International Journal of Statistics and Probability; Vol. 3, No. 4; 2014
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Evaluating Variables as Unbiased Proxies for Other Measures: Assessing the Step Test Exercise Prescription as a Proxy for the Maximal, High-Intensity Peak Oxygen Consumption in Older Adults

Jonathan D. Mahnken[1,2], Xueyi Chen[1], Alexandra R. Brown[1], Eric D. Vidoni[2,3], Sandra A. Billinger[4] & Byron J. Gajewski[1]

[1] Department of Biostatistics, The University of Kansas Medical Center, Kansas City, KS, USA

[2] The University of Kansas Alzheimer's Disease Center, Kansas City, KS, USA

[3] Department of Neurology, The University of Kansas Medical Center, Kansas City, KS, USA

[4] Department of Physical Therapy and Rehabilitation Science, The University of Kansas Medical Center, Kansas City, KS, USA

Correspondence: Jonathan D. Mahnken, Department of Biostatistics, The University of Kansas Medical Center, Kansas City, KS, USA. Tel: 1-913-588-4703. E-mail: jmahnken@kumc.edu

Received: June 24, 2014 Accepted: September 11, 2014 Online Published: September 23, 2014

doi:10.5539/ijsp.v3n4p25 URL: http://dx.doi.org/10.5539/ijsp.v3n4p25

Abstract

To assess validity of a low-intensity measure of fitness (X) in a population of older adults as a proxy measure for the original, high-intensity measure (Y), we used ordinary least square regression with the new, potential proxy measure (X) as the sole explanatory variable for Y. A perfect proxy measure would be unbiased (i.e., result in a regression line with a y-intercept of zero and a slope of one) with no error (variance equal to zero). We evaluated the properties of potential biases of proxy measures. A two degree-of-freedom approach using a contrast matrix in the setting of simple linear ordinary least squares regression was compared to a one degree-of-freedom paired t test alternative approach. We found that substantial improvements in power could be gained through use of the two degree-of-freedom approach in many settings, while scenarios where no linear bias was present there could be modest gains from the paired t test approach. In general, the advantages of the two degree-of-freedom approach outweighed the benefits of the one degree-of-freedom approach. Using the two degree-of-freedom approach, we assessed the data from our motivating example and found that the low-intensity fitness measure was biased, and thus was not a good proxy for the original, high-intensity measure of fitness in older adults.

Keywords: bias, linear contrast, ordinary least squares regression, paired *t* test

1. Introduction

Proxy measures evaluation has been presented in the context where a true, underlying measure may be unobservable, so this measure may be replaced by an observable proxy (e.g., Kotz & Johnson, 1986, p. 323; Trenkler & Stahlecker, 1996; Hawkins, 2002). It is often of interest of scientific investigators to develop a new measure to serve as a proxy for another in the setting where the original is observable. Proxy measures can be of great importance, particularly when costs, broadly defined, are substantially less for measuring the new proxy compared to the original measure. A systematic review by Dickinson, Hrisos, Eccles, Francis and Johnston (2010), however, found problems with methods frequently used to evaluate proxy measures.

The focus of this work was on the evaluation of whether a new measure, X, served as a good proxy for the original measure Y. We evaluated the proxy for the setting where $Y|X$ was (at least approximately) a continuous measure. In the Methods section that follows, we first defined our criteria for assessing whether a measure serves as a valid proxy. We used simple linear regression of Y onto X to provide a formal means to assess for validity as a proxy, and compared this approach to a simpler strategy of taking the difference between measures for each subject, and testing the null hypothesis that the mean of these differences is zero using a paired t test.

The motivating example was the assessment of a lower-impact, self-paced, submaximal step test (the Step Test Exercise Prescription, or STEP test; Petrella, Koval, Cunningham & Paterson, 1998; Petrella Koval, Cunningham & Paterson, 2001; Petrella & Wight, 2000) as a proxy for the rigorous, maximal, high-intensity peak oxygen consumption (VO$_2$ peak) test to measure fitness levels. More specifically, we were interested in the validity of this potential proxy in an older population than had been previously examined that included subjects with Alzheimer's Disease (Vidoni et al., 2013). A major research focus of The University of Kansas Alzheimer's Disease Center (KU ADC), which is funded by the National Institute on Aging, is on modifiable lifestyle risk factors and Alzheimer's Disease prevention. Thus, a submaximal fitness assessment could serve as a key measure of interest for research at the KU ADC. Discussions compared and contrasted these two approaches for assessing proxy measures in general settings, and conclusions from our motivating example were presented.

2. Methods

For a variable, X, to serve as a valid proxy measure for another (the original) variable, Y, the relationship between these two variables would be approximately $Y = X$. In a simple linear regression framework, this relationship would ideally be $(Y \mid X = x) = x + E$, where E represents the random error and most of the distribution of E falls at or near zero (Kotz & Johnson, 1986, p. 323). A normal distribution for E with a mean of zero and small variance parameter meets these criteria; and is also advantageous given the wealth of literature on the theory and practice of ordinary least squares regression that fit under this simple linear regression paradigm (e.g., Draper & Smith, 1998, Chap. 1; Kleinbaum, Kupper, Muller & Nizam, 1998, Chap. 5; Kutner, Nachtsheim & Neter, 2004, Chap. 1), though as noted by Lin (1989), settings where the variance of E is large may be problematic for this approach. Bell-shaped distributed errors seem plausible for the difference between two measures—regardless of the underlying distributions of the individual random variables X and Y—if Y indeed is a valid proxy for X. It is noteworthy that the assumptions of this model did not require both Y and X to be continuous, so long as the conditional distribution of $Y|X$ is approximately normally distributed.

We formally defined the model using the simple linear regression paradigm for assessment of X as a proxy measure for Y as

$$(Y|X = x) = \beta_0 + \beta_1 x + E \tag{1}$$

where $E \sim N(0,\ \sigma^2)$. As noted above, one of our properties of a valid proxy was a small variance; however, the property of unbiasedness in and of itself may be important, for example in some clinical settings, where there is motivation to report a proxy without adjustment. Hence, the focus of this work was on the property of the unbiasedness of the proxy measure; specifically, that $\beta_0 = 0 \cap \beta_1 = 1$ in (1). To formally test the null hypothesis that H_0: $\beta_0 = 0 \cap \beta_1 = 1$, we used a contrast matrix to perform the simultaneous, two degree-of-freedom test.

2.1 Linear Transformation

To facilitate the two approaches investigated, we first re-parameterized this two degree-of-freedom approach using a linear transformation. From (1), we subtracted the observed value of X (i.e., x) to define the random variable $Z_1|X$, or

$$(Z_1|X = x) = (Y - X|X = x) = \beta_0 + \beta_1 x - x + E$$

$$(Z_1|X = x) = \beta_0 + (\beta_1 - 1)x + E. \tag{2a}$$

Re-labeling the parameters yielded

$$(Z_1|X = x) = \gamma_0 + \gamma_1 x + E. \tag{2b}$$

From (2b), assessment for unbiasedness of the proxy measure as $\beta_0 = 0 \cap \beta_1 = 1$ in (1) was equivalent to testing $\gamma_0 = \gamma_1 = 0$. Under the assumptions defined above regarding the error, $E \sim N(0,\ \sigma^2)$, testing for unbiasedness can be done by comparing

$$F = \frac{\left(\underline{\widehat{\gamma}} - \underline{\gamma}\right)^T \left(\underline{\widehat{\gamma}} - \underline{\gamma}\right)/2}{\left(\underline{z_1} - X\underline{\widehat{\gamma}}\right)^T \left(\underline{z_1} - X\underline{\widehat{\gamma}}\right)/(n-2)}, \tag{3a}$$

where $\underline{\gamma}$, which represented the true, unknown parameters $(\gamma_0,\ \gamma_1)^T$, to a central F-distribution with two and $n - 2$ degrees of freedom in the numerator and denominator, respectively (e.g., using Rencher, 2000, pp. 184-190). In (3a), $\underline{z_1}$ represented the observed vector of differences between the n original (Y) and proxy (X) measures, and $\underline{\widehat{\gamma}}$ the ordinary least squares maximum likelihood estimates (MLEs) for $\underline{\gamma}$. In the Appendix we demonstrated that this analogous transformed approach produced equivalent inferences as $\widehat{\gamma}_0 = \widehat{\beta}_0$ and $\widehat{\gamma}_1 = \widehat{\beta}_1 - 1$.

To test for unbiasedness we defined the null hypothesis as H_0: $\underline{\gamma} = \underline{0}$, which yielded

$$F = \frac{\widehat{\underline{\gamma}}^T \widehat{\underline{\gamma}}/2}{\left(\underline{z_1} - X\widehat{\underline{\gamma}}\right)^T \left(\underline{z_1} - X\widehat{\underline{\gamma}}\right)/(n-2)}. \tag{3b}$$

Under the alternative hypothesis, $\underline{\gamma} \neq \underline{0}$ (i.e., the proxy was not valid as the proxy was a biased measure of the original), the distribution of F (3b) followed a non-central F-distribution with non-centrality parameter

$$\lambda = \underline{\gamma}^T X^T X \underline{\gamma}/2\sigma^2 = (n\gamma_0^2 + 2n\overline{x}\gamma_0\gamma_1 + \gamma_1^2 \underline{x}^T \underline{x})/2\sigma^2 \tag{4a}$$

(Rencher, 2000, p. 185). This followed from the fact that the contrast matrix was the identity matrix for testing this null hypothesis. We further refined (4a) by substituting for $\underline{x}^T \underline{x}$ using the formula for the sample variance of the proxy measure (X),

$$\underline{x}^T \underline{x} = (n-1)\, s^2 + n\overline{x}^2,$$

which yielded

$$\lambda = \left\{n\gamma_0^2 + 2n\overline{x}\gamma_0\gamma_1 + \gamma_1^2 \left[(n-1)\, s_x^2 + n\overline{x}^2\right]\right\}/2\sigma^2. \tag{4b}$$

[The subscript was added to s_x^2 in (4b) to more clearly indicate it as the sample variance from the proxy measure, X, which represented the explanatory variable in underlying models of (1) and (2b).]

2.2 Paired t Test Approach

We evaluated an alternative approach for assessing whether the proxy measure was biased (and therefore not a valid proxy) by conducting the one-sample paired t-test on $(Z_2|X = x)$, which from (2a) was defined as $(Y - X \mid X = x)$. In contrast to the model used by (2a), this alternative approach excluded the linear term (γ_1) altogether. The test for unbiasedness to assess the validity of the proxy measure was similarly constructed as in (3a-b). However, the design matrix on the right hand side of the equation, X^* (say), was equal to the $n \times 1$ vector $\underline{1}$; thus, the MLE $\widehat{\underline{\gamma}}^*$ (say) for this alternative approach was the scalar $\left(\underline{1}^T \underline{1}\right)^{-1} \underline{1}^T \underline{z_2} = \overline{z}_2$. The true, unknown mean parameter vector was also a scalar, so

$$\underline{\gamma}^* = \gamma^* = \gamma_0 + \gamma_1\overline{x}$$

where γ_0 and γ_1 were the same parameters from the model of (2b) (Rencher, 2000, p 154). This resulted in

$$F^* = \frac{\left(\widehat{\underline{\gamma}}^* - \underline{\gamma}^*\right)^T \left(\widehat{\underline{\gamma}}^* - \underline{\gamma}^*\right)/1}{\left(\underline{z_2} - X\widehat{\underline{\gamma}}^*\right)^T \left(\underline{z_2} - X\widehat{\underline{\gamma}}^*\right)/(n-1)} = \frac{(n-1)(\overline{z}_2 - \gamma_0 - \gamma_1\overline{x})^2}{\left(\underline{z_2} - \overline{z}_2\underline{1}\right)^T \left(\underline{z_2} - \overline{z}_2\underline{1}\right)}, \tag{5a}$$

which was compared to a central F-distribution with one and $n - 1$ degrees of freedom in the numerator and denominator, respectively. To test for unbiasedness under this paired one-sample t test approach we defined the null hypothesis as

$$H_0 : \gamma^* = 0 \Longleftrightarrow \gamma_0 + \gamma_1\overline{x} = 0 \Longleftrightarrow \gamma_0 = -\gamma_1\overline{x},$$

which yielded

$$F^* = \frac{(n-1)\overline{z}_2^2}{\left(\underline{z_2} - \overline{z}_2\underline{1}\right)^T \left(\underline{z_2} - \overline{z}_2\underline{1}\right)}. \tag{5b}$$

Under the alternative hypothesis, $\gamma^* \neq 0$ (i.e., the proxy was not valid as it was a biased measure of the original), the distribution of F^* (5b) followed a non-central F-distribution with non-centrality parameter

$$\lambda^* = n(\gamma_0 + \gamma_1\overline{x})^2/2\sigma^2 \tag{6}$$

(Rencher, 2000, p. 185). This followed from the fact that the contrast matrix was the scalar value of one for this null hypothesis. Of note, the null hypothesis for this paired t test approach, H_0: $\gamma_0 = -\gamma_1\overline{x}$, indicated it would be unable, in some settings, to reject the null hypothesis of a biased proxy even when it was truly biased (i.e., $\gamma_0 \neq 0 \cup \gamma_1 \neq 0$). This represented a potential shortcoming of the paired t test approach.

2.3 Power of the Two Approaches

From (3b), (4b), (5b), and (6), power functions were derived based on a given design matrix. Specifically, the probability of rejecting the null hypothesis (that X was a valid proxy for Y because it was unbiased) when in fact it was not unbiased followed a non-central $F_{2,\, n-2,\, \lambda}$ distribution, and for the alternative approach it followed a non-central $F_{1,\, n-1,\, \lambda^*}$ distribution. The non-centrality parameters (λ and λ^*) were functions of the true parameters γ_0, γ_1, [or analogously β_0 and β_1 using the form of (1)], σ^2, n, and $\overline{x}$. For the initial proposed approach (with two degrees of freedom in the numerator), λ was also a function of s_x^2. Expanding the polynomial term in λ^* and using substitution, we wrote λ as

$$\lambda = \lambda^* + (n-1)\, s_x^2 \gamma_1^2 / 2\sigma^2.$$

Writing the non-centrality parameter, λ, in this way implied power advantages in the proposed two degree-of-freedom approach to detect linear deviations from unbiasedness, particularly: 1) when there was more variation in the proxy measure collected (X) as quantified by s_x^2; or 2) when the absolute value of γ_1 was large (assuming a sample size of $n > 1$). However, when these values were small relative to σ^2 (the variance in the differences given X), the non-centrality parameter λ approached that of λ^*, which gave potential advantage to the paired t test approach based on the differences in the degrees of freedom for the respective F tests.

To facilitate our comparison of the theoretical results for these two approaches, a receiver operating characteristics (ROC) curve approach was used similar to that done in Mahnken, Wick, Gajewski and Mayo (2010). Specifically, we plotted the type I error along the horizontal axis and the power on the vertical axis. Notably, in the setting of the valid proxy actually being unbiased (i.e., when the null H_0: $\gamma_0 = \gamma_1 = 0$ was true) the power was expected to equal to the type I error, and so both approaches were anticipated to produce identical curves in this setting where the power equaled the type I error.

3. Motivating Example

A cohort of 102 adult research participants was recruited through the KU ADC Registry cohort and a concurrent study in the Research in Exercise and Cardiovascular Health (REACH) Laboratory (Billinger, van Swearingen, McClain, Lentz, & Good, 2012). The protocols for obtaining the original (Y, VO$_2$ peak) and proxy (X, STEP) measures were described in Vidoni et al. (2013). While differences in the KU ADC Registry and REACH Laboratory measure collections differed slightly as described by Vidoni et al., these data provide a useful example for the statistical discussions presented here. Further, we treated all subjects from the two study sources as homogeneous for our motivating example sections below. Approval was provided for both parent studies that obtained the source data measures by The University of Kansas Medical Center Human Subjects Committee (#11132 and #12460).

Simple linear regression was used to assess the STEP test (X) as an unbiased proxy for VO$_2$ peak (Y). The GLM procedure in SAS version 9.3 (SAS Institute Inc., Cary, NC, 2002-2010) was used for this analysis. To test for unbiasedness of our proxy measure, we used the form $\gamma_0 = \gamma_1 = 0$ for our null hypothesis corresponding to (2b). This enabled the use of the CONTRAST statement within the GLM procedure, which limited hypothesis testing to be subject to the constraint

$$L\underline{\beta}_0 = \underline{0}.$$

For our hypothesis, this was only the case using the linear transformation approach [i.e., the model of (2b)]. As stated above and demonstrated in the Appendix, this linear transformation approach produced identical inferences for corresponding hypothesis tests. Model assessment by residual analysis was not further described nor their results presented as this was not the focus of the work presented here. More formal presentation of the motivating example results, including subgroup-specific assessments, was presented previously in Vidoni et al. (2013). The alternative, paired t test approach was also performed for comparison.

4. Results

Plots were generated to compare the power curves using an ROC curve approach. This enabled comparisons of power across varying levels of type I error. While presented over the entire range of [0, 1], the focus of the results will often be around those portions of the plots with lower type I error. Curves were generated with varying values of β_0 and β_1 of -0.25(0.25)0.25 and 0.75(0.25)1.25, respectively such that they spanned null and alternative hypotheses–including non-null cases where the paired t test approach fit for testing the true, underlying model. The effect of the variance, σ^2, of the original model (1) was held fixed at one across all models.

4.1 Power as a Function of $\overline{x}$

Sample mean ($\overline{x}$) values for the proxy measures (X) varied within each plot in Figure 1 from -1(1)1. In each case where the slope parameter was not equal to one (i.e., where the alternative approach under-fit the relationship between the original and its proxy) the proposed (two degree-of-freedom) approach outperformed or did no worse than the alternative paired t test (one degree-of-freedom) approach. This was demonstrated by the dashed curves being closer to the point coordinate (0, 1) on the plots than the correspondingly-colored solid curves, or the two corresponding curves overlaying one another. In the cases where there was no linear bias in the proxy measure (i.e., where $\beta_1 = 1$), the one degree-of-freedom paired t test approach outperformed the two degree-of-freedom approach except in the case where the proxy was completely unbiased. In this latter case (i.e., where $\beta_0 = 0$ and $\beta_1 = 1$) the non-centrality parameters λ and λ^* where both equal to zero, so the power had a 1:1 relationship with the type I error in that case.

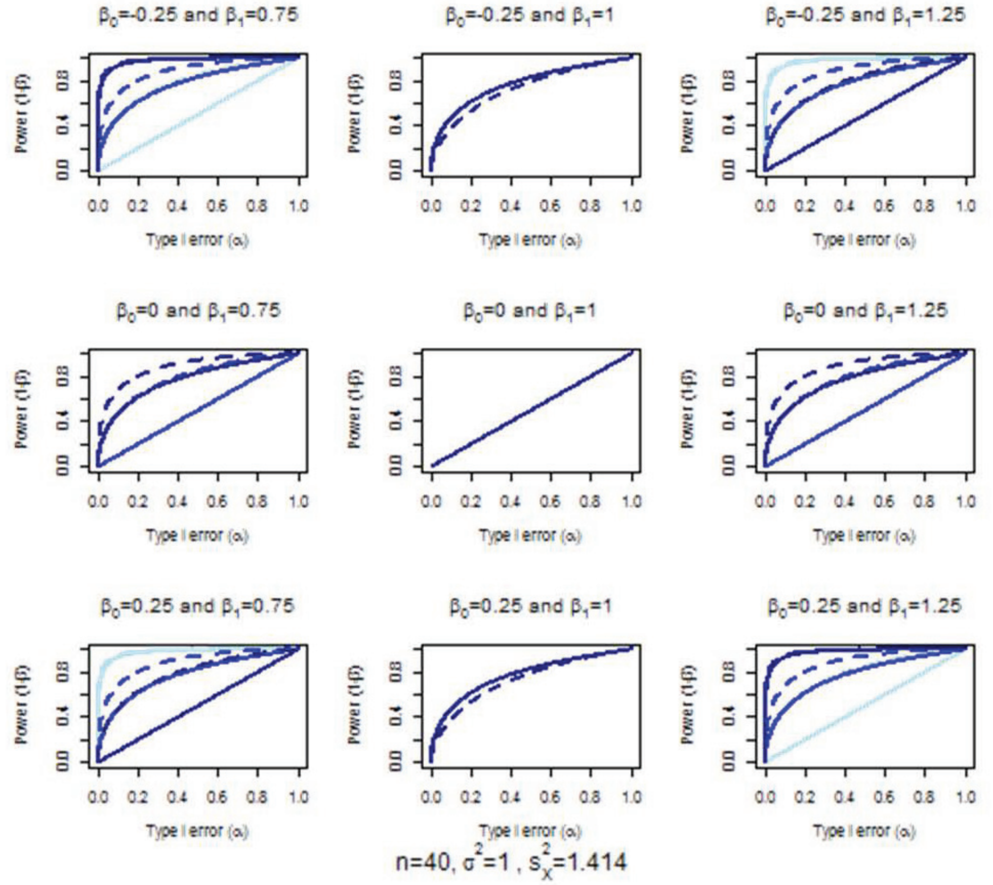

Figure 1. Comparison of the two degree-of-freedom (dashed) versus one degree-of-freedom, paired t test (solid) approach over $\overline{x} = -1,\ 0$, and 1 (light to dark)

Note. Underlying model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \sim N(0,\ \sigma^2)$; power and type I error correspond to the test of the null hypothesis that the proxy is not biased, which for the two degree-of-freedom approach is H_0: $\beta_0 = 0 \cap \beta_1 = 1$ and is H_0: $\beta_0 = -(\beta_1 - 1)\overline{x}$ for the one degree-of-freedom, paired t test approach.

4.2 Power as a Function of s_x^2

Sample variance (s_x^2) values for the proxy measures (X) varied within each plot in Figure 2, using values of 0.1, 1, and 2. In these plots, the power curves for the paired t test (one degree-of-freedom) approach completely overlaid one another. This followed from the fact that the non-centrality parameters for these models, λ^*, were not functions of s_x^2 [see (6)]. We again saw a slight power advantage with the paired t test (one degree-of-freedom) approach over that of the proposed method for the cases where no linear bias was present, and both power functions reverting to the type I error rate when the new measure was an unbiased proxy for the original. A slight advantage of the paired t test approach was seen in other scenarios when the sample variance of the proxy measure was close to zero; but this advantage was lost as the sample variance increased. For the case where the bias of the proxy was only in the linear term the proposed (two degree-of-freedom) approach always outperformed the paired t test (one degree-of-freedom) approach, with the greater advantage observed as the sample variance of the proxy (X) increased.

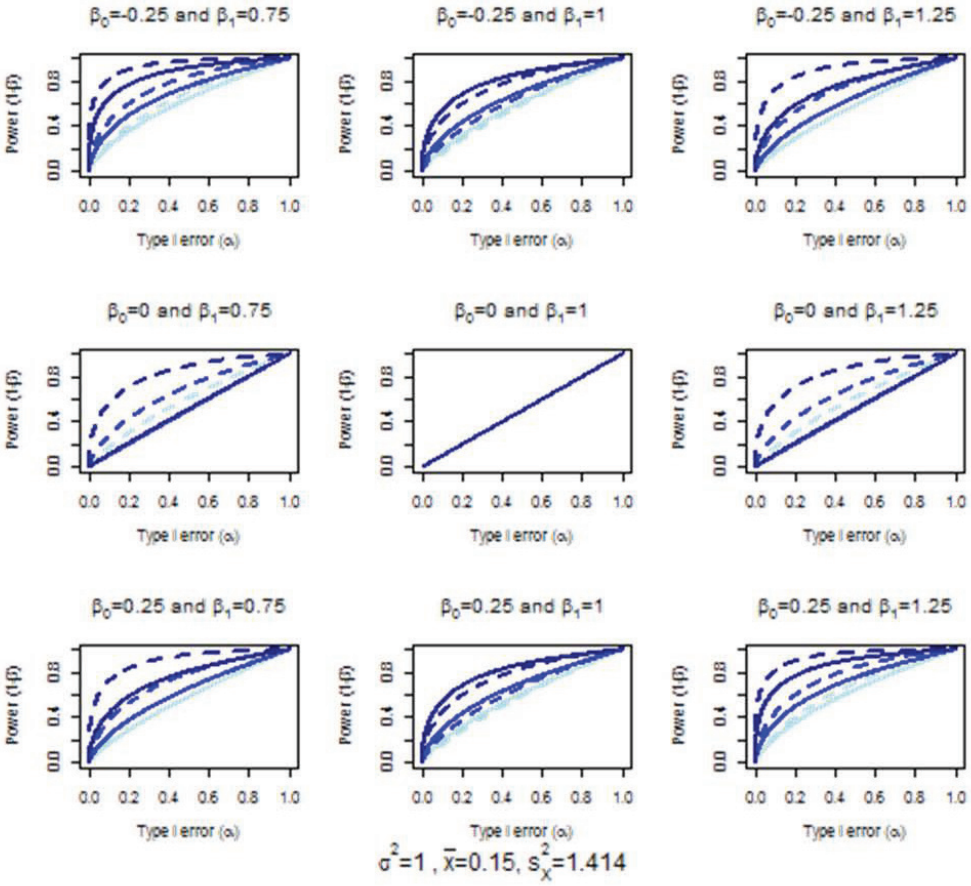

Figure 2. Comparison of the two degree-of-freedom (dashed) versus one degree-of-freedom, paired t test (solid) approach over $n = 10,\ 20$, and 50 (light to dark)

Note. Underlying model $y_i = \beta_0+\beta_1 x_i+\varepsilon_i$ where $\varepsilon_i \sim N(0,\ \sigma^2)$; power and type I error correspond to the test of the null hypothesis that the proxy is not biased, which for the two degree-of-freedom approach is H_0: $\beta_0 = 0 \cap \beta_1 = 1$ and is H_0: $\beta_0 = -(\beta_1 - 1)\overline{x}$ for the one degree-of-freedom, paired t test approach.

4.3 Power as a Function of n

For the comparison of these approaches over increasing sample sizes (n; Figure 3), the proposed approached had greater power in each scenario where a linear bias was present, compared to slight advantages with the pair t test (one degree-of-freedom) approach under the settings where there was no linear bias (i.e., $\beta_1 = 1$). Also, for the paired t test approach the power did not improve as a function of sample size when the bias was only in the linear term (i.e., when $\beta_0 = 0$, but $\beta_1 \neq 1$). As with the evaluations of power as functions of other parameters, when the proxy (X) truly was an unbiased measure of the original (Y), the type I error was preserved under either approach.

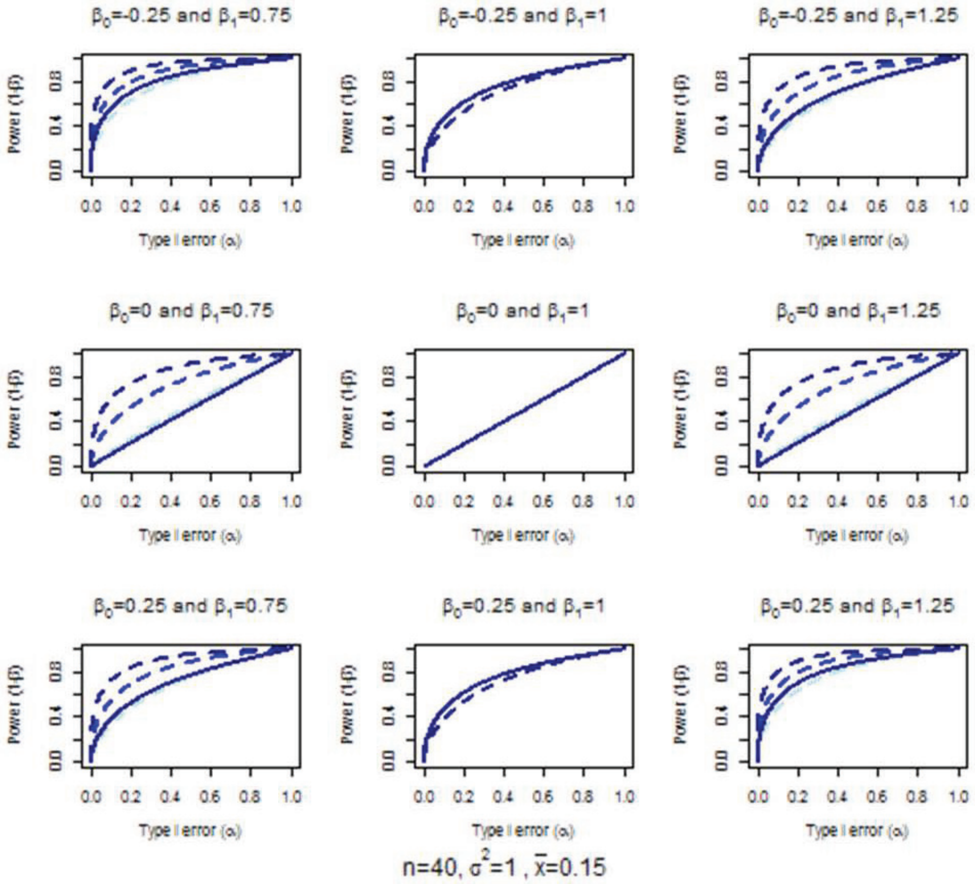

Figure 3. Comparison of the two degree-of-freedom (dashed) versus one degree-of-freedom, paired t tested (solid) approach over $s_x^2 = 0.1,\ 1$, and 2 (light to dark)

Note. Underlying model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \sim N(0,\ \sigma^2)$; power and type I error correspond to the test of the null hypothesis that the proxy is not biased, which for the two degree-of-freedom approach is H_0: $\beta_0 = 0 \cap \beta_1 = 1$ and is H_0: $\beta_0 = -(\beta_1 - 1)\bar{x}$ for the one degree-of-freedom, paired t test approach.

4.4 Assessment of the STEP Test as a Proxy for VO$_2$ Peak Fitness Measures

A scatter plot of the STEP test versus the VO$_2$ peak test was presented in Figure 4. The ordinary least squares regression line estimated from these data [of the (1) form] was plotted (solid), as was a $y = x$ reference line (dashed). This figure indicated a general bias toward the STEP test overestimating the fitness level of the VO$_2$ peak testing. The maximum likelihood estimates for the original scale (1) were

$$\widehat{\beta_0} = -5.37 \quad \text{and} \quad \widehat{\beta_1} = 1.04;$$

or equivalently on the linear transformed scale

$$\widehat{\gamma_0} = -5.37 \quad \text{and} \quad \widehat{\gamma_1} = 0.04.$$

The F test of the null hypothesis that the STEP measure was an unbiased proxy for VO$_2$ peak was rejected (p < 0.0001). In this example, the alternative, paired t test approach also rejected this null hypothesis (H_0: $\gamma^* = 0$; p < 0.0001).

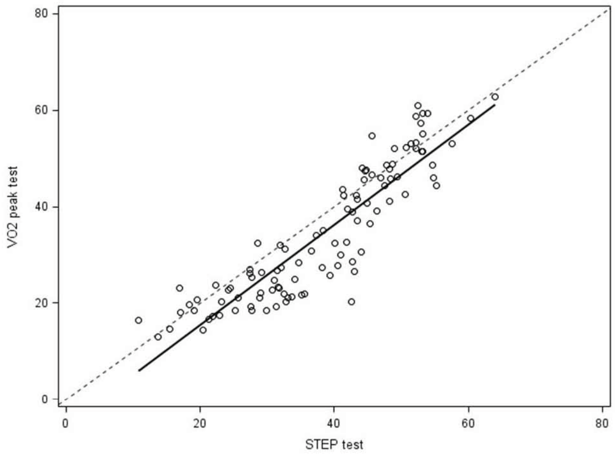

Figure 4. Scatter plot of the potential proxy STEP test measure by the true fitness measure VO_2 peak

5. Discussion

Proxy measures are important tools. Less costly alternative means to assess important measures have substantial importance in scientific research, clinical practice, business, and sociology, among other areas of application. A simple approach to evaluating a measure as a proxy is to test for differences in central tendency by applying the one sample paired t test to the observed differences. However, this approach can overlook important systematic forms of bias that, when left uncheck, could lead to poorly capturing of the underlying construct of interest by use of a proxy. Previous studies of methods for evaluating proxy measures has also been described by Bland and Altman (1999) and Dickenson et al. (2010).

Our investigations demonstrated that when linear biases were present the power to detect that a measure was a poor proxy was greatly reduced with the paired t test approach. Especially notable was the fact that when the bias was entirely on the slope parameter, increases in sample size did not correspondingly increase the power of the paired t test (one degree-of-freedom) approach for detecting that the proxy was invalid due to bias (see Figure 2 plots where $\beta_0 = 0$). While the cases where there was no linear bias in the proxy indicated slight increases in power, the magnitudes were much greater for changes when the bias did, in fact, include linear effects. This advantage followed from estimating only one mean parameter for the intercept rather than two–one for the intercept and one for the slope. We were unable to find an analytic solution to the probability that the alternative, one degree-of-freedom approach would have greater power than the two degree-of-freedom approach when no linear term was needed (i.e., $\beta_1 = 1$).

As noted above, for our motivating example the conclusions did not change with the differing approaches. However, the null hypothesis of the one degree-of-freedom paired t test approach was

$$H_0 : \gamma^* = 0 \Longleftrightarrow \gamma_0 + \gamma_1 \overline{x} = 0 \Longleftrightarrow \gamma_0 = -\gamma_1 \overline{x}.$$

This implied that even in the presence of strong linear bias, the null hypothesis using the one degree-of-freedom approach would not be rejected when the sample mean for the proxy (X) is nearly centered. This was highlighted by the results of Figure 1, where the one degree-of-freedom approach had power equal to the type I error for the scenarios: where $\overline{x} = -1$ for the plots where $\beta_0 = -0.25 \cap \beta_1 = 0.75$ and $\beta_0 = 0.25 \cap \beta_1 = 1.25$ (light blue solid lines); where $\overline{x} = 0$ for the plots where $\beta_0 = 0 \cap \beta_1 = 0.75$ and $\beta_0 = 0 \cap \beta_1 = 1.25$ (blue solid lines); and where $\overline{x} = 1$ for the plots where $\beta_0 = 0.25 \cap \beta_1 = 0.75$ and $\beta_0 = -0.25 \cap \beta_1 = 1.25$ (dark blue solid lines). In our motivating example $\overline{x} \approx 38$, so there are an infinite number of possible (β_0, β_1) pairs that could produce values with little power to reject the null hypothesis even in cases of highly biased, poor proxy measures. These pairs would be all those that fall along the line

$$\gamma_0 = -\gamma_1 \overline{x} \Longleftrightarrow \beta_0 = -(\beta_1 - 1)\, 38 \Longleftrightarrow \beta_0 = -38\beta_1 + 38.$$

Although there are an infinite number of pairs that fall on this line, these solutions would still represent a restricted subset of the entire two-dimensional space of all possible (β_0, β_1) pairs. We further note that samples with centered proxy measures (i.e., $\overline{x} = 0$) also suffer from a power disadvantage for detecting systematic linear biases. Thus, overall, we believe these findings support the use of the two degree-of-freedom approach advocated here over the risk of a slight loss of power due to using an $F_{2,n-2}$ distribution as opposed to an $F_{1,n-1}$ distribution (paired t test approach) for inference about whether a proxy is a biased measure of the original.

Acknowledgements

Drs. Mahnken and Vidoni are supported in part by The University of Kansas Alzheimer's Disease Center, which is funded by a grant from the National Institute on Aging (P30AG035982). Dr. Billinger is supported in part by K01HD067318. Dr. Vidoni is also supported in part by KL2TR000119, and Drs. Gajewski and Vidoni are supported in part by a Clinical and Translational Science Award from the National Center for Research Resources and National Center for Advancing Translational Sciences awarded to The University of Kansas Medical Center for Frontiers: The Heartland Institute for Clinical and Translational Research Clinical and Translational Science Unit (UL1 TR000001).

References

Billinger, S. A., van Swearingen, E., McClain, M., Lentz, A. A., & Good, M. B. (2012). Recumbent stepper submaximal exercise test to predict peak oxygen uptake. *Medicine and Science in Sports and Exercise, 44*(8), 1539-1544. http://dx.doi.org/10.1249/MSS.0b013e31824f5be4

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*, 135-160. http://dx.doi.org/10.1191/096228099673819272

Dickinson, H. O., Hrisos, S., Eccles, M. P., Francis, J., & Johnston, M. (2010). Statistical considerations in a systematic review of proxy measures of clinical behavior. *Implementation Science, 5*(20). http://dx.doi.org/10.1186/1748-5908-5-20

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York, NY: John Wiley & Sons, Inc. http://dx.doi.org/10.1002/9781118625590

Hawkins, D. M. (2002). Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine, 21*, 1913-1935. http://dx.doi.org/10.1002/sim.1013

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods* (3rd ed.). Pacific Grove, CA: Duxbury Press.

Kotz, S., & Johnson, N. L. (Eds.) (1986). *Encyclopedia of statistical sciences* (Vol. 7). New York, NY: John Wiley & Sons.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear regression models* (4th ed.). Boston, MA: McGraw-Hill Irwin.

Lin, LI-K. (1989). A concordance coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255-268. http://dx.doi.org/10.2307/2532051

Mahnken, J. D., Wick, J. A., Gajewski, B. J., & Mayo, M. S. (2010). A study design with conditional, serially assessed co-primary endpoints: An application to a single-arm, pilot non-Hodgkin's lymphoma trial. *Drug Development Research, 71*, 395-403. http://dx.doi.org/10.1002/ddr.20387

Petrella, R., Koval, J., Cunningham, D., & Paterson, D. (1998). Predicting VO_2max in community dwelling seniors using a self-paced step test. *Medicine and Science in Sports and Exercise, 30*(5), 76. http://dx.doi.org/10.1097/00005768-199805001-00433

Petrella, R. J., Koval, J. J., Cunningham, D. A., & Paterson, D. H. (2001). A self-paced step test to predict aerobic fitness in older adults in the primary care clinic. *Journal of the American Geriatrics Society, 49*(5), 632-638. http://dx.doi.org/10.1046/j.1532-5415.2001.49124.x

Petrella, R. J., & Wight, D. (2000). An office-based instrument for exercise counseling and prescription in primary care. The step test exercise prescription (STEP). *Archives of Family Medicine, 9*(4), 339-344. http://dx.doi.org/10.1001/archfami.9.4.339

Rencher, A. C. (2000). *Linear models in statistics*. New York, NY: John Wiley & Sons, Inc.

Trenkler, G., & Stahlecker, P. (1996). Dropping variables versus use of proxy variables in linear regression. *Journal of Statistical Planning and Inference, 50*, 65-75. http://dx.doi.org/10.1016/0378-3758(95)00045-3

Vidoni, E. D., Mattlage, A., Mahnken, J., Burns, J. M., McDonough, J., & Billinger, S. A. (2013). Validity of the step test for exercise prescription does not extend to a larger age range. *Journal of Aging and Physical Activity, 21*, 444-454.

Appendix

Proof of Equivalence of Inferences for $\widehat{\gamma_0} = \widehat{\beta_0}$ and $\widehat{\gamma_1} = \widehat{\beta_1} - 1$.

Let $Y_i|X_i = x_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$, so x_i is treated as a constant and $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim N\left(0,\ \sigma^2\right)$. The maximum likelihood estimators for β_0 and β_1 are

$$\widehat{\beta_0} = \frac{\sum_i^n y_i}{n} - \beta_1 \frac{\sum_i^n x_i}{n} = \overline{y} - \beta_1 \overline{x}$$

and

$$\widehat{\beta_1} = \frac{\sum_i^n x_i y_i}{\sum_i^n x_i^2} - \frac{\beta_o}{\sum_i^n x_i}.$$

For the transformed $Z_i|X_i$, $(Y_i - X_i|X_i = x_i) \sim N(\beta_0 + [\beta_1 - 1]\, x_i,\ \sigma^2)$, so

$$y_i - x_i = \beta_0 + [\beta_1 - 1]\, x_i + \varepsilon_i,$$

where $\varepsilon_i \sim N\left(0,\ \sigma^2\right)$. Let $\gamma_0 = \beta_0$ and $\gamma_1 = \beta_1 - 1$, then

$$y_i - x_i = \gamma_0 + \gamma_1 x_i + \varepsilon_i.$$

Now to show that $\widehat{\gamma_0} = \widehat{\beta_0}$,

$$\widehat{\gamma_0} = \frac{\sum_i^n (y_i - x_i)}{n} - \gamma_1 \frac{\sum_i^n x_i}{n}$$

$$\widehat{\gamma_0} = \frac{\sum_i^n y_i - \sum_i^n x_i}{n} - (\beta_1 - 1) \frac{\sum_i^n x_i}{n}$$

$$\widehat{\gamma_0} = \overline{y} - \overline{x} - (\beta_1 \overline{x} - \overline{x})$$

$$\widehat{\gamma_0} = \overline{y} - \beta_1 \overline{x};$$

so $\widehat{\gamma_0} = \widehat{\beta_0}$. Finally, to show that $\widehat{\gamma_1} = \widehat{\beta_1} - 1$,

$$\widehat{\gamma_1} = \frac{\sum_i^n x_i (y_i - x_i)}{\sum_i^n x_i^2} - \frac{\gamma_o}{\sum_i^n x_i}$$

$$\widehat{\gamma_1} = \frac{\sum_i^n x_i y_i - \sum_i^n x_i^2}{\sum_i^n x_i^2} - \frac{\gamma_o}{\sum_i^n x_i}$$

$$\widehat{\gamma_1} = \frac{\sum_i^n x_i y_i}{\sum_i^n x_i^2} - 1 - \frac{\beta_o}{\sum_i^n x_i}$$

$$\widehat{\gamma_1} = \frac{\sum_i^n x_i y_i}{\sum_i^n x_i^2} - \frac{\beta_o}{\sum_i^n x_i} - 1,$$

so $\widehat{\gamma_1} = \widehat{\beta_1} - 1$.

Copyrights