Opinion

“17% at or above the 95th percentile” — What is wrong with this statement?

Weimo Zhu

Department of Kinesiology & Community Health, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received 2 July 2012; accepted 5 July 2012

When reporting the prevalence of childhood obesity in the USA a few years ago, the magazine U.S. News & World Report stated:

“...some 17 percent of kids are now obese, which means they’re at or above the 95th percentile for weight in relation to height for their age; an additional 17 percent are overweight, or at or over the 85th percentile.”¹

Anyone with some basic training in measurement or statistics will realize that this statement is incorrect. This is because the percentile is defined as the value below which a certain percent of observations fall in a population. For example, the 15th percentile is the value (or score) below which 15 percent of the observations in a population may be found. If the percentile value in the above statement is correct, 5%, rather than 17%, should be at or above the 95th percentile. Unfortunately, similar statements can be found everywhere in scientific literature, especially when describing the prevalence of childhood obesity using the growth chart developed by the U.S. Centers for Disease Control and Prevention (CDC).²³ How could this happen?

To fully understand what went to wrong in this statement and similar reporting practices, a quick review on commonly used evaluation frameworks should be helpful. After getting a value or score from a measurement scale, we can make a judgment of the value either by comparing it with the values of others or with an absolute standard. The former is known as the norm-referenced (NR) evaluation and the latter is called as the criterion-referenced (CR) evaluation. When employing the NR evaluation framework, a person’s performance is compared with his/her peers, often by gender and age. Therefore, the nature of the NR evaluation is “relative.” The Presidential Physical Fitness Award (PPFA) in the U.S. President’s Challenge program is a good example of an NR evaluation, in which students must score at or above the 85th percentile on all five fitness test items to qualify for the award. In contrast, when employing the CR evaluation framework, a person’s performance is compared with a predetermined value or standard known as the “criterion” or “criterion behavior” (e.g., if a student has mastered the skill taught in a specific sport or if a child meets a minimal needed physical activity level). The nature of the CR evaluation, therefore, is “absolute.” Determining if a person’s blood pressure is normal bused on his/her systolic and diastolic pressures is a good example of a CR evaluation.

When the measurement interest is on “the more (e.g., number of pull-ups a student can do), or less (e.g., how fast a student can finish a one-mile run/walk test), the better”, the NR evaluation is more appropriate. Constructing an NR evaluation is relatively easy as long as a large, current and representative sample of a population 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, several major limitations often associated with the NR evaluation framework.

First, it is difficult to update norms regularly due to cost, time, and manpower constraints. As an example, the PPFA’s norms were based on the 1985 National School Population Fitness Survey and there have been no major national fitness studies in the USA since the 1980s. As a result, these outdated values likely do not reflect current norms (e.g., an 80th percentile from the 1980s may now be equivalent to the 90th percentile), but rather how the present values compare to the previous norms, making them inaccurate in its original evaluation framework and the key “percentage” information no longer exists.

Second, the interpretation under the NR evaluation depends on the “normal” status of the reference population. The designations of “average” or “above average” have limited meaning if the majority of a population is not normal (e.g., obese, unfit or unhealthy).

Third, the selection of a percentile associated with health outcome measures (e.g., 85th or 95th percentiles as the cutoff values for “overweight” or “obese”) is often arbitrary with...
little scientific foundation. It is likely that other percentiles (say 83rd vs. 97th) may be the more appropriate values when connecting these cut-off values with outcome variables of interest (e.g., health outcomes such as metabolic syndrome).

Fourth, the employment of the NR evaluation framework 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 PPFA) 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 very low. If unfit students are less motivated during physical fitness testing they may come to perceive physical education classes, especially physical fitness testing, as a punitive, rather than enjoyable.

The problem of the “17% in the 95th percentile” statement noted earlier is a good example of the first three limitations of the NR evaluation. According to CDC’s current standard or growth chart, a child is defined as overweight if their body mass index (BMI) is at or above the age- and gender-specific 85th percentile, and obese if their BMI is at or above the 95th percentile of their peers. If this norm was 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 with a third (33%) of the U.S. children and adolescents identified as overweight and obese. The difference in prevalence estimate is explained by the fact that the CDC’s growth chart was derived from data collected in the 1970s and 1980s.4 Thus, about 12% (17 – 5 = 12) of children could be misclassified as not being obese if we use the 95th percentiles standards based on today’s norms of a relative unhealthy population (Fig. 1). Clearly, these outdated percentiles have lost their associations with the meaning of “percentages” and now function as cut-off scores with an “absolute” meaning under the CR framework.

Fortunately, the four major limitations related to NR evaluation can be eliminated by employing the CR evaluation framework, in which a person’s performance or status is compared with an absolute criterion. First, because the criterion is defined independently and not impacted by changes in a population, the limitation of “population dependence” in the NR evaluation is eliminated. Second, while there are always some test takers classified as below average, average, and above average in an NR evaluation, there is a possibility that all test takers could be classified as “pass” or “fail” based on a criterion (i.e., it is possible for everyone to either meet or not meet the CR standards, or be fit or not fit in the context of physical fitness testing). As a result, the limitation of “the population has to be normal” in the NR evaluation is eliminated. Third, setting a standard for a CR evaluation is either based on the contributions of a panel of experts or some correlation studies, hence the arbitrariness in standard setting is greatly reduced. Finally, since the focus in a CR evaluation is often on the “minimal competency”, the evaluation standard established is often attainable by any test takers as long as an effort is made. Thus, the limitation of discouraging “low-percentile” participants associated with the NR evaluation is minimized. Since it was introduced in 1980s,5–7 the CR evaluation has been employed in kinesiology for evaluation standard setting. Setting the standards for FITNESSGRAM®, a fitness testing and education program, is perhaps the best example of such an application (see a recent special issue of the American Journal of Preventive Medicine, Vol. 41(4, Suppl. 2), 2011 for more details8). Meanwhile, CR evaluation is not without its own challenges. Setting and validating an appropriate standard, known as the cut-off score, often takes years of research efforts and accumulations.

Several lessons can be learned from the incorrect usage of NR evaluation information:

1. To maintain the “percentage” meaning of a norm, it should be generated from a large, current, and representative sample and kept updated;
2. Whenever a norm is used, the time when the norm was developed must be reported;
3. When an “outdated” norm is used, the “percentage” meaning of the values in a norm no longer exists, therefore: (a) they should not be called “percentiles” and (b) they should be interpreted as the “absolute” standard except for when comparing cross-year percentage shifts/changes;
4. When a norm is used for classification, the cut-off percentiles should not be selected based on some conventional practice (e.g., using 85th and 95th percentiles); rather, selections should be based on the established relationship between percentile(s) and associated outcome measures.

In summary, the confusion in the “17% in the 95th percentile” statement is caused by employing an outdated norm, in which values no longer maintain their associations with percentages. Whenever a norm is used, the time when it was developed must be reported simultaneously. Since cut-off
percentiles in NR evaluation are often selected arbitrarily, they should not be directly used for classification before establishing their relationship with meaningful external outcome measures. While the NR framework has its role in the practice of evaluation, it has several known limitations. Users should be aware of these limitations and interpret the results with caution. Fortunately, these limitations can be eliminated or minimized by employing the CR evaluation framework. Setting and validating appropriate standards in the CR evaluation, however, take systematic efforts.

Acknowledgment

Some thoughts in this article were generated from a fitness testing section at the 2008 American Alliance for Health, Physical Education, Recreation and Dance (AAHPERD) national convention organized by Dr. James R. Morrow, Jr. and the American Journal of Preventive Medicine (Vol. 41(4, Suppl. 2), 2011) article I co-authored with Drs. Matthew T. Mahar, Gregory J. Welk, Scott B. Going, and Kirk J. Cureton.

References