HHS Public Access

Author manuscript

Med Sci Sports Exerc. Author manuscript; available in PMC 2019 April 01.

Published in final edited form as:

Med Sci Sports Exerc. 2018 April ; 50(4): 837–845. doi:10.1249/MSS.0000000000001481.

A Primer on the Use of Equivalence Testing for Evaluating Measurement Agreement

Philip M. Dixon[1], **Pedro F. Saint-Maurice**[2,6], **Youngwon Kim**[3], **Paul Hibbing**[4], **Yang Bai**[5], and **Gregory J. Welk**[2]

[1]Department of Statistics, Iowa State University, Ames, IA

[2]Department of Kinesiology, Iowa State University, Ames IA

[3]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, United Kingdom

[4]Department of Kinesiology, Recreation, and Sport Studies, University of Tennessee, Knoxville TN

[5]Department of Rehabilitation and Movement Science, University of Vermont, Burlington, VT

[6]Metabolic Epidemiology Branch, National Cancer Institute, NIH, Bethesda, MD

Abstract

Purpose—Statistical equivalence testing is more appropriate than conventional tests of difference to assess the validity of physical activity (PA) measures. This paper presents the underlying principles of equivalence testing and gives three examples from PA and fitness assessment research.

Methods—The three examples illustrate different uses of equivalence tests. Example 1 uses PA data to evaluate an activity monitor's equivalence to a known criterion. Example 2 illustrates the equivalence of two field-based measures of physical fitness with no known reference method. Example 3 uses regression to evaluate an activity monitor's equivalence across a suite of 23 activities.

Results—The examples illustrate the appropriate reporting and interpretation of results from equivalence tests. In the first example, the mean criterion measure is significantly within +/−15% of the mean PA monitor. The mean difference is 0.18 METs and the 90% confidence interval of [−0.15, 0.52] is inside the equivalence region of [−0.65, 0.65]. In the second example, we chose to define equivalence for these two measures as a ratio of mean values between 0.98 and 1.02. The estimated ratio of mean VO2 values is 0.99, which is significantly (p=0.007) inside the equivalence region. In the third example, the PA monitor is not equivalent to the criterion across the suite of activities. The estimated regression intercept and slope are −1.23 and 1.06. Neither confidence interval is within the suggested regression equivalence regions.

Corresponding Author: Name: Gregory J. Welk, Mailing Address: Department of Kinesiology, Iowa State University, 257 Forker Building, Ames, IA, 50011, Telephone: 1 - 515-294-3583, Fax: 1 - 515 - 294 - 8740, gwelk@iastate.edu.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplemental Digital Content
Code and Data.zip

Conclusions—When the study goal is to show similarity between methods, equivalence testing is more appropriate than traditional statistical tests of differences (e.g., ANOVA and t-tests).

Keywords

Calibration; validation; criterion validity; convergent validity

Introduction

Measurement research is an implicit and essential aspect of almost all areas of science. Studies in many fields depend on the availability of valid and reliable measures that can minimize error and bias. Increased precision is particularly important for answering more complex research questions since measurement error directly limits statistical power. Other fundamental needs are to develop alternative measures that can save costs, facilitate analyses and enable data collection in the field. This need is evident in many areas of science but it is particularly common in exercise science research where there is a continued need for better field-based measures of physical activity and physical fitness.

Two frequent measurement needs are to evaluate the validity of a new method (by comparing it to an established criterion) and to evaluate agreement between two alternative methods. The most common approach for evaluating validity and measurement agreement has been the Bland Altman method which enables error and bias to be visualized across a range of scores (1). These established principles of evaluating agreement (2, 3) have clearly advanced measurement research, but a limitation is that it does not enable the degree of agreement to be directly quantified. To address this, researchers routinely use standard statistical tests of mean differences (e.g. ANOVA, t-tests) to test the absolute differences between measures. However, this is a fundamentally flawed approach since these tests are designed to detect differences, not equivalence. Failure to reject the null hypothesis of “no difference” does not necessarily provide evidence of equivalence. Because of the nature of difference tests, studies with large samples are more likely to find statistically significant “differences” leading to conclusions that two measures do not agree. The opposite problem is more troubling, since smaller samples are less likely to detect differences, which ultimately leads to erroneous conclusions that a measure is “valid,” i.e. that two measures agree. Tests of mean difference are common in various lines of measurement agreement research, but that does not make them correct.

Previous papers have called for the use of more appropriate analytical methods when studying the measurement properties of physical activity/fitness assessment tools (4, 5). In this paper, we describe the use of “equivalence testing” as a more appropriate method for evaluating agreement among measures. A number of papers by our team have utilized this approach (6–15), but detailed descriptions of the methodology are needed to facilitate broader adoption. The statistical principles underlying equivalence testing will first be provided followed by examples of applications in exercise science research.

General principles of equivalence testing

Equivalence testing has been developed as a statistical approach to directly provide empirical “evidence of equivalence”, rather than implying “no evidence of differences” between different measurement tools. In performing an equivalence test, the traditional null and alternative hypotheses are reversed, meaning the null hypothesis is that two methods are not equivalent (i.e., the difference between them is large). Thus, if the observed difference is sufficiently small, that null hypothesis is rejected in favor of the alternative, that the difference is considered (on subject-matter grounds) to be equivalent to 0.

An immediate complication is that a reversed null hypothesis of “not zero” is practically impossible to reject, since normal variability will almost always lead to some degree of difference between two measures, regardless of whether that difference has practical significance or not. Thus, for equivalence testing the user must define an equivalence region, which is defined as the set of differences between population means that are considered equivalent to zero. Because we focus on the difference of population means, the methods we describe evaluate what is called average equivalence in the equivalence testing literature (30, pp. 314–315).

Choosing and justifying the equivalence region is one of the most difficult aspects of an equivalence test. The equivalence region may be specified in absolute terms, e.g., two methods are equivalent when the mean for a test method is within 5 units of the mean for a reference method, or in relative terms, e.g., two methods are equivalent when the mean for a test method is within 10% of the reference mean. In either case, the magnitude (5 units, or 10%) is selected either based on prior evidence on the clinical or practical importance of the value, or somewhat arbitrarily by the investigators. Our suggestions for various types of studies are discussed later.

If equivalence is specified as a test mean within 5 units of the reference mean, the equivalence region for the difference in means (δ) is $-5 < \delta < 5$. The null hypothesis for our equivalence test is that the difference is large, either $\delta \leq -5$ or $\delta \geq 5$, and the alternative hypothesis is that $-5 < \delta < 5$. There are many different statistical methods to test the equivalence null hypothesis (16), and we present two, the Two-One-Sided-Tests method and the confidence interval method.

Two-One-Sided-Tests method

In the Two-One-Sided-Tests (TOST) method (17), the null hypothesis of non-equivalence, either $\delta \leq -5$ or $\delta \geq 5$, is divided into two one-sided null hypotheses: *Ha*: $\delta \leq -5$ and *Hb*: $\delta \geq 5$. Each hypothesis, *Ha* and *Hb*, is tested by a one-sided test at level α. The null hypothesis of non-equivalence is rejected at level α only if both one-sided null hypotheses (*Ha* and *Hb*) are rejected at level α. The larger of the p-values for these individual one-sided tests of *Ha* and *Hb* is the p-value for the test of the overall null hypothesis of non-equivalence, which follows from the Intersection-Union method for constructing a statistical test. This method says that if a null hypothesis can be written as a union of two parts, e.g. $\delta \leq -5$ or $\delta \geq 5$, it can be tested at a significance level of α, e.g. 5%, by testing each part at level α (18). In other words, it is not necessary to report the p-values from each one-sided test, since the

overall null hypothesis test has a single p-value, which is equal to the larger of the two one-sided p-values.

This test procedure is illustrated in Figure 1a. The one-sided test of *Ha*: $\delta \leq -5$ rejects *Ha* when the observed difference is sufficiently larger than −5, i.e. any value along the right-pointing arrow. The one-sided test of *Hb*: $\delta \geq 5$ rejects *Hb* when the observed difference is sufficiently smaller than 5, i.e. any value along the left-pointing arrow. The rejection region for a statistical hypothesis test is the set of observed differences for which the test rejects the null hypothesis. The rejection region for the equivalence test contains the observed differences around 0 that are in both one-sided rejection regions.

Figure 1 also shows that the TOST approach is conservative in the sense that a test with a type I error rate of 5% will actually reject the null hypothesis at a probability somewhat less than 5%. This can be seen by considering the one-sided test of *Ha*, carried out at $\alpha = 0.05$. This means there is a 5% probability of finding an observed difference in the rejection region for *Ha* when the true difference is exactly −5. Since the rejection region for the two-part equivalence test is a subset of the rejection region for *Ha*, the probability of finding an observed difference in the rejection region of the equivalence test is less than 5%. Hence, the true type-I error rate of a nominal 5% equivalence test is something less than 5%.

This conservative characteristic of equivalence testing can lead to an interesting property of the TOST method when the standard error of the estimated difference, is large. When the standard error is large, the rejection region for test *Ha* may only include large positive values and the rejection region for test *Hb* may only include large negative values of the observed difference, such that there is no overlap between the two one-sided rejection regions and, consequently, no rejection region. Figure 1 illustrates this situation. This situation has sometimes been seen as a fallacy of the TOST approach, which can be avoided by more complicated test methods (18). However, we consider this interpretation of the TOST results to still be reasonable. It makes scientific sense; if the difference is poorly estimated (large standard error), it does not seem reasonable to claim confidently that the difference is close to 0.

Confidence interval method

A second view of an equivalence test is based on the confidence interval for the difference in means. Comparing the equivalence region and a confidence interval for the difference in means provides information about the p-value of the TOST equivalence test. The null hypothesis of non-equivalence is rejected at level α if the $100(1 - 2\alpha)\%$ confidence interval for the difference in means lies entirely with the equivalence region. This is somewhat unintuitive, since the α level of the equivalence test differs from the confidence level of the confidence interval. For example, if we are interested in an $\alpha = 5\%$ test of equivalence (often called 95% equivalence testing), we would calculate a 90% confidence interval for the difference in means. When the equivalence region is (−5, 5), we would reject the null hypothesis of non-equivalence and claim evidence of equivalence if the 90% confidence interval was (−4, 2), (−3, −1) or (2.9, 4.9). We would not reject the non-equivalence null hypothesis if the 90% confidence interval was (−7, −3) or (−6, 2) or (−10, 10). In all three of the non-reject cases, the confidence interval includes values outside the equivalence region.

This ability to conduct a test by calculating a confidence interval makes it possible to easily test equivalence even when a study has a complicated design. All that is necessary is that you can obtain an appropriate 90% confidence interval for the difference in means. That confidence interval is then used to test equivalence at $\alpha = 5\%$. For example, a study of youth activity may include multiple classes in multiple schools. Classes and schools are usually considered random effects, so the statistical model for comparing two measurement methods is a mixed model. Most mixed model software does not easily compute a one-sided test of the difference in means = 5. But, it is easy to get a 90% confidence interval for the difference from mixed model software.

Methods and assumptions

The methods and assumptions for equivalence testing depend on whether a surrogate measure is compared to a known reference value (i.e. criterion measure). When there is a known reference value, the equivalence region can be expressed as a mean response or a difference from the reference method mean. In this case, the bounds of the equivalence region are known exactly on the scale of the response variable. For example, if a reference value is known to be 120, an equivalence region could be defined as (105, 135) if it were known that deviations of 15 units had little to no practical importance. Or, the equivalence region could be defined as +/− 10% of the reference value, i.e. (108, 132) for a reference value of 120 if it were known that deviations of 10% had little to no practical importance. In both cases, the bounds of the equivalence region are known exactly because the reference value, or the mean for the reference method, is presumed to be known exactly.

Applying the relationship between two one-sided tests and the confidence interval for the mean difference, the surrogate measurement would be considered equivalent to the reference at $\alpha = 0.05$ if the 90% confidence interval for the surrogate measurement fell entirely inside the known equivalence range. If the data are paired, e.g., simultaneous measured by the surrogate and reference methods, the analysis would start by estimating the differences between the surrogate measurement and the constant reference value for each individual. The paired-data confidence interval for the mean difference would be compared to an equivalence region that is expressed in terms of the difference between the two methods. If a deviation of 15 units was considered to be of no practical importance, the surrogate method would be considered equivalent to the reference at $\alpha = 0.05$ if the 90% confidence interval for the difference fell entirely within the range (−15, 15).

Slightly different approaches are needed when neither method can be treated as an accurate reference. When the equivalence region is specified in absolute terms, e.g. (−15, 15), the equivalence test is conducted similarly, based on a confidence interval for the difference in means, using a two-sample or paired-data method, depending on the design. The analysis is more difficult if the equivalence region is expressed as a percent error (i.e. standardized difference between the two measures). As before, we wish to show that the mean responses for method A and method B are within 10% of each other. This specification has two interpretations: $\mu_A > (1 - 0.1)\ \mu_B$ or $\mu_A < (1 + 0.1)\ \mu_B$. Because neither “method A” nor “method B” is a criterion measure, we need to make sure that “within 10%” means the same thing for both μ_A/μ_B and μ_B/μ_A. The first interpretation above specifies the equivalence

bounds as $0.9 < \mu_A/\mu_B < 1/0.9$; the second specifies the bounds as $1/1.1 < \mu_A/\mu_B < 1.1$. The choice is arbitrary, but has little practical importance for a 10% difference because the two specifications of within 10% are very similar. They are more different for "within 20%" and even more different for "within 50%". For illustration, we will adopt the first specification of "within 10%": $0.9 < \mu_A/\mu_B < 1.1111$.

When both methods are measured simultaneously on a subject, we need a way to test the one-sided null hypothesis that $\mu_A/\mu_B < 0.9$ and the one-sided null hypothesis that $1.1111 < \mu_A/\mu_B$. When the sample average for method B (the denominator in the ratio) is a random variable, the ratio of sample averages, $\bar{Y}_A/\bar{Y}_B$, is not normally distributed and it is difficult to construct a confidence interval for that difference. However, each of the one-sided hypotheses can be re-expressed as a linear combination of normally distributed random variables, which avoids all the problems with ratios. The hypothesis $\mu_A/\mu_B < 0.9$ is equivalent to the hypothesis $\mu_A - 0.9\mu_B < 0$. The random variable $\bar{Y}_A - 0.9\ \bar{Y}_B$ is normally distributed when the two sample averages are normally distributed. Hence, the one-sided hypothesis that $\mu_A/\mu_B < 0.9$ can be tested by computing $D_A = Y_A - 0.9\ Y_B$ for each participant and doing a one-sample T-test using the D_A values. The null hypothesis is rejected if the average of the D_A values is sufficiently greater than 0. Similarly, the other bound can be tested by computing $D_B = Y_A - (Y_B/0.9)$ for each subject. The second one-sided hypothesis is rejected if the average of the D_B values is sufficiently smaller than 0.

If the two measurements are log-normally distributed, i.e., the log transformed values are normally distributed, the equivalence test can be simplified. If, as is commonly found, the two measurements have the same variance, the log ratio of the means for each method is the difference in their log transformed measurements. Hence, the equivalence criterion $0.9 < \mu_A/\mu_B < 1/0.9$ can be expressed as $\log 0.9 < \mu_{\log A} - \mu_{\log B} < -\log 0.9$. This can be tested at $\alpha = 0.05$ either by calculating a 90% confidence interval for the mean difference of log transformed values and comparing it to (log 0.9, −log 0.9), or by computing two one-sided tests, one with a null hypothesis value of log 0.9 and the second with a null hypothesis value of −log 0.9.

When planning a study that will use equivalence testing, it is still important to determine an appropriate sample size. The principles are the same as those for determining a sample size for a test of no difference. The sample size depends on the error variance, the number of observations, properties of the statistical test or confidence interval, and the expected mean difference. The details of the calculation have been described elsewhere (Chao, Shao, and Wang 2003, pp. 52 and 59). Briefly, for paired data, the number of paired observations is given by

$$n = \frac{(t_\alpha + t_{\beta/2})^2 \sigma^2}{(\delta - |\varepsilon|)^2}$$

where t_α and $t_{\beta/2}$ are the indicated quantiles of a t distribution with n−1 degrees of freedom, α is the type I error rate (often 0.05), $1 - \beta$ is the desired power (often 0.8), σ^2 is the error variance, δ is the upper bound of the equivalence region, and ε is the true difference in

means. This equation usually has to solved iteratively because the t quantiles depend on the choice of *n*. The use of this formula is demonstrated at the end of example 1 (below).

Examples of Equivalence Testing

We provide three examples from exercise science research that illustrate the principles discussed above and their application. The examples are based on challenges that arise in trying to evaluate the absolute and relative validity of different measures of physical activity and physical fitness, but these are provided just as illustrations of the methodology. The three examples include 1) equivalence of a PA monitor and known criterion (n=15; 7–11yrs) for energy expenditure during walking; 2) equivalence between two field-based measures of aerobic capacity with no known reference method (n=680; 13–17yrs); and 3) equivalence of a PA monitor and a criterion measure across 23 activities (n=43; 7–11yrs) using regression. The principles and approaches described in the examples can be used in any area of measurement research, but these examples will help explain how to actually conduct the comparisons.

Example 1: Comparison of a surrogate measure to a known criterion measure

This example demonstrates how equivalence testing procedures can be used to evaluate agreement in physical activity measures when a known criterion is available. The data were from a study by Kim et al. (12), which evaluated the relative validity of an accelerometry-based activity monitor for estimating energy expenditure. In this study, participants (43 children ages 7–11) wore two different monitors while performing a series of 12 different lab-based activities (randomly selected from a pool of 24 activities) and while being monitored with a portable indirect calorimetry system which served as the criterion measure. The goal in this type of study is to determine if the surrogate accelerometry measure is equivalent to the criterion, indirect calorimetry.

In the first part of this example, we consider one activity, brisk walking. This activity was measured on 15 individuals. The measurements from the activity monitor are made simultaneously with the indirect calorimetry measurement, so the data are paired. The mean MET using indirect calorimetry is 4.34, so to be equivalent using a ±15% equivalence region requires that the confidence interval for the difference between the surrogate and reference measures falls between −0.65 MET and 0.65 MET. For the activity monitor, the mean MET is 4.16. The standard error of the difference is 0.19, so the 90% confidence interval for the difference is (−0.52, 0.15). For this activity, the estimate from the monitor instrument is significantly equivalent to the reference with $p < 0.05$ because the 90% confidence interval of (−0.52, 0.15) is completely inside the equivalence region of (−0.65, 0.65).

If the actual p-value for the equivalence test is required, it is important to first calculate the p-values for the two one-sided tests. For the hypothesis Ha: $\delta < -0.65$, the T statistic is computed as $T = (4.16 – 4.34 – (-0.65)) / 0.19 = 2.47$. The associated one-sided p-value is the probability that a T random variable with 14 degrees of freedom is larger than 2.47, which is 0.013. For the hypothesis Hb: $\delta > 0.65$, the T statistic is computed as $T = (4.16 – 4.34 – 0.65) / 0.19 = -4.40$. The associated one-sided p-value is the probability that a T

random variable with 14 degrees of freedom is less than −4.40, which is 0.0003. The p-value for the equivalence test is the larger of the two one-sided p-values, i.e. 0.013.

To illustrate that equivalence requires more than just a non-significant difference in traditional tests between two measurements, we consider another activity, light calisthenics, which was measured in 19 subjects. The mean MET using indirect calorimetry is 3.77, so the 15% equivalence region is that the difference falls between −0.57 and 0.57. The estimated difference, 0.04, is within that region and the p-value for the usual paired t-test of no difference is 0.93. However, the standard error of the difference is 0.45, which is sufficiently large that the 90% confidence interval, (−0.74, 0.83) crosses outside the equivalence region. Hence, the estimate from this instrument is not significantly equivalent ($p > 0.05$) for this activity. The p-values for the two one-sided hypotheses are 0.098 for Ha: $\delta < -0.57$, and 0.13 for Hb: $\delta > 0.57$, so the p-value for the equivalence test is 0.13.

The decision about equivalence depends on the definition of the equivalence region. Returning to the brisk walking example, if equivalence was defined more narrowly, e.g. as within 10% of the reference measurement, the equivalence region would be (−0.43, 0.43). With this definition of equivalence, the activity monitor cannot be shown to be equivalent to indirect calorimetry. The p-values for the two one-sided tests for brisk walking are 0.10 and 0.0029, so the p-value for the equivalence test is 0.10. Conversely for light calisthenics, if equivalence was considered more liberal, e.g. as within 25% of the reference measurements, the equivalence region would be (−0.94, 0.94). With this definition of equivalence, the activity monitor is significantly equivalent when assessed with light calisthenics. The p-values for the two one-sided tests are 0.022 and 0.031, so the p-value for the equivalence test is 0.031.

When appropriate, the equivalence region can be set up asymmetrically around 0. For example, if under-estimating activity is considered more serious than over-estimating activity, the equivalence region could be defined as (−5%, 10%) of the reference mean activity. When the equivalence region is defined this way, a surrogate measure with a mean that is 6% lower than the reference cannot be equivalent but a surrogate that is 6% higher may be equivalent if the estimated difference in means is sufficiently precise.

A key question in this case is to determine how many subjects are needed to demonstrate equivalence within 10% of the reference measurement (i.e. an equivalence region of (−0.43, 0.43)). If we consider an $\alpha = 5\%$ test, and want a power of 80%, then $\beta = 0.2$, We believe that the error standard deviation will be similar to that seen in the current data, so $\sigma = 0.74$. The smallest sample size is when the true mean difference is $\varepsilon = 0$. In that situation, the sample size formula for paired data given by Chao, Shao, and Wang (2003, pp. 52) is as follows:

$$n = \frac{(t_{0.95} + t_{0.1})^2 0.74^2}{(0.43 - 0)^2}.$$

In this case, a sample size of n=27 subjects (26 df for the t quantiles) is needed. When the true difference is not 0, the appropriate sample size increases considerably. For example, if the true difference is $\varepsilon = 0.10$ (instead of 0), then a sample size of $n = 43$ is needed.

Example 2: Comparison of two methods when neither is a reference method

This example will demonstrate the utility of equivalence testing when a known criterion is not available. The example uses data from a study by Saint-Maurice et al. (9) which evaluated the agreement between two common field measures of aerobic fitness in youth. The data set includes a sample of 680 participants (324 boys and 356 girls) 7th through 12th grade that completed two different field based assessments of aerobic capacity as part of their normal physical education programming. One measure is from the Progressive Aerobic Cardiovascular Endurance Run (PACER) and the other is from the established Mile Run assessment. The validity of both assessments has been well established (19) but they have different measurement characteristics and motivational factors. The PACER test requires that participants complete a series of 20-meter shuttle runs at a progressively faster cadence until voluntary exhaustion (similar to a maximal treadmill test). The number of completed laps is then used to predict maximal oxygen consumption using established prediction equations (8). The Mile Run test is an alternative assessment of aerobic capacity that requires participants to complete a 1 mile distance as quickly as possible with estimates of aerobic capacity determined based on a prediction algorithm developed by Cureton et al (20). To evaluate the agreement between the two assessments, youth completed both tests (within 10 days of each other) in a counter balanced design. Both tests provide the same outcome indicator of aerobic capacity (VO2 in $ml \cdot kg^{-1} \cdot min^{-1}$) but neither can be considered a true criterion. Thus, this data set provides a useful example to demonstrate an equivalence test when neither method is a reference.

Summary statistics for the PACER and Mile Run assessments of fitness are shown in Table 1a. The mean difference of −0.39 ml/kg/min is approximately 1% of the mean for each method. Because of the large sample size, we will consider differences of up to 2% as equivalent. That means the null hypothesis is that $\mu_{pacer}/\mu_{mile} < 0.9800$ or $1.0204 < \mu_{pacer}/\mu_{mile}$. As described above, we compute $D_A = Y_{pacer} - 0.98\ Y_{mile}$ for each participant and do a one-sample T-test using the D_A values. We also compute $D_B = Y_{pacer} - 1.0204\ Y_{mile}$ for each participant and do a second one-sample T-test using the D_B values. The T statistic for the one-sided test of D_A is large and the T-statistic for the one-sided test of D_B is negative (Table 1a), so the p-values for both tests are small (0.0070 for D_A and < 0.0001 for D_B). The p-value for the equivalence test is the larger of the two p-values from the one-sided tests, i.e. 0.0070. There is strong evidence that the two assessments are equivalent.

Example 3: Comparison of surrogate and reference measures for many activities

This example is based on the same data used for Example 1 from Kim et al. (12), but highlights how equivalence testing can be used to assess overall agreement across a range of indicators or comparisons. A common strategy in research on activity monitors is to examine the validity for individual activities (often termed “point estimates”) but this has limited value since the overall goal is typically to determine the overall accuracy for an array of activities. Instead of testing mean equivalence for each separate activity, we really want to

capture a single evaluation of equivalence that takes into account all of the activities. When there is a reference measure, one way to do this is to fit a regression model, $surrogate = \beta_0 + \beta_1\ reference$, to the pairs of values (reference mean, surrogate mean) for each activity. If the surrogate method is equivalent to the reference method across multiple activities, then the intercept of this regression, β_0, is close to 0 and the slope, β_1, is close to 1.

The average of the estimates from the surrogate monitor and the indirect calorimetry system for each of the tested activities are plotted in Figure 2. The estimated regression line predicting the surrogate mean measurements from the reference mean measurements has an intercept of −1.23 and slope of 1.06 (Table 2). The intercept is not significantly different from 0 ($p = 0.0608$) and the slope is not significantly different from 1 ($p = 0.69$). The correlation of surrogate and reference means across the 23 activities is 0.82 ($p < 0.0001$). Using the traditional null hypotheses of no difference, it appears that the average activity measured by the surrogate monitor is similar to that measured by indirect calorimetry.

As with a single comparison of means, failing to reject a traditional null hypothesis (such as intercept = 0 and/or slope = 1) does not demonstrate equivalence. Robinson et al. (21) propose using equivalence tests and concluding equivalence when the intercept is equivalent to 0 and the slope is equivalent to 1. To claim equivalence across the array of activities at $\alpha = 0.05$ requires that two 90% confidence intervals, one for the intercept and one for the slope, fall inside their respective equivalence regions. Robinson et al. (21) suggested regression-based equivalence regions as +/− 10% of the reference mean for the intercept and (0.9, 1.1) for the slope (i.e., +/− 10% of the slope of 1 that would be expected for equal means on the two measures). As before, conclusion(s) about equivalence depend on the choice of equivalence region, which should be informed by its clinical or practical relevance or subject-matter considerations.

We recommend a revision of the Robinson et al. (21) approach when assessing physical activity monitors. When a regression is fit to data points where the X values are the mean reference measurement for an activity, the regression intercept estimates the mean surrogate measure when the reference measure equals 0. For physical activity, this X value is on the edge of possible values and is quite likely to be far outside a relevant range of activity. We recommend centering the X values by subtracting the overall reference mean (averaged over all activities) from the mean reference values. The intercept now describes an average activity. To maintain an intercept of 0 when two methods are equivalent, the overall reference mean (average X) is also subtracted from all Y values. The regression slope is unchanged by centering. Table 2 illustrates effects of this adjustment on the estimated regression intercept and slope.

We evaluate equivalence for the collection of activities using both regressions. We use the equivalence regions suggested by Robinson et al. (21) which for the slope is (0.9, 1.1). The overall average of all 23 activities is 3.73 MET, so the intercept equivalence region is (−0.37, 0.37). The 90% confidence intervals for both the intercept and the slope (Table 2) cross outside their respective equivalence regions, so we have not demonstrated equivalence. The adjusted regression leads to the same conclusion (Table 2). The intercept is more precisely estimated in the adjusted regression (Table 2), but the confidence interval for the intercept

now falls completely outside its equivalence region. The null hypothesis of non-equivalence cannot be rejected. Even though the intercept and slope are not significantly different from their target values, the estimates of those parameters are not sufficiently precise to show equivalence.

When a p-value is required for the regression test of equivalence, it is essential to examine the individual tests that associated with the 90% confidence interval. The regression test of equivalence requires four one-sided tests: two that compare the intercept to its equivalence region and two that compare the slope to its equivalence region. The Intersection-Union test principle (18) says that the p-value for the equivalence test is the largest of the p-values from these four one-sided tests. Using the adjusted regression, the p-values for the one-sided tests of the intercept are 1.000 and < 0.0001 and the p-values for the one-sided tests of the slope are 0.16 and 0.41. Hence, the p-value for the equivalence test of both the intercept and slope is 1.000.

The statistical methods used to estimate the 90% confidence interval or test one-sided hypotheses can be simple when appropriate or complicated when necessary. For example, different numbers of subjects completed each activity, from a minimum of 6 subjects to a maximum of 36 subjects. Thus, the activity means for the activities done by large numbers of subjects are more precisely estimated. This can be accounted for by fitting a meta-regression mixed model (22). The results of fitting that mixed model are estimates and confidence intervals for the intercept and the slope. For these data, the consequences of unequal numbers of observations for each activity are minor: the 90% confidence intervals from the mixed model with the adjusted X and Y, (−1.19, −0.68) for the intercept and (0.75, 1.32) for the slope, are almost the same as those from the simple regression model. The important point is that once confidence intervals are available, the evaluation of equivalence can be done without considering the details of how those intervals were computed.

Discussion

Science progresses through incremental advances in knowledge as well as through improved methods for evaluating and studying phenomena. Researchers often follow methods used in past studies but this has limitations if the past methods are inherently flawed. This paper documents that there are major limitations to the “standard” methods used to evaluate validity and measurement agreement in exercise science and physical activity research. The key point is that traditional tests of differences (e.g. ANOVA and t-tests) are inappropriate for evaluating agreement. Significant correlations are not sufficient to document validity, and failing to show a difference between two methods does not demonstrate that measures are equivalent. There are numerous examples in the exercise science literature of validity being inferred based on inappropriate analyses and it is particularly concerning when results from poorly-conducted studies get cited as evidence and become further reinforced in the literature over time. To advance clinical and research practice more attention needs to be paid to the selection of statistical tests being used.

The present study documents the advantages and specific applications of equivalence testing (as opposed to tests of differences) for evaluating measurement agreement. A number of

recent papers provide research-based examples of the specific advantages and applications of equivalence tests for methodological comparisons (7, 8, 10, 12, 13, 15, 17, 21, 23), but the present paper provides the background and concepts needed to promote broader adoption and utilization in measurement research..

This paper focused on average equivalence by providing specific examples of how equivalence testing can be used in different types of exercise science research. Examples of the applications were provided for evaluating the utility of field measures of physical activity and physical fitness, but the method has applications for any realm of measurement research focused on agreement. While equivalence testing should not be used alone to define agreement, if reported consistently in validation studies, this approach would facilitate systematic reporting and evaluations of different measurement instruments and methods. To facilitate adoption of the method, sample SAS code and R code (and associated Excel data files) are provided for each of the examples presented in the text (see Supplemental Digital Content 1, Code and Data.zip).

A few additional comments are warranted to help ensure appropriate use of these methods. A challenge for some in interpreting equivalence tests is that (like a Bland-Altman plot) the visual examination of the confidence intervals can seem somewhat subjective. However, this is not the case with equivalence tests since the ranges are determined and judged empirically. As described, p values can be computed for equivalence tests but we very strongly recommend reporting the confidence interval since it provides much more information than the p-value. However, if statistical significance is defined using traditional methods (i.e. a p-value < 0.05), this translates to reporting 90% confidence intervals when evaluating equivalence. In most situations, the confidence interval is easier to obtain from statistical software. The one commonly occurring exception is when equivalence is expressed in terms of a ratio without a known criterion value (e.g., example 2). In that case, it is easy to construct tests for specific ratios, but more difficult to estimate a confidence interval for the ratio.

It is important to acknowledge that the concept of “statistical equivalence” is heavily influenced by the choice of the equivalence region. This should be determined by the intended use of the surrogate measure. Some guidelines are available from other application areas. The US Food and Drug Administration requires that a proposed generic drug satisfies an average equivalence comparison to a patented drug, which is the reference measure, with an equivalence region of (80%, 125%) (24). The US Occupational Safety and Health Administration has an individual-like equivalence criterion that at least 90% of surrogate measurements be within (75%, 125%) of a reference measurement, with 95% confidence (25). The FDA bioequivalence criterion is the result of practical experience and extensive discussions between drug developers and regulators. The appropriate choice, or choices, of equivalence region for exercise science research depend on the application. Readers should understand that the selection of confidence interval for equivalence is arbitrary in the same way as the selection of $p < .05$ is arbitrary for significance tests.

In conclusion, the present study provides a detailed background on equivalence testing and its potential to advance measurement research in exercise science. The rationale for this

approach has been well documented but further discussions are needed towards reconciling how to best use this information to guide researchers developing/testing new measures and researchers interested in the direct applications of these same measures. A good illustration of such efforts includes those aimed at standardizing procedures for measure selection or those aimed at harmonizing activity outcomes generated by different measures of physical activity (26, 27). Similar efforts could be applied to develop standardized approaches for evaluating new and existent measures of physical activity or fitness while providing clear documentation for the implications (regarding measurement error) of choosing one measure over another. Thus, this framework could demonstrate value for improving the quality of the measures employed in exercise science research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The manuscript was conceived over many years of conducting collaborative measurement agreement research in the Physical Activity and Health Promotion Lab at Iowa State University (www.physicalactivitylab.org). Sample data files along with associated code for each of the 3 examples (in R and SAS) is available on the website to facilitate adoption and utilization of the methods. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation, and statement that results of the present study do not constitute endorsement by ACSM.

Source of Funding: There were no sources of funding contributing to this paper.

References

1. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PLoS One. 2012; 7(5):e37908. [PubMed: 22662248]
2. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet. 1986; 1(8476):307–10. [PubMed: 2868172]
3. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999; 8(2):135–60. [PubMed: 10501650]
4. Staudenmayer J, Zhu W, Catellier DJ. Statistical considerations in the analysis of accelerometry-based activity monitor data. Med Sci Sports Exerc. 2012; 44(1 Suppl 1):S61–7. [PubMed: 22157776]
5. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. Med Sci Sports Exerc. 2009; 41(1):3–13. [PubMed: 19092709]
6. Saint-Maurice PF, Kim Y, Hibbing P, Oh A, Perna F, Welk GJ. Calibration and Validation of the Youth Activity Profile: the FLASHE study. American Journal of Preventive Medicine. In Press.
7. Saint-Maurice PF, Welk GJ. Validity and Calibration of the Youth Activity Profile. PloS one. 2015; 10(12):e0143949. [PubMed: 26630346]
8. Saint-Maurice PF, Welk GJ, Finn KJ, Kaj M. Cross-Validation of a PACER Prediction Equation for Assessing Aerobic Capacity in Hungarian Youth. Res Q Exerc Sport. 2015; 86(Suppl 1):S66–73. [PubMed: 26054958]
9. Saint-Maurice PF, Anderson K, Bai Y, Welk GJ. Agreement Between VO2peak Predicted From PACER and One-Mile Run Time-Equated Laps. Res Q Exerc Sport. 2016; 87(4):421–6. [PubMed: 27586563]
10. Kim Y, Welk GJ. The accuracy of the 24-h activity recall method for assessing sedentary behaviour: the physical activity measurement survey (PAMS) project. J Sports Sci. 2017; 35(3): 255–61. [PubMed: 27019092]

11. Kim Y, Welk GJ. Criterion Validity of Competing Accelerometry-Based Activity Monitoring Devices. Med Sci Sports Exerc. 2015; 47(11):2456–63. [PubMed: 25910051]
12. Kim Y, Crouter SE, Lee JM, Dixon PM, Gaesser GA, Welk GJ. Comparisons of prediction equations for estimating energy expenditure in youth. J Sci Med Sport. 2016; 19(1):35–40. [PubMed: 25459235]
13. Lee JM, Kim Y, Welk GJ. Validity of consumer-based physical activity monitors. Med Sci Sports Exerc. 2014; 46(9):1840–8. [PubMed: 24777201]
14. An HS, Kim Y, Lee JM. Accuracy of inclinometer functions of the activPAL and ActiGraph GT3X +: A focus on physical activity. Gait Posture. 2017; 51:174–80. [PubMed: 27780084]
15. Bai Y, Welk GJ, Nam YH, et al. Comparison of Consumer and Research Monitors under Semistructured Settings. Med Sci Sports Exerc. 2016; 48(1):151–8. [PubMed: 26154336]
16. Wellek, S., editor. Testing statistical hypotheses of equivalence and noninferiority. 2. Boca Raton: CRC Press; 2010. p. 415xvi
17. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987; 15(6):657–80. [PubMed: 3450848]
18. Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science. 1996; 4:283–319.
19. Plowman, S., Meredith, M. Fitnessgram/Activitygram Reference Guide. Dallas, TX: The Cooper Institute; 2014. Available from: The Cooper Institute
20. Cureton KJ, Sloniger MA, O'Bannon JP, Black DN, McCormack WP. A generalized equation for prediction of VO2 peak from one-mile run/walk performance in youth. Medicine and Science in Sports and Exercise. 1995; 27:445–51. [PubMed: 7752874]
21. Robinson AP, Duursma RA, Marshall JD. A regression-based equivalence test for model validation: shifting the burden of proof. Tree Physiol. 2005; 25(7):903–13. [PubMed: 15870057]
22. Fletcher D, Dixon PM. Modelling data from different sites, times or studies: weighted vs. unweighted regression. Methods Ecol. Evol. 2012; 3(1):168–76.
23. Welleck, S. Testing Statistical Hypotheses of Equivalence and Noninferiority. Boca Raton, FL: Chapman and Hall / CRC Press; 2010. Multisample tests for equivalence.
24. Food and Drug Administration, FDA. Guidance for Industry: Statistical Approaches to Establishing Bioequivalence. MUF Rockville: Center for Drug Evaluation and Research; 2001. p. 45editor2001
25. Krishnamoorthy K, Mathew T. Statistical methods for establishing equivalency of a sampling device to the OSHA standard. AIHA J. (Fairfax, Va). 2002; 63(5):567–71.
26. Brazendale K, Beets MW, Bornstein DB, et al. Equating accelerometer estimates among youth: The Rosetta Stone 2. J Sci Med Sport. 2016; 19(3):242–9. [PubMed: 25747468]
27. Strath SJ, Kaminsky LA, Ainsworth BE, et al. Guide to the assessment of physical activity: Clinical and research applications: a scientific statement from the American Heart Association. Circulation. 2013; 128(20):2259–79. [PubMed: 24126387]

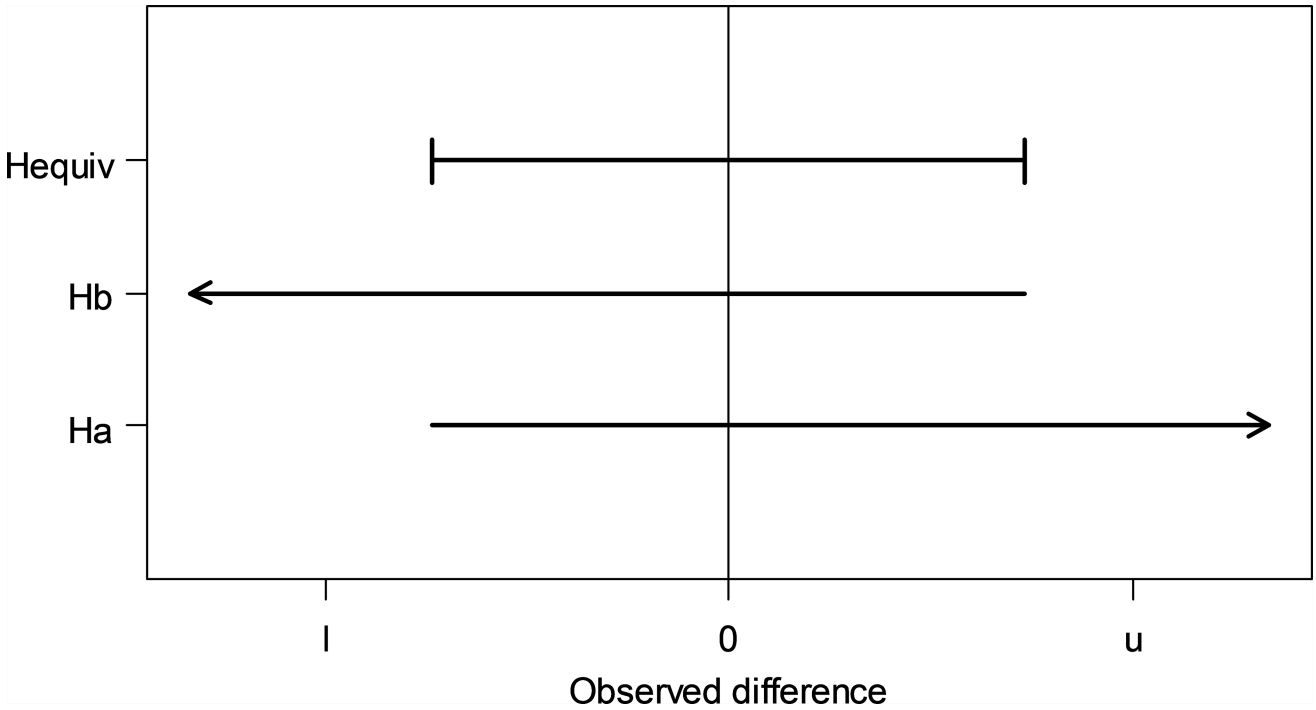

Figure 1.

Rejection regions for the two one-sided T tests and for the equivalence test. The equivalence region is (l, u). Ha is the hypothesis that $Ha: \delta < l$ and Hb is the hypothesis that $Ha: \delta > u$. The equivalence null hypothesis, Hequiv, is that $Ha: \delta < l$ *or* $\delta > u$. Each line segment shows the set of observed differences for which that null hypothesis is rejected.

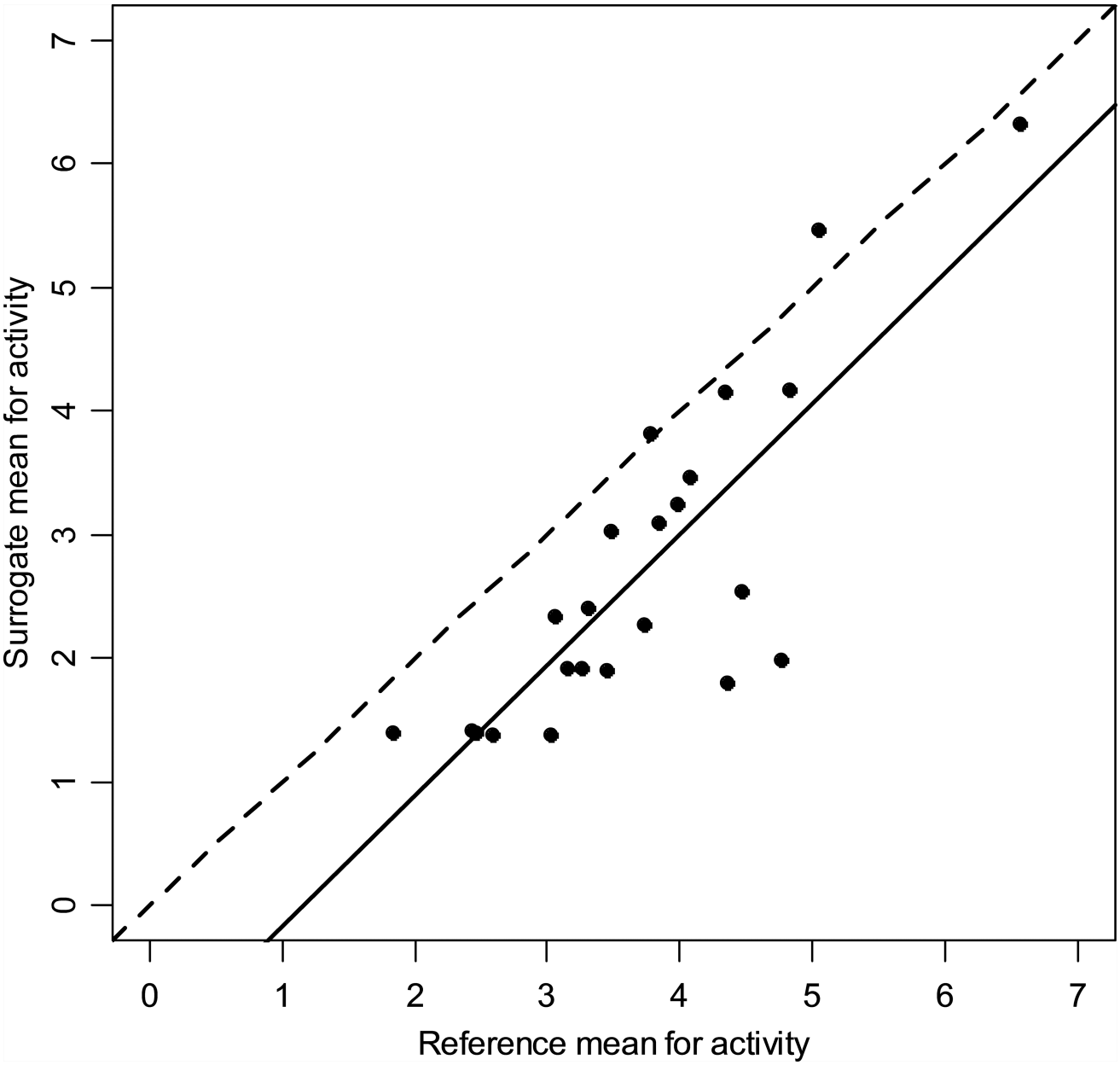

Figure 2.
Plot of the mean measurements for the reference method and the surrogate method for 23 activities. The solid line is the fitted regression line. The dashed line shows equality of the two measurements.

Table 1

Summary statistics for the pacer and mile run assessments of fitness on 680 school children. Both pacer and mile run values are expressed as VO_2 for comparability. Da and Db statistics are the quantities used in each one-sided test.

Statistic	average	se	T statistic	one-sided p-value
Pacer	43.14	0.27		
mile run	43.54	0.26		
Difference	−0.39	0.20		
Da	0.48	0.19	2.46	0.0070
Db	−1.28	0.20	−6.50	< 0.0001

Table 2

Results from regression of mean novel measurement on mean reference measurement for 23 activities.

	Intercept			Slope		
Model	**Estimate**	**se**	**90% CI**	**Estimate**	**se**	**90% CI**
Unadjusted	−1.23	0.62	(−2.31, −0.16)	1.06	0.16	(0.79, 1.34)
Adjusted X and Y values	−0.99	0.16	(−1.28, −0.71)	1.06	0.16	(0.79, 1.34)

CI: Confidence Interval