validation, errors associated with classifications, and consequences must be understood.

According to Kane (2001), standard setting involves two basic tasks: (a) define performance levels, and (b) estimate cutoff scores that correspond to each performance standard. To validate the standard is to provide evidence that these two tasks have been performed successfully. To support validity of standard setting, Kane (2001, p. 59) proposed to collect four kinds of validity evidence, including (a) the conceptual coherence of the standard-setting process (e.g., if the standard-setting method and related assessment procedure are consistent with the conception of achievement underlying the decision procedure), (b) procedural evidence for the descriptive and policy assumptions (e.g., if the standards were set up in a reasonable way by persons who are knowledgeable about the purpose of the standards and familiar with the standard-setting procedure), (c) internal consistency evidence (e.g., if the presumed relationship between a performance standard and (d) a cutoff score can be confirmed), and agreement with external criteria (e.g., if the decision made is consistent with other assessment-based decision procedures). In addition, the role of consequences in standard setting and associated arbitrariness in standards must be examined.

It should be noted that any classification, no matter how well the standards were set up, will lead to some errors or misclassifications, and this is also true for the classifications based on NR or CR evaluations. When an NR evaluation is used for just the classification of "ranking by percentage," the error (incorrect placement of a test taker within the population) should be minor as long as the norm is based on a current/updated representative sample. When an NR evaluation is used as an ordinal "competency" scale with only a few categories based on outdated norms, the related errors could be substantial. First, the "rank percentage" label is inaccurate and therefore meaningless; second, the nature of NR evaluation is lost when standards are based on an outdated norm and the standards start to play an "absolute" CR classification role; finally, many of the existing ordinal "competency" scales are based on an arbitrary selection of a percentile point and have not been validated. The current Centers for Disease Control and Prevention (CDC) body mass index (BMI, obesity) chart, which is often used in practice for childhood and adolescence's obesity classification, is a good example of problematically using the NR evaluation for classification. According to the 2000 CDC Growth Charts (Kuczmarski et al., 2002), a child will be classified as "overweight" if the child's BMI for age $\geq$ 95th percentile, "at risk of overweight" if $\geq$ the 85th percentile but $<$ 95th percentile, "normal weight" if $<$ 85th but $>$ 5th percentile, and "underweight" if $<$ 5th percentile. Because the percentiles were developed based on 1963–1995 data, the meaning of percentage is no longer accurate, and that is why an illogical conclusion like "17% at or above the 95th percentile" could be drawn (Zhu, 2012). In addition, these percentiles were likely conveniently selected, and true values for the correct classifications are likely different percentiles (Himes & Bouchard, 1989).

When a CR evaluation is used for a pass–fail classification, two kinds of errors could occur: "false positive," in which a noncompetent test taker is classified as a competent one, and "false negative," in which a competent test taker is classified as a noncompetent one. When the classification is multicategory, the misclassification could be more complex although the errors will likely occur between adjacent categories. Because classification errors are unavoidable, the consequences of misclassification have to be considered when setting standards (e.g., to purposely set the standard higher or lower to avoid a particular error). Usually, the error of "false positive" (e.g., issuing personal trainer certification to an individual who is not qualified) is considered the more serious error and the test developer may purposely set the "bar" higher to avoid such an error. The tradeoff of this practice, however, will likely increase the "false-negative" error.

STANDARD-SETTING PRACTICE IN KINESIOLOGY

A Historical Overview

Like the long and rich history of testing and measurement in kinesiology (Safrit, 1989), the practice of standard setting in kinesiology can also be traced back to the early days of the field. Topics in setting standards were covered in all major early measurement textbooks. The following quotes are good examples of the early interest:

> A norm is valuable only in interpreting test scores, and necessity for a norm, or a given type of norm, is determined by the information which is desired regarding the ability of an individual or a group. To know that John can jump three feet nine inches means little without interpretation. But it is not likely that John's jump will be measured unless the records made by his classmates also are measured. With those, the average jump for John's grade or age is found to be three feet six inches and John's jump can now be interpreted as "better than average." If John's jump is to be interpreted with reference to the city, the state, or the nation, there must be norms for the children of those areas .... the more varied the available norms, the more extensive may be the interpretation of any record. (Glassow & Broer, 1938, pp. 53–54)

> Pass or fail. In this type of scoring, a standard is set. The individual is recorded as having achieved this standard or as not having achieved this standard. (McCloy, 1939, p. 89)

In fact, many tests with performance standards were developed in those days. The Universal Test for Strength, Speed, and Endurance by Sargent (1902), Physical Capacity Tests by Rogers (1931), Achievement Scales in Physical Education Activities by Neilson and Cozens (1934) and by Cozens, Cubberley, and Neilson (1937), and National Physical Achievement Standards for Girls by Howland (1936) are just a few examples. All of the standards during this period were based on the NR evaluation framework. In fact, the NR evaluation was the only method for standard setting until recently, and some well-known national and international examples with NR evaluation include the to-be-retired President's Challenge Physical Fitness Test (President's Council on Fitness, Sports & Nutrition, 2013), YMCA Fitness Testing and Assessment (Golding, 2000), Test of Gross Motor Development (TGMD; Ulrich, 2000), EURFIT (1988), and American Alliance for Health, Physical Education, Recreation and Dance developed sport skill test batteries (e.g., Hopkins, Shick, & Plack, 1984; see also Burton & Miller, 1998; Hoffman, 2006; Kirby, 1991; Ostrow, 2002, for a set of collections of tests and norms in kinesiology). Except for a few tests (e.g., TGMD, which was/is being updated), most of the norms or standards in these tests are outdated; therefore, their percentage order meaning is based on something that no longer exists, and at the very least, the measures (e.g., percentiles and percentile ranks generated) are not accurate.

In contrast with the long and rich history of the NR evaluation, the CR evaluation was not introduced to the field of kinesiology until much later by Safrit and others' pilot works (Looney, 1989; Safrit, 1981, 1989; Safrit, Baumgartner, Jackson, & Stamm, 1980). After these preliminary works, much CR evaluation-related research has been reported (e.g., Kalohn, Wagoner, Gao, Safrit, & Getchell, 1992; Rutherford & Corbin, 1994), and large standardized tests started to adopt the CR evaluation for their standard setting. Some well-known examples of CR evaluation tests include Fitnessgram (Plowman et al., 2006), a health-related youth fitness test, the Brockport Physical Fitness Test (Winnick & Short, 1999), a health-related test for youths with physical and mental disabilities, and the Senior Fitness Test (Rikli & Jones, 2013), which, in fact, contains both NR and CR evaluations.

PE Metrics is another example in physical education of a test that used CR evaluation. Designed as an assessment bank for the *National Standards for Physical Education* developed by the National Association for Sport and Physical Education (NASPE, 1995, 2004), PE Metrics employed a unique item-centered method to set cutoff scores. Rather than asking the experts to make a judgment on items after they were developed (in some cases, the data were collected), the experts were asked to develop the scoring rubric at the beginning of the item development. For each item, a five-level (0–4) scoring rubric was constructed, and Level 3 was defined as the competency level. Experts were asked to define criterion of each level with a targeted grade level in mind and then modified them if necessary after the pilot data were collected and analyzed. As an example, the task "dribble with hand and jog" was developed as an assessment in "Standard 1" for Grade 2 students' competency in motor skills, and a scoring rubric was developed to rate students' performance in three aspects, including "form," "space and distance," and "ball control." The criterion for the "form" competency (Level 3) is defined as (NASPE, 2010):

> Dribbles with selected essential elements:
> a) pushing action of finger pads
> b) ball at approx. waist height
> c) ball in front of body and to the 'dribble hand' side of the midline. (p. 67)

More technical details of PE Metrics and related standard setting can be found in a special section published in *Measurement in Physical Education and Exercise Science* (Dyson et al., 2011; Fox et al., 2011; Zhu, Fox, et al., 2011; Zhu, Rink, et al., 2011). Two additionally important, yet often overlooked, areas that actively involve CR evaluation-based decisions are pre-employment physical testing (Jackson, 2006) and diagnostic classification in clinical settings (Looney, 2006). Many professionals, such as firefighters, police officers, and military personnel, are required to participate in some sort of pre-employment tests. In addition to meeting the usual requirements of psychometric quality, developing defendable performance standards and related cutoff scores are perhaps the most important, as well as challenging, part of the test construction of a pre-employment test. Issues such as legal considerations, federal employment laws (e.g., discrimination litigation, disparate impact, and business necessity), and test fairness have to be carefully addressed in test development. Similar "seriousness" can be found in test development for the clinical setting where the consequence of a diagnostic test could be life and death. Along the decision theory, statistical procedures, such as the receiver-operating characteristic (ROC) curve, have long been employed in standard setting in both pre-employment and clinical diagnostic testing. Interested readers are referred to excellent introductions to the standard settings in these areas by Constable and Palmer (2000), Jackson (2006), and Looney (2006).

Finally, another area actively involved in cutoff score setting is the assessment of physical activity using wearable monitors. Specifically, the interest is to quantify MVPA using a wearable monitor, such as accelerometers or

pedometers. In fact, a group of cutoff points has been developed. Unfortunately, few of these studies were conducted under a carefully thought-out CR standard-setting framework. As a result, there has been a great inconsistency among the cutoff points reported. For example, based on a systematic review of cutoff points set for youth using ActiGraph accelerometers, Kim, Beets, and Welk (2012) reported that "no cut-points accurately classified moderate-to-vigorous physical activity (MVPA) across all ranges of physical activity intensity levels in comparison to a criterion measure" (p. 311). Rather than using the traditional single-regression approach to keep getting various cutoff scores, the new recommendation in this area is to apply "pattern recognition" approaches to train the algorithms on various activities to get better estimation of energy expenditure (Bassett, Rowlands, & Trost, 2012).

Although more than 30 years have passed since Safrit et al. (1980) introduced CR evaluation to the field, standard-setting practice based on the CR evaluation framework, except for a few examples mentioned earlier, is rather limited. There are several reasons for this. First, physical education is often treated as an unimportant subject area and most tests in kinesiology are used for making decisions that have little impact concerning education or polices. As a result, classifications made by a physical education-related test are often ignored by the general public, as well as students and their parents. Second, many test developers were not well trained in standard setting. As a result, the procedures for standard setting were not well thought-out and developed. Third, setting and validating a defendable standard take tremendous efforts and investment, including both research and practical training and implementation. As a result, many tests were developed as a "one-shot" effort, which could not meet the vigorous merits needed for standard setting. Fourth, modern standard setting is rather complex and often sophisticated in terms of methodological requirement (e.g., demanding on the measurement, statistical expertise, and computing skills). An interdisciplinary team with content and measurement/statistical expertise is often essential. To fully understand constraints and challenges related to CR standard setting, a careful look at the standard setting involved in Fitnessgram may be helpful.

Setting Standards for Fitnessgram

Initially designed by Charles L. Sterling as a physical fitness "report card" in 1977, it was renamed Fitnessgram after it was adopted by the Cooper Institute in 1982 as a standardized health-related fitness test (see Plowman et al., 2006, for a historical review of Fitnessgram). Very recently, Fitnessgram became the U.S. "national" fitness educational assessment and reporting software program.[1] There were several unique features at the start of the Fitnessgram development, which made it, including its standard setting, successful. First, it was designed specifically for "health-related physical fitness" to meet societal and public health needs (Jackson, 2006). Second, it selected an appropriate measurement model (i.e., CR measurement) as its foundation for the test construction, development, and score interpretation. Third, a unique concept of the "Health Fitness Zone" (HFZ) was created for Fitnessgram, which in nature is a multiple-category ordinal "fitness-proficient" scale. Rather than using a single cutoff score, a range of cutoff scores was created (e.g., the HFZ for a 5-year-old boy's percent body fat measured by skinfold measurements or bioelectric impedance analyzer ranged "from 18.8 and 8.9" [Cooper Institute, 2013]). In addition, based on the nature of the construct being measured, the number of categories and corresponding category labels varied. For aerobic capacity, there are three categories (HFZ, Need Improvement [NI], and NI-Health Risk); an extra category (Very Lean) was added for body composition; and only two categories (HFZ and NI) were employed for other fitness components. Fourth, an individualized computerized physical fitness "report card" with exercise prescription information was developed. Fifth, a group of excellent scientists with interdisciplinary expertise was recruited to serve on the scientific advisory board of Fitnessgram to guide its test development and ongoing revisions. Finally, a continuous and systematic effort was made to keep improving the test, including its standard setting.

While setting and validating HFZ can be considered the "best practice" of standard setting in kinesiology, it is not without some significant challenges. First, to develop a health-related fitness CR test, its connection with health must be established. Specifically, fitness and health components and measures must be operationally defined and selected. As illustrated in Figure 3, there are many fitness/health components, as well as many related measures to each component, that one can select. This part of the decision is usually completed based on a comprehensive literature review and the consensus of a panel of experts.

Second, the relationship between health and fitness may not be clear in children even when the relationship has been well established in the adult population. This is because health-deficient or disease symptoms may take years to develop. When Cureton was asked to be in charge of developing cutoff scores for aerobic capacity in 1994, he decided to choose morbidity and mortality as the health outcome because their relationship with aerobic capacity has been well supported by research literature for adults (Cureton, 1994). Because morbidity (caused mainly by unwanted pregnancy, substance abuse, physical/sexual abuse, stress, etc.) and mortality (caused mainly by accidents, suicide, and homicide) in children and youth are not directly related to physical fitness, cutoff scores

[1] After the 2012–2013 school year, the NR-based President's Challenge Physical Fitness Test will be replaced by the Presidential Youth Fitness Program (i.e., Fitnessgram)

cannot be directly related to children's morbidity and mortality data. Instead, Cureton derived the cutoff scores based on the information of both adult morbidity and mortality and age/growth-related changes in maximal oxygen consumption (VO_2max; Cureton, 1994; Cureton & Warren, 1990). Incredibly, the derived cutoff scores were well supported later by the cross-validation studies by Welk, Saint-Maurice, Laurson, and Brown (2011), which employed children's data and used different health outcome measures. Cureton's theoretical driven method in determining cutoff scores, which is named the "relative position" method in a recent Institute of Medicine (IOM) report on youth fitness measures and health outcomes (IOM, 2012), is illustrated in Figure 4.

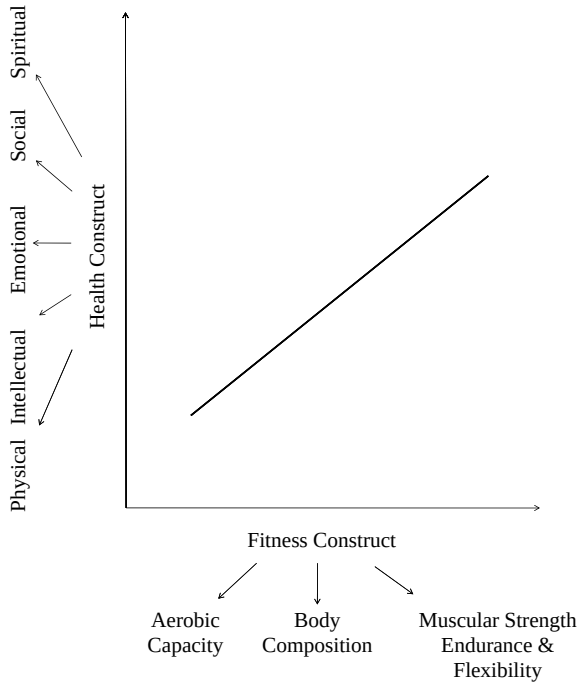

FIGURE 3 Fitness and health constructs and their components.

The third challenge in setting health-related fitness standards is that multiple cutoff scores are often needed. As illustrated in Part A in Figure 5, to determine the cutoff score of aerobic capacity, the cutoff scores of health outcome measures (e.g., metabolic syndrome) and fitness criterion measures (e.g., VO_2max) must be first determined, and an agreement between these two measures must be established. Then, a field test's cutoff score (e.g., 1-mile run/walk score [1,609.344 m]) must also be determined and its classification agreement with the criterion measure should be confirmed (Part B in Figure 5).

The fourth challenge is related to inconsistent classifications among field tests. In practice, more than one field test is often used to measure the same construct. As a result, cutoff scores for each individual test were set up. Due to variability of the tests and study samples, cutoff scores set up varied from each other, and as a result, passing ratings among different field tests are often inconsistent. For

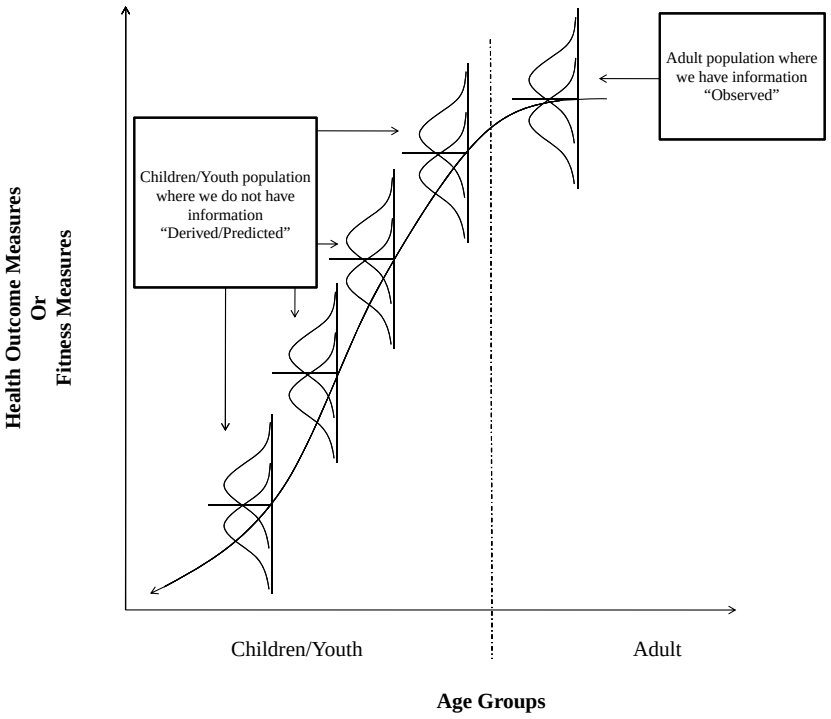

FIGURE 4 Theoretically driven, or relative position, method in deriving standards.

example, both the 1-mile run/walk and Progressive Aerobic Cardiorespiratory Endurance Run fitness test (PACER) are used in practice to measure aerobic capacity in Fitnessgram, but the standard equivalencies between them were rather poor. Mahar et al. (1997) reported that 34% of fourth- and fifth-grade girls who achieved PACER standards failed to pass the 1-mile run/walk standards (see also Beets & Pitetti, 2006). To address this problem, Zhu, Plowman, and Park (2010) proposed a primary test-centered equating method, in which PACER test scores were first transferred onto the scale of the 1-mile run/walk test and were then converted into 1-mile run/walk scores that were then used to predict VO_2max using the equation developed for the 1-mile run/walk (see Part C in Figure 5). The method was verified further by Welk et al. (2011) and was used in the setting of the current version of the HFZ (Cooper Institute, 2013). Together from A to C of Figure 5,[2] a conceptually sound and technically solid method called the "health-centered" standard-setting method was developed for setting standards for health-related fitness tests (Zhu, Mahar, et al., 2011).[3] With some modifications, it is expected that the method should be applicable to many other physical fitness and activity assessment tests when the interests are health-related.

The fifth challenge is related to the relationship between a fitness test and health outcomes. When the relationship can be confirmed in youth, the data-mining method (e.g., ROC curve) can be employed to determine the standards; when the relationship cannot be confirmed in youth, but in adults, the relative position method can be used to derive the standards, which have been supported by Cureton's work (1994); finally, when the relationship cannot be confirmed in both youth and adults and standards have to be set up, an alternative is to use the comparatively relative position method, in which a percentile determined from another fitness measure is borrowed. The validity of this last method, however, has not been verified. Interested readers may refer to a recent IOM (2012) report for more information about using these methods to set up standards.

Finally, the sixth challenge is related to a lack of the understanding of the consequences of misclassification. As mentioned earlier, there will always be misclassifications when an assessment serves a classification role, no matter how well the related cutoff score is set up. In the context of fitness testing, either false-positive classification (e.g., an unfit test taker misclassified as fit) or false-negative classification (a fit test taker misclassified as unfit) could happen. The question is which misclassification has a more serious consequence. In general, it is believed that the false-positive classification may be a more serious error in this case because the misclassified test takers may get the wrong impression that they are fit enough already and may therefore not exercise at a desirable level and consequently fail to reduce or even increase their risk for disease (Cureton & Warren, 1990). This, however, may not be true all the time. As an example, when an NR evaluation is employed, the standards set up usually discourage students whose fitness levels might be moderate or low because only a small percentage of students will be able to meet the standards under such an evaluation framework. In other words, the false-negative errors were likely to have occurred. This, in fact, has been verified. For example, when the President's Challenge Award was employed, only a very small proportion of test takers could get the award (i.e., scored in the 85th percentile or higher for all five tests; Corbin, Lovejoy, Steingard, & Emerson, 1990). In this case, the standards were likely set too high and the classification error of false-negative dominated. 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 focus on their level compared to others. As a result, students are able to enhance their motivation and self-confidence. However, because the CR-based standards often focus on minimal required fitness, one concern is that the performance standard set could be so low that it may fail to motivate the top fit students. Because most top fit students may already have successful experiences in and be motivated by participating in sport activities outside of the school physical education environment, the impact of the concern should be minimal. Meanwhile, the consequence of misclassification is clearly an understudied area (Corbin et al., 1990; Cureton & Warren, 1990).

Addressing these challenges, according to Kane's (2001) validation framework for setting standards, is in fact collecting validity evidences for the standards set up. For example, the development of the Fitnessgram technical manual and standard-setting methodology (Welk & Meredith, 2008; Zhu, Mahar, et al., 2011) can be considered evidence for "the conceptual coherence of the standard-setting procedure" described earlier; Cureton and Warren's (1990) work on how to set standards using the relative position method and Plowman et al.'s (2006) review of the history of Fitnessgram can be considered the "procedural evidence," all related works in Parts B and C of Figure 5 (e.g., Mahar et al., 1997; Zhu et al., 2010) can be considered the "internal consistency evidence," and all related works in Part A of Figure 5 can be considered the evidence of "agreement with external criteria" (e.g., the articles in the Fitnessgram supplement in the *American Journal of Preventive Medicine*, edited by Morrow, Going, and Welk

[2] Although it may not be the best practice to use an alternative field test to correlate with a health outcome measure, it was done sometimes in practice, as illustrated in Part D of Figure 5.

[3] Unfortunately, because BMI is required for the 1-mile run/walk prediction equation and many states or schools do not allow BMI measurements of their students, the standard set for the aerobic capacity based on this advanced method will be dropped from the next version of the HFZ (http://www.cde.ca.gov/ta/tg/pf/healthfitzones.asp).

[2011], on development of CR standards for aerobic capacity and body composition).

LESSONS LEARNED AND FUTURE DIRECTIONS

By reviewing the long history of the practice of standard setting in kinesiology and examining successful and failed examples, a number of lessons can be learned:

1. Successful standard setting depends on a combination of science and art. By combining experts' thoughtful recommendations and input with sophisticated scientific models/methods, challenges typically related to standard setting can be addressed and overcome.
2. An interdisciplinary team with content and measurement expertise is essential for successful standard setting.
3. The purpose of the test/assessment must be clearly defined, and the corresponding evaluation framework—NR, CR, or both—should then be selected accordingly.
4. A carefully designed standard validation plan should be drafted before setting standards and cutoff scores so that related evidence can be collected purposely.
5. "Rome was not built in a day," and developing a defendable standard/cutoff score takes time and systematic efforts. The collection of evidence to support a standard/cutoff scores is needed when standard setting.
6. Interpretation of a classification must be based on the nature of the evaluation framework upon which a classification is derived. For example, an outdated norm cannot be interpreted on its relative rank in percent meaning.
7. Validity, reliability, and other psychometric issues should be taken into consideration when setting standards.
8. Classification error is unavoidable, but the impact of the misclassification can be controlled and reduced. To do so, understanding the consequences of a misclassification is essential.
9. A pilot study should be conducted when setting up a new standard or modifying an old one.
10. There is an urgent need to improve our graduate students' training in how to set standards appropriately.

Regarding the research needs and direction, the study needed most is to understand the consequence of misclassification, which has been a long-understudied issue in standard setting. Considering Fitnessgram has become a national assessment and award program, to fully understand the impact of classification on children's physical activity behavior, which in turn impacts their fitness, becomes extremely important. Corbin et al. (1990) constructed a well thought-out theoretical foundation on studying the impact of the fitness

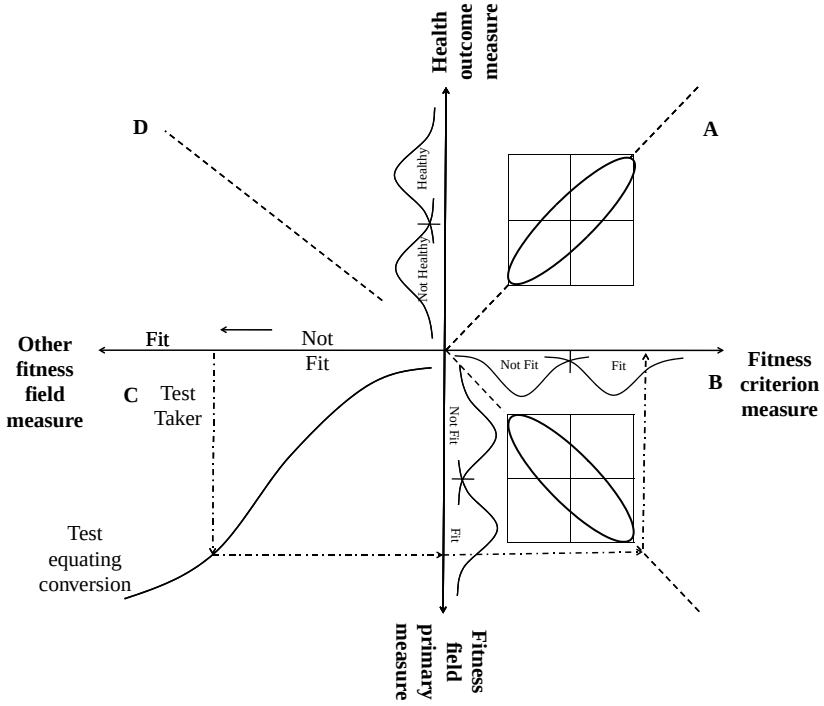

FIGURE 5 Health-centered method for standard setting.

awards, which, even after 25 years, can still serve as a useful guideline for design-related studies.

One limitation of both NR and CR evaluations is that neither take a test taker's initial position or change into consideration in the evaluation or classification. Fortunately, a new evaluation method called "student growth percentiles" (SGP) could help address this gap. Developed by Betebenner (2008, 2009), SGP belongs to a "student growth" methodology with a focus on how much a student has improved or grown from 1 year to the next as compared with the student's academic peers with similar starting scores or performances. SGP thus is a relative measure that focuses on the rate of change in comparison to a student's academic peers. The rate of change is expressed using a "percentile" that can range from 1 to 99. Lower numbers indicate lower growth/change and higher numbers show higher growth. The scores in the middle represent moderate growth. As a result, every student has an opportunity to demonstrate high or low growth or improvement. Thus, SGP can be conceptually considered as a local change norm. Because pretest and posttest scores in SGP are compared to corresponding absolute criterion and differences between pretest and posttest scores were evaluated based on a norm, SGP can also be considered a "mixed" evaluation approach that takes the advantages of NR and CR assessments and considers pretest and posttest change. With these unique features, SGP should address one of education's key interests: How much has a student learned from their starting point last year? In addition, SGP has been used as a "value-added" method to determine teacher effectiveness. SGP should have great potential for assessment and evaluation practice in physical education. For example, SGP could help answer key questions of parents and stakeholders: Did my child make a year's worth of progress? Is my child growing as much in aerobic fitness as in muscular strength? How close are my students to being "proficient?" Are our students making appropriate strides toward meeting state/national standards? There is an urgent need to study and apply SGP to the practice of measurement and evaluation in kinesiology.

Finally, kinesiology researchers should give a careful look at machine-learning/pattern recognition techniques and explore their application potential in measurement and evaluation practice. Machine learning is a branch of computer-based artificial intelligence, or the science of how to get computers to act without being explicitly programmed for that act (Hastie, Tibshirani, & Friedman, 2001). Although many are not aware, machine learning already has supplied self-driving cars, practical speech recognition software, effective Web searching, and a vastly improved understanding of the human genome. Although initial attempts to utilize machine learning have been made in physical activity research, the field of kinesiology has not yet taken full advantage of this powerful technique. This article therefore calls for more research studies to explore machine learning, especially for classification and evaluation research and practice.

CONCLUSION

Setting standards and cutoff scores, especially those based on the NR evaluation framework, has a long and rich history in the field of kinesiology. Yet, most of the norms used to develop standards and cutoff scores are outdated and can no longer serve their relative percentage order meaning. Although CR evaluation has several advantages over NR evaluation, it has only been used by a few large-scale test applications. In addition, few efforts have been made to systematically collect validation evidence for the standards or cutoff scores developed. Through the analysis of a successful CR evaluation application, challenges and needed solutions to develop defendable standards and cutoff scores have been described in detail. Finally, lessons learned from past and current setting of standard and cutoff scores have been described, and future research needs and directions have been outlined.

REFERENCES

Bassett, D. R. Jr., Rowlands, A., & Trost, S. G. (2012). Calibration and validation of wearable monitors. *Medicine and Science in Sports and Exercise*, *44*(Suppl. 1), S32–S38.

Beets, M. W., & Pitetti, K. H. (2006). Criterion-referenced reliability and equivalency between the PACER and 1-mile run/walk for high school students. *Journal of Physical Activity and Health*, *3*(Suppl. 2), 21–33.

Berk, R. A. (Ed.). (1980a). *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (Ed.). (1980b). *A guide to criterion-referenced test construction*. Baltimore, MD: Johns Hopkins University Press.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York, NY: Taylor & Francis.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*, 42–51.

Burton, A. W., & Miller, D. E. (1998). *Movement skill assessment*. Champaign, IL: Human Kinetics.

Calaprice, A. (Ed.). (2010). *The ultimate quotable Einstein*. Princeton, NJ: Princeton University Press.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting*. Thousand Oaks, CA: Sage.

Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, *11*, 1305–1319.

Constable, S. H., & Palmer, B. (Eds.). (2000). *The process of physical fitness standards development*. Dayton, OH: Human Systems Information Analysis Center.

Cooper Institute. (2013). *Healthy Fitness Zone Standards*. Dallas, TX: Author. Retrieved from http://www.cooperinstitute.org/healthyfitness zone

Corbin, C. B., Lovejoy, P. Y., Steingard, P., & Emerson, R. (1990). Fitness awards: Do they accomplish their intended objectives? *American Journal of Health Promotion*, *4*, 345–351.

Cozens, F. W., Cubberley, H. J., & Neilson, N. P. (1937). *Achievement scales in physical education activities for secondary school girls and college women*. New York, NY: A.S. Barnes.

Cureton, K. J. (1994). Aerobic capacity. In J. R. Morrow, Jr. , H. B. Falls, & H. W. Kohl, III (Eds.), *The Prudential FITNESSGRAM technical reference manual* (pp. 33–55). Dallas, TX: Cooper Institute for Aerobics Research.

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

Dyson, B., Placek, J. H., Graber, K. C., Fisette, J. L., Rink, J., Zhu, W., & … Park, Y. (2011). Development of PE metrics elementary assessments for National Physical Education Standard 1. *Measurement in Physical Education and Exercise Science*, *15*, 100–118.

EURFIT. (1988). *European tests of physical fitness*. Rome, Italy: Council of Europe, Committee for the Development of Sport.

Fox, C., Zhu, W., Park, Y., Fisette, J. L., Graber, K. C., Dyson, B., … Raynes, D. (2011). Related critical psychometric issues and their resolutions during development of PE Metrics. *Measurement in Physical Education and Exercise Science*, *15*, 138–154.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, *18*, 519–522. doi: 10.1037/h0049294

Glassow, R. B., & Broer, M. R. (1938). *Measuring achievement in physical education*. Philadelphia, PA: W.B. Saunders.

Golding, L. A. (Ed.). (2000). *YMCA fitness testing and assessment manual* (4th ed.). Champaign, IL: Human Kinetics.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Himes, J. H., & Bouchard, C. (1989). Validity of anthropometry in classifying youths as obese. *International Journal of Obesity*, *13*, 183–193.

Hoffman, J. (2006). *Norms for fitness performance and health*. Champaign, IL: Human Kinetics.

Hopkins, D. R., Shick, J., & Plack, J. J. (1984). *Basketball for boys and girls: Skills test manual*. Reston, VA: American Alliance for Health, Physical Education, Recreation and Dance.

Howland, A. R. (1936). *National physical achievement standards for girls*. New York, NY: National Recreation Association.

Institute of Medicine. (2012). *Fitness measures and health outcomes in youth*. Washington, DC: National Academics.

Jackson, A. S. (2006). The evolution and validity of health-related fitness. *Quest*, *58*, 160–175.

Kalohn, J. C., Wagoner, K., Gao, L. G., Safrit, M. J., & Getchell, N. (1992). A comparison of two criterion-referenced standard setting procedures for sports skills testing. *Research Quarterly for Exercise and Sport*, *63*, 1–10.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–461.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kim, Y., Beets, M. W., & Welk, G. J. (2012). Everything you wanted to know about selecting the "right" Actigraph accelerometer cut-points for youth, but …: A systematic review. *Journal of Science and Medicine in Sport*, *15*, 311–321.

Kirby, R. F. (Ed.). (1991). *Kirby's guide to fitness and motor performance tests*. Cape Girardeau, MO: BenOak Publishing.

Kuczmarski, R. J., Ogden, C. L., Guo, S. S., Grummer-Strawn, L. M., Flegal, K. M., Mei, Z., & … Johnson, C. L. (2002). 2000 CDC Growth Charts for the United States: Methods and development. *Vital and Health Statistics*, *11*, 246. Retrieved from http://www.cdc.gov/nchs/data/series/sr_11/sr11_246.pdf

Livingston, S. A., & Zieky, M. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Looney, M. A. (1989). Criterion-referenced measurement: Reliability. In M. J. Safrit & T. M. Woods (Eds.), *Measurement concepts in physical education and exercise science* (pp. 137–152). Champaign, IL: Human Kinetics.

Looney, M. A. (2006). Measurement issues in the clinical setting. In T. Wood & W. Zhu (Eds.), *Measurement theory and practice in kinesiology* (pp. 297–312). Champaign, IL: Human Kinetics.

Mahar, M. T., Rowe, D. A., Parker, C. R., Mahar, F. J., Dawson, D. M., & Holt, J. E. (1997). Criterion-referenced and norm-referenced agreement between the mile run/walk and PACER. *Measurement in Physical Education and Exercise Science*, *1*, 245–258.

McCloy, C. H. (Ed.). (1939). *Tests and measurements in health and physical education*. New York, NY: F.S. Crofts.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.

Morrow, Jr., J. R., Going, S. B., & Welk, G. J. (Eds.). (2011). FITNESSGRAM® development of criterion-referenced standards for aerobic capacity and body composition [Supplement]. *American Journal of Preventative Medicine*, *41*(4, Suppl. 2).

National Association for Sport and Physical Education. (1995). *Moving into the future: National standards for physical education: A guide to content and assessment*. Reston, VA: Author.

National Association for Sport and Physical Education. (2004). *Moving into the future: National standards for physical education* (2nd ed.). Reston, VA: Author.

National Association for Sport and Physical Education. (2010). *PE Metrics: Assessing National Standards 1–6 in elementary school*. Reston, VA: Author.

Neilson, N. P., & Cozens, F. W. (1934). *Achievement scales in physical education activities for boys and girls in elementary and junior high schools*. Sacramento, CA: California State Department of Education.

Ostrow, A. C. (Ed.). (2002). *Directory of psychological tests in the sport and exercise sciences* (2nd ed.). Morgantown, WV: Fitness Information Technology.

Plowman, S. A., Sterling, C. L., Corbin, C. B., Meredith, M. D., Welk, G. J., & Morrow, J. R. Jr. (2006). The history of FITNESSGRAM®. *Journal of Physical Activity & Health*, *3*(Suppl. 2), S5–S20.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

President's Council on Fitness, Sports & Nutrition (2013). *The President's Challenge*. Rockville, MD: Author. Retrieved from http://www.presidentschallenge.org/challenge/physical/benchmarks.shtml

Rikli, R., & Jones, C. J. (2013). *Senior Fitness Test manual* (2nd ed.). Champaign, IL: Human Kinetics.

Rogers, F. R. (1931). *Physical capacity tests*. New York, NY: A.S. Barnes.

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

Safrit, M. J. (1981). *Evaluation in physical education*. Englewood Cliffs, NJ: Prentice Hall.

Safrit, M. J. (1989). Criterion-referenced measurement: Validity. In M. J. Safrit & T. M. Wood (Eds.), *Measurement concepts in physical education and exercise science* (pp. 117–135). Champaign, IL: Human Kinetics.

Safrit, M. J., Baumgartner, T. A., Jackson, A. S., & Stamm, C. L. (1980). Issues in setting motor performance standards. *Quest*, *32*, 152–162.

Safrit, M. J., & Wood, T. M. (1995). *Introduction to measurement in physical education and exercise science* (3rd ed.). St. Louis, MO: Mosby.

Sargent, D. A. (1902). *Universal test for strength, speed, and endurance of the human body.* Cambridge, MA: Powell.

Ulrich, D. A. (2000). *Test of Gross Motor Development* (2nd ed.). Austin, TX: Pro-Ed.

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

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

Winnick, J. P., & Short, F. X. (1999). *The Brockport Physical Fitness Test manual*. Champaign, IL: Human Kinetics.

Zhu, W. (2006). Constructing tests using item response theory. In T. Wood & W. Zhu (Eds.), *Measurement theory and practice in kinesiology* (pp. 53–76). Champaign, IL: Human Kinetics.

Zhu, W. (2012). '17% at or above the 95th percentile'—What is wrong with this statement? *Journal of Sport and Health Science*, *1*, 67–69.

Zhu, W., Fox, C., Park, Y., Fisette, J. L., Dyson, B., Graber, K. C., … Raynes, D. (2011). Development and calibration of an item bank for PE Metrics assessments: Standard 1. *Measurement in Physical Education and Exercise Science*, *15*, 119–137.

Zhu, W., Mahar, M. T., Welk, G. J., Going, S. B., & Cureton, K. J. (2011). Approaches for development of criterion-referenced standards in health-related youth fitness tests. *American Journal of Preventive Medicine*, *41* (Suppl. 2), S68–S76.

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

Zhu, W., Rink, J., Placek, J. H., Graber, K. C., Fox, C., Fisette, J. L., … Raynes, D. (2011). Physical education metrics: Background, testing theory, and methods. *Measurement in Physical Education and Exercise Science*, *15*, 87–99.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.