Educational and Psychological Measurement
72(5) 862–888

DOI: 10.1177/0013164411430225
http://epm.sagepub.com

The Assessment of Reliability Under Range Restriction: A Comparison of α, ω, and Test–Retest Reliability for Dichotomous Data

Dustin A. Fife[1], Jorge L. Mendoza[1], and Robert Terry[1]

Abstract

Though much research and attention has been directed at assessing the correlation coefficient under range restriction, the assessment of reliability under range restriction has been largely ignored. This article uses item response theory to simulate dichotomous item-level data to assess the robustness of KR-20 (α), ω, and test–retest under varying selection ratios. These estimators, both corrected and uncorrected for range restriction, were compared in terms of both bias and precision. Test–retest reliability was usually the best estimator of reliability across a variety of conditions. Only under indirect range restriction did KR-20 and ω performed well. All estimators suffered imprecision as a function of range restriction, above and beyond the reduction in sample size. Based on the results, a set of recommendations are proposed.

Keywords

classical test theory, coefficient alpha, coefficient omega, range restriction, reliability, test–retest reliability

Range restriction is a common occurrence in educational, governmental, and organizational settings. It occurs when there is less variance in the sample than in the population because of a selection procedure (Sackett, Laczo, & Arvey, 2002), such as

[1]University of Oklahoma, Norman, OK, USA

Corresponding author:
Dustin Fife, 455 W. Lindsey Street, Dale Hall Tower, Room 705, Norman, OK 73019, USA
Email: dfife@ou.edu

self-selection, response bias, personnel selection, attrition, and so on. The prevalence of range restriction is unfortunately difficult to estimate because often the variance of the unselected population is unobtainable. For example, if the variance of a sample has been attenuated through self-selection, the selection variable is unmeasured, which would make selection ratio calculations impossible. However, a few studies have used various means of estimating the distribution of selection ratios (e.g., Alexander, Carson, Alliger, & Cronshaw; 1989; Schmidt & Hunter, 1977). For example, Schmidt and Hunter (1977) did a general (nonquantitative; Schmidt, Oh, & Le, 2006) review of the literature and estimated that the average ratio of restricted to unrestricted variance is .59, meaning that one can expect only 59% of the unrestricted variance to be present in the selected sample. Alexander et al. (1989), on the other hand, estimated selection ratios empirically. They did so by comparing sample standard deviations with published standard deviations for standardized tests. They found selection ratios ranging from approximately .7 to .91, depending on the domain. Although an exact selection ratio is impossible to obtain, these studies show that range restriction is a common problem and is likely more serious than what can be empirically investigated (Alexander et al., 1989).

Range restriction becomes problematic when researchers seek to estimate reliability or validity coefficients from selected samples. Because the variance is altered under range restriction, and because both correlation and reliability coefficients rely on variances for their computations, these coefficients are generally also affected. Indeed, Pearson (1903) stated that not only does range restriction ''determine the amount of correlation, but that it is probably the chief factor in the production of correlation'' (p. 2).

Pearson (1903) was one of the first to discuss range restriction and provided a ''correction'' for range restriction. This correction seeks to estimate the unrestricted or unselected correlation coefficient. Since 1903, a wealth of research attention has been directed at estimating unrestricted correlation coefficients (e.g., Alexander et al., 1989; Bobko, 1983; Dunbar & Linn, 1991; Lord & Novick, 1968; Sackett & Yang, 2000; Thorndike, 1949). What has received much less attention is the estimation of *reliability* under range restriction. Indeed, many researchers have derived various statistical procedures assuming that restricted reliability estimates are unbiased (e.g., Schmidt et al., 2006; Raju & Brand, 2003). Sackett et al. (2002) have addressed the issue of estimating restricted reliability but limited their investigation to interrater reliability. Additionally, their study focused on the population and did not address estimation bias or standard errors. Because the variance of corrected estimators are generally larger than the variance of the uncorrected ones (Gross & Kagen, 1983), standard errors are also important to consider. Our study addresses estimation as well as standard errors.

In this article, we add to the work done by Sackett et al. (2002) by addressing other estimates of reliability. Our study focuses on measures that are dichotomously scored because they are found in many tests that are used for selection (e.g., cognitive tests). Additionally, we explore the sampling distribution of restricted reliability estimates,

both corrected for range restriction and uncorrected. We show that both raw and corrected reliability estimates are often biased and/or imprecise.

Reliability

Reliability refers to the stability of a measurement, and the degree to which a measurement randomly varies is an indication of measurement error (Cortina, 1993; Nunnally, 1967). The classical test theory definition of reliability (Lord & Novick, 1968) is given by

$$\rho_{xx} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}, \tag{1}$$

where σ_T^2 is the true score variance and $\sigma^2{}_E$ the error variance. Or, equivalently, $\rho_{xx} = \sigma_T^2/(\sigma_X^2)$, where σ_x^2 is equal to total test variance, assuming that $X = T + E$.

The classical test theory definition of reliability makes two key assumptions: (a) true scores and errors are assumed to be independent and (b) the estimated reliability coefficient is bounded between zero and one (McDonald, 1999). We show that under direct range restriction, each of these assumptions is not tenable.

Several methods are used for estimating reliability such as test–retest, parallel forms, split-half (Brown, 1910; Spearman, 1910), KR-20 (coefficient α; Cronbach, 1951; Guttman, 1945), and McDonald's ω (McDonald, 1970, 1999). In this article, we limit our discussion of reliability to three estimators: KR-20 (which we will refer from now on as coefficient α), coefficient ω, and test–retest.

Coefficient α

Coefficient α is the more general equation for the Kuder–Richardson Formula 20 (KR-20; Kuder & Richardson, 1937). It is intended to be a measure of, to use Cronbach's (1951) words, equivalence. A coefficient of equivalence measures how much two measures of the same factor agree (Cronbach, 1951), or as is the case with α, how much multiple measures (i.e., items) of the same factor agree. However, as pointed out by Cortina and others (Cortina, 1993; McDonald, 1999; Sijtsma, 2009), a measure may have a high α coefficient and not be unidimensional. Therefore, α is simply a measure that indicates the degree to which a test is free from error (Cortina, 1993; Revelle & Zinbarg, 2009).

The α coefficient is a lower bound to reliability (Cortina, 1993; Guttman, 1945; Lord & Novick, 1968; McDonald, 1999; see also Komaroff, 1996), with equivalence being achieved only under the τ-equivalent model, where item variances and covariances are equal.

We have elected to use coefficient α (KR-20) in our study because it is readily available in most software packages and because it is the most popular estimate of reliability in the psychological literature (Cortina, 1993; Hogan, Benjamin, & Brezinski, 2000). Indeed, between 1966 and 1997, Cronbach's (1951) article had

been cited approximately 60 times per year (Cortina, 1999). Hogan et al. (2000) estimated that 70% of psychological measures use α or KR-20 to estimate reliability. Our study focuses on KR-20, or the dichotomous form of α.

Coefficient ω

McDonald's ω[1] (McDonald, 1970, 1999) is not a popular estimator of reliability; of the 696 tests sampled in the Hogan et al. (2000) study, only three tests (0.4%) used coefficient ω. However, we have elected to use ω because it is said to be a higher lower bound than α (McDonald, 1999), and therefore may be more robust under range restriction.

Coefficient ω is defined as the ratio of the squared sum of factor loadings to total variance for a homogeneous test (McDonald, 1970, 1999). Or, equivalently, ω is defined as

$$\omega = \frac{\left(\Sigma\lambda_j\right)^2}{\left(\Sigma\lambda_j\right)^2 + \Sigma\psi_j^2}, \tag{2}$$

where λ_j is the factor loading for the *j*th item and $\Psi 2_j$ is the error variance for the *j*th item.

According to McDonald (1999), ω is a lower bound to population reliability, and α is a lower bound to ω. ω is equal to population reliability only when the measure is unidimensional (i.e., when there is only one factor in a measure), and α is equal to ω only when all the factor loadings are of equal value (McDonald, 1999). α and ω both equal the population reliability when the items are τ-equivalent (McDonald, 1999). For this study, we have limited our study of reliability to near τ-equivalent models so that we could easily assess the performance of each estimator of reliability; under τ-equivalence, each estimator should be equivalent.

Test–Retest

We have also included test–rest reliability in our study for two reasons: first, although it is not as popular as α, it is still frequently used; 19% of the tests sampled by Hogan et al. (2000) used test–retest as an estimate of reliability. Second, test–retest does not assume independence between true scores and error scores in order to be estimated, which is important under range restriction (Mendoza & Mumford, 1987).

Reliability Estimation Under Range Restriction

Most of what is known about range restriction comes from the body of literature that has sought to estimate unrestricted correlation coefficients. Many researchers have adopted the nomenclature of Pearson (1903), who classified restriction as either direct or indirect. Direct range restriction occurs when selection is made on the basis

of a predictor score, X. Direct range restriction alters the variance of X, which affects both true and error variances, and creates a correlation between true and error scores (Mendoza & Mumford, 1987). Indirect range restriction occurs when the variance of a variable is altered, not through selection of the variable itself, but through the selection of a variable with which it is correlated. For example, if selection is made on X, and X is correlated with Y, then Y will be *indirectly* restricted. Indirect range restriction, in contrast to direct range restriction, only affects the variance of the true score. In this article, we investigate the effects of direct and indirect range restriction on reliability estimates. We will also investigate selection under a double-hurdle situation. Double-hurdle selection occurs when individuals are directly selected on two variables, such as GRE scores and GPA. We have chosen double-hurdle selection because it is a common selection procedure (Roth, Bobko, Switzer, & Dean, 2001). We will see that the type of selection has differential effects on estimates of reliability as well as the variances.

Estimating the unrestricted reliability from restricted reliability has typically been done using the following equation (Kelley, 1921; Lord & Novick, 1968):

$$\rho_{xx} = 1 - \frac{\sigma'^2_X}{\sigma^2_X}(1 - \rho'_{xx}). \tag{3}$$

Equation 3 requires both the unrestricted and restricted variances. This equation can also be used to estimate the reliability of Y. When the unrestricted variance of Y is not known, which is common, it is estimated with

$$\sigma_Y = \sigma'_Y \left(1 - r'^2_{xy} + r'^2_{xy} \frac{\sigma^2_X}{\sigma'^2_X} \right). \tag{4}$$

Equation 3 makes two important assumptions (Kelley, 1921; Lord & Novick, 1968):

1. True scores and errors remain independent after range restriction.
2. $\sigma_E = \sigma'_E$, or the size of the residual variance, is unaffected by selection.

When estimating population reliability for a variable that has been indirectly selected (i.e., Y), Assumptions 1 and 2 are not problematic. However, under direct range restriction, true scores and errors are correlated (Mendoza & Mumford, 1987) violating the first assumption. Consequently, this correction is likely to yield biased results, especially when the selection ratio is small. The second assumption fails when selection is made on X, because X is a composite of T and E. The effect is to reduce the error variance, σ'_E. Lord and Novick (1968) note that when $\sigma'_E < \sigma_E$, Equation 3 will overestimate ''global'' (unrestricted) reliability.

There is, however, a correction that does not rely on the assumption of independence or equal error variance:

Table 1. Correction Formulas Used for Varying Conditions of Range Restriction

Selection Type	α	ω	Test–Retest
Direct	Equation 3	Equation 3	Equation 5
Double-hurdle	Equation 3	Equation 3	Equation 3
Indirect	Equation 3, Equation 4	Equation 3, Equation 4	NA

$$\rho_{xx} = \frac{\mathrm{cov}(x_1, x_2)}{\sigma^2_{x_1}}. \quad (5)$$

However, this estimator requires a second administration (see Stauffer & Mendoza, 2001). This estimator relies on the assumption that the slope of the regression equation between X_1 and X_2 remains the same under direct range restriction (Stauffer & Mendoza, 2001).

In this article, we use Equation 3 to correct α and ω for range restriction (direct, indirect, and double hurdle). In the case of test–retest, Equation 3 will only be used under indirect and double-hurdle range restriction. When the restriction is direct, we will use Equation 5 to correct the test–retest estimator. When Equation 3 was used in the indirect case, the variance of Y was estimated using Equation 4.[2] See Table 1 for a summary.

Local Versus Global Reliability

In this article, a distinction is made between reliability in the unselected (entire) population and reliability in the selected (restricted) population. Note that the selected population is a subset of the entire population. We refer to the reliability of the test in the entire population as global reliability, whereas local reliability is the reliability of the test in the selected population. Since the selected population is generally going to have a larger mean and smaller variance than the unselected population, these reliabilities are different. Generally speaking, global reliability is the reliability of the test within the population of applicants, and local reliability is the reliability of the test within the population of incumbents (i.e., those who were admitted).

Most uncorrected estimates of reliability are assumed to be local reliability estimates (unless the researcher's sample is truly random). This estimate may or may not approximate population local reliability well. Likewise, when the researcher corrects the local estimate for range restriction, they are estimating global reliability. This too may or may not be biased. In this article, we investigate bias at both levels: local and global. We make this distinction because local reliability can be of interest on its own. In situations where an instrument is only administered and used in a restricted population, local reliability is more relevant than global reliability. For example, if an institution develops a performance evaluation for graduate students (who are selected undergraduates), they would not be interested in the reliability of

the instrument in the population of students containing both graduate and undergraduate students. Similarly, if we develop a measure of job performance to be used with incumbents, we would not be interested in the reliability of the measure in the applicant population. Consequently, investigation of bias at both levels is important.

To mathematically define local reliability, it is important to remember the effects of direct versus indirect selection on the variances of T and E. When selection is indirect, local reliability is simply the ratio of restricted true score variance over total variance, (σ'_T/σ'_X). This is because the variance of E is unaffected by selection, only the variance of T. However, under direct range restriction, selection is made on the basis of observed scores (X), which creates a correlation between T and E. Because of this correlation, the observed variance is no longer equal to the sum of true score variance and error variance, making the definition of local reliability more difficult. Therefore, local reliability can be conceived of in two ways:

$$\rho' = \frac{\sigma'^2_T}{\sigma'^2_T + \sigma'^2_E + 2 \times \mathrm{cov}(T,E)} = \frac{\sigma'^2_T}{\sigma'^2_X} \tag{6a}$$

or

$$\rho' = \frac{\sigma'^2_T}{\sigma'^2_T + \sigma'^2_E}. \tag{6b}$$

The problem with using Equation (6a) as the definition of local reliability is that selection on X reduces the variance of X faster than it reduces the variance of T. As X becomes more selected, estimates of reliability begin to exceed unity, which is theoretically impossible. Thus, we use Equation (6b) as our definition of local reliability. Note that when selection is on T, the unrestricted variance of E (σ^2_E) replaces the restricted variance σ'^2_E in Equation (6b).

Item Response Theory

Item response theory (IRT) is an alternative to classical test theory. IRT gives the probability of a correct response given the item parameters (a and b) and the ability of the individual (θ) as

$$p(\mathrm{correct}|\theta) = \frac{1}{1 + e^{-1.7 \times a \times (\theta - b)}}. \tag{7}$$

(The 1.7 in Equation 7 is used to approximate the normal parameterization.) Within the IRT framework, the standard error of measurement is a function of the test's discrimination parameter and the ability level θ. In general, the steeper the slope α, the smaller the standard error of measurement for a given θ and the higher the precision of the ability estimate. Because the standard error of measurement is conditional on the ability of the individual, θ, the IRT model does not provide a single measure of

reliability. However, an equivalent measure can be found by averaging over the distribution of ability.

Although a ''reliability'' measure can be obtained within the framework of IRT, that is not the purpose of this study. This study focuses on the classical test theory (CTT) definition of reliability, because most reliability measures found in the literature are based on CTT. Unfortunately, many of the CTT estimates of reliability (e.g., α and ω) require item-level data to be estimated. Fortunately, IRT provides a method for generating item-level data and in particular for generating dichotomous items (which are frequently used in cognitive tests with correct/incorrect choices). Consequently, we use IRT to *generate* item-level data and then use CTT to *estimate* reliability.

Method

Monte Carlo Simulation

To simulate the sampling distributions of each reliability estimator, the following steps were performed.

Step 1: Generating item parameters under an IRT model. Twenty items were generated with the idea of simulating a dichotomously scored cognitive ability test. We chose three a (discrimination) values—0.45, 0.65, and 1.69—that resulted in three levels of reliability (low = .7, medium = .8, high = .94).[3] We chose these levels of reliability following Nunnally's (1967) reliability benchmarks: .7 for tests in development, .8 for basic research, and .90 to .95 for instruments that are used to make important decisions. In addition, the simulated reliabilities match common reliabilities found in the research literature, with the majority of estimates (approximately 81%) falling between .7 and .94 (Hogan et al., 2000). We also limited the number of items to 20 because the median number of items in the research literature is approximately 20 (Hogan et al., 2000).

Variability was added to the slopes (a) in order to simulate the behavior of items in a real test. However, the variability was sufficiently small so that the items were essentially τ-equivalent. The b parameters for each test were sampled from a normal distribution with a mean of zero and a standard deviation of 0.25. We limited the variability of the b parameter because it is common practice in testing to eliminate items that are either very easy or very difficult. Note that the IRT function was only used to simulate sample responses. These responses were then used to compute the reliability of the test.

Step 2: Generating ability levels for the predictor and criterion. To generate a sample of abilities on the predictor and criterion (θ_x and θ_y, respectively), we sampled $n = 200$ values from a bivariate normal distribution with mean zero and variance one. The correlation between the two θ values was set to .6.

Step 3: Generating the subject's response to each item on the test. To simulate responses for each of the 20 items, we first obtained the probability of getting the item correct given θ_x and θ_y using Equation 7. Once we obtained the probability of a

correct response we drew a uniform random number on the 0 to 1 interval; if the random number exceeded the probability value, the response was coded as a 0 (*incorrect answer*), and if the random number did not exceed the probability, the response was coded as a 1 (*correct answer*). Total scores (X_{tot} and Y_{tot}) were then calculated for each subject by summing responses across all 20 items. In addition, true scores (T_{xtot} and T_{ytot}) were calculated for each subject by summing up the probabilities from Equation 7 across the items. Responses for X' and Y' (the retest portions) were simulated in the same manner as for X and Y, respectively. This procedure was repeated for each subject and test.

Step 4: Computing the population reliability. Global (unrestricted) population reliability was computed on a sample of 10,000 subjects. It was calculated using the classical test theory definition of reliability (Equation 1) where σ_T is defined as the variance of T, and $T = \sum_{j=1}^{j=20} p_j(\text{correct} \mid \theta)$. E as usual was defined as $X_{ij} - T_{ij}$.

Next, local population reliability parameters were obtained using a net sample size of 10,000 (after selection). Selection was performed by systematically selecting performers on the basis of X from the top 90th percentile, then 80th percentile, then 70th, and so on, until only the top 10th percentile remained. For double-hurdle selection, the same basic procedure was followed except the subject's score had to fall in the target percentile in both X and Y. Reliability was computed based on the variance that remained after the selection. Again σ'_T was defined in the same was as σ_T, except that it was computed on the restricted population.

We also computed the ratio of standard deviations for each selection ratio,

$$U = \frac{\sigma'_x}{\sigma_x}. \tag{8}$$

U is a common metric in the range restriction literature, allowing comparisons across studies. Furthermore, the validity generalization literature has discussed possible U distributions (Alexander et al., 1989; Schmidt & Hunter, 1977), which can be used to assess how frequently a particular selection ratio is expected to occur.

Step 5: Computing KR-20 (α), test–retest, and ω in the unrestricted and restricted samples. After the binary responses for X and Y were simulated in Step 3, selection was performed on the basis of the total test score. At each level of restriction α, ω, and test–retest reliability were calculated. For ω, we bounded the communalities to be less than one to avoid inflating ω. We also corrected each of the estimates for range restriction using Equation 3 for α and ω (for direct, indirect, and double hurdle). The test–retest reliability (direct only) was corrected using Equation 5 (see Table 1). The computations were repeated 10,000 times, while tracking of the means and variances of the reliability estimates.

Step 6: Assess bias. The mean uncorrected estimate of the sampling distribution was compared with local population reliability while corrected estimates were compared with global population reliability. These comparisons were made numerically and graphically. Bias was computed as

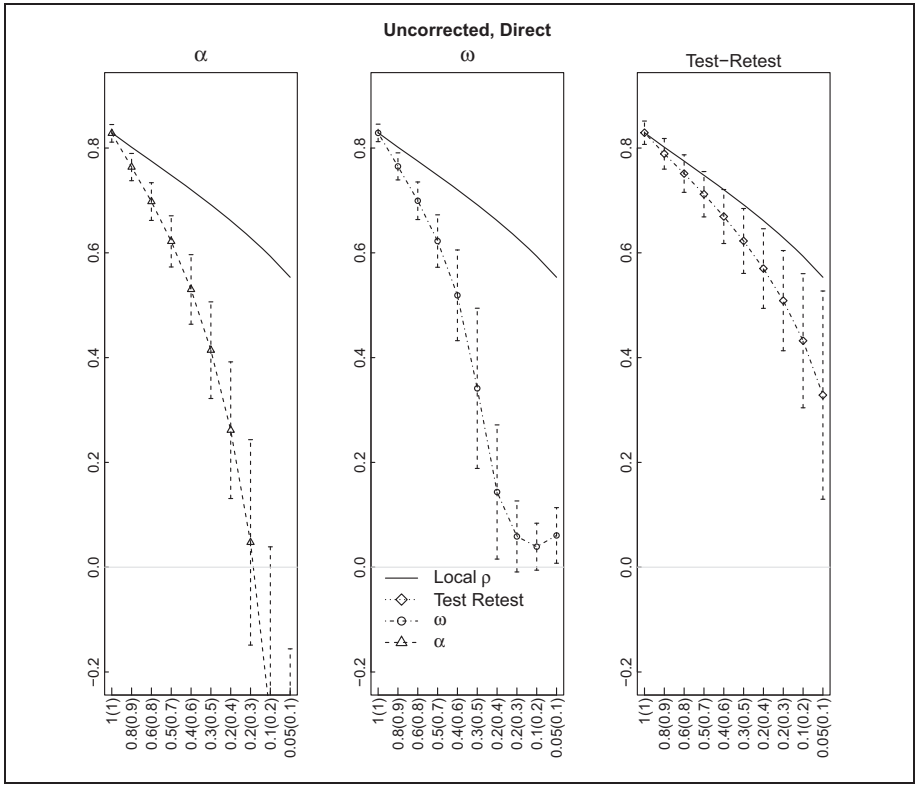

Figure 1. Uncorrected estimates of local reliability for direct range restriction
Note: *y*-axis, reliability estimate; *x*-axis, restriction amount (in parentheses) and corresponding *U* value. Each estimate is banded with a 95% confidence interval.

$$\text{Bias} = 100 \times \frac{\text{estimate} - \text{parameter}}{\text{estimate}}. \quad (9)$$

To simplify interpretation, we selected somewhat arbitrary bias thresholds of −.10 and .02 for underestimation and overestimation, respectively. Anything that exceeded these thresholds we noted as practically significant. The graphical comparison was done by plotting the average ω, α, and test–retest alongside the population reliabilities.

Step 7: Assess precision. The standard deviations (standard error) of the 10,000 sample estimates were computed for α, ω, and test–retest (corrected and uncorrected) at each level of restriction to assess the degree of variability in estimation.

Results

Uncorrected Population Estimates

Direct range restriction. Figure 1 shows that under direct range restriction all three estimates are biased and imprecise when estimating local reliability (the solid,

concave downward line), particularly α and ω. (Note: because the results were similar across all reliability conditions, only mid reliability results are shown.) In many cases when the selection ratio was small, α was negative. ω did not go negative because the communalities were bound to be less than one in the study. Under some selection ratios, α was larger than ω. This was unexpected because α is supposed to be a lower bound to ω (McDonald, 1999). Furthermore, the standard errors increased substantially in size as restriction increased. Test–retest was also biased and imprecise when estimating local population reliability but not to the degree of α and ω.

Table 2 shows the percentage of bias in estimating local reliability as a function of reliability (low, medium, and high) and selection ratio for each local estimator. α and ω showed significant bias (i.e., exceeded the 10% threshold for underestimation) as early as a selection ratio of .9. These estimators showed less bias as reliability increased, but even at the highest reliability these estimators exceeded the −.10 threshold at a selection ratio of .5. Test–retest was much more robust, but it showed significant bias at selection ratios of .7, .4, and .3 for low, medium, and high reliability, respectively. From these results it is clear that under severe direct range restriction local reliability cannot be estimated accurately. In situations when local reliability must be estimated under direct range restriction, test–retest reliability is preferable, showing the least biased.

Double-hurdle range restriction. Figure 2 shows local reliability estimates under double-hurdle selection. Notice how Figure 2 look nearly identical to Figure 1. The effects of double-hurdle selection mirror those of direct range restriction: all estimators estimated local reliability poorly.

Indirect range restriction. Figure 3 shows that all estimators estimated local reliability quite well under indirect range restriction. The amount of bias increased as restriction became more severe, but it was negligible. Although the standard errors also increased as restriction increased, the increase was not large (i.e., not greater than .1) until the selection ratio reached .1. Again, test–retest approximated local reliability slightly better than the other estimators. In addition, Table 3 shows that none of the estimators exceeded the 10% bias threshold.

Corrected Population Estimates

Direct/double-hurdle range restriction. Figure 4 shows the corrected reliability estimators under direct range restriction. None of the estimators consistently estimated global reliability (global reliability is the solid horizontal line). Corrected α and ω estimates underestimated for low to moderate range restriction and overestimated for high range restriction. It seems that Equation 3 is not robust to violations of independence. ω showed the most bias, followed by α then test–retest. The standard errors actually *decreased* as restriction became more severe, but this is only because the corrected estimates were bounded to be less than one. The corrected test–retest estimator (using the Equation 5) did moderately well until restriction became severe (selection ratios $< .3$). However, it had the largest standard errors of the three estimators. Table 4 shows the percentage of bias for corrected reliability estimates under

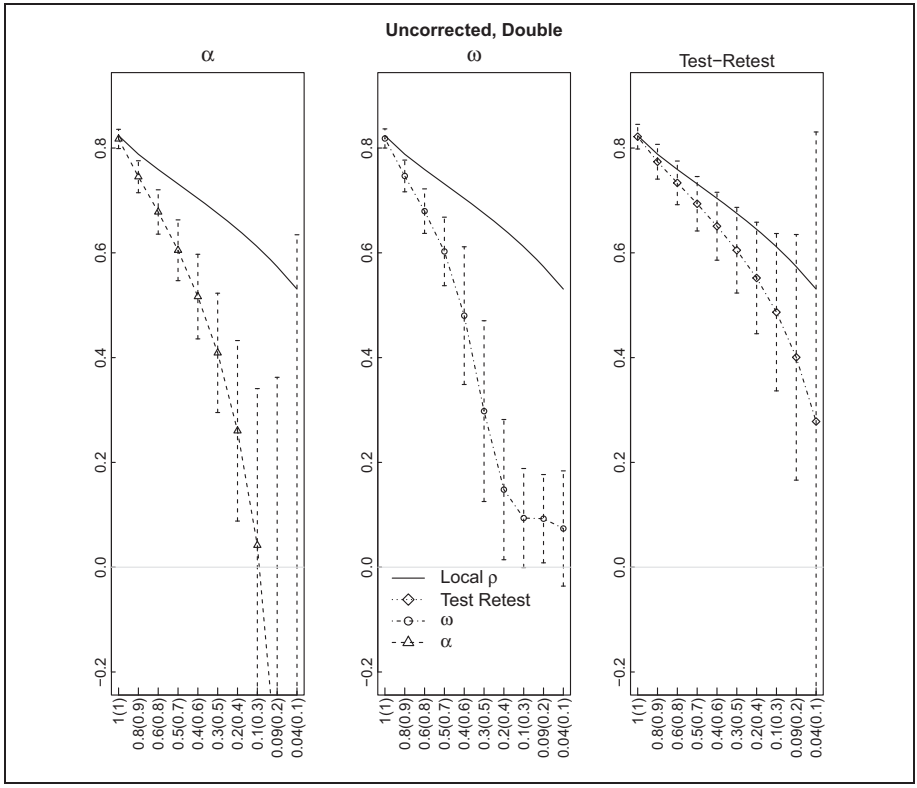

Figure 2. Uncorrected estimates of local reliability for double-hurdle direct range restriction
Note: *y*-axis, reliability estimate; *x*-axis, restriction amount (in parentheses) and corresponding *U* value. Each estimate is banded with a 95% confidence interval.

direct range restriction, again showing that none of the estimators estimated global reliability accurately when the selection ratio was smaller than .3.

Corrected double-hurdle estimates (Figure 5) mirrored the corrected direct range restriction estimates. With the exception of test–retest reliability, all estimators underestimated under low to moderate (1-.6) restriction, then overestimated with selection ratios of less than .6. The standard errors also mirrored the behavior of those under direct range restriction, increasing with range restriction.

Indirect range restriction. Figure 6 shows the corrected estimators under indirect range restriction. The corrected estimators under indirect range restriction approximated global reliability accurately, much better than those under direct range restriction. All estimators estimated global reliability well until the selection ratios were less than .2 when they overestimated. Test–retest overcorrected the most of the three estimators. Finally, the standard errors also were acceptable, not exceeding .2 (with the exception of ω). Table 5 shows that all corrected estimates safely estimated global reliability until restriction became quite severe (selection ratios less than .2), rarely even exceeding 1% in bias.

Table 2. Percentage of Bias of Each Estimate for Local Reliability Under Direct Range Restriction

	α			ω			Test–retest		
Selection ratio	Low (.7)	Medium (.8)	High (.94)	Low (.7)	Medium (.8)	High (.94)	Low (.7)	Medium (.8)	High (.94)
1	−0.18	−0.14	−0.2	0.05	−0.04	−0.18	0.06	0	0.01
.9	−12.36[a]	−4.73	−0.58	−12.02[a]	−4.56	−0.56	−4.01	−1.57	−0.09
.8	−24.57[a]	−9.96	−1.19	−25.03[a]	−9.76	−1.21	−7.28	−3.07	−0.24
.7	−38.98[a]	−16.89[a]	−2.36	−45.22[a]	−16.8[a]	−2.52	−10.6[a]	−4.86	−0.52
.6	−56.67[a]	−26.44[a]	−4.82	−71.19[a]	−27.99[a]	−5.36	−14.09[a]	−7.12	−1.12
.5	−79.26[a]	−40.11[a]	−10.32[a]	−87.48[a]	−50.63[a]	−11.91[a]	−17.96[a]	−9.96	−2.47
.4	−109.17[a]	−60.47[a]	−24.32[a]	−93.65[a]	−78.3[a]	−34.14[a]	−22.56[a]	−13.83[a]	−5.74
.3	−152.09[a]	−92.47[a]	−61.68[a]	−95.65[a]	−90.67[a]	−75.26[a]	−28.01[a]	−19.11[a]	−13.91[a]
.2	−220.26[a]	−148.67[a]	−146.26[a]	−96.12[a]	−93.41[a]	−81.52[a]	−34.69[a]	−27.16[a]	−33.16[a]
.1	−363.5[a]	−270.2[a]	−124.13[a]	−92.9[a]	−89.04[a]	−99.49[a]	−45.76[a]	−40.65[a]	−51.52[a]

[a]Exceeds the 10% underestimation threshold.

Table 3. Percentage of Bias of Each Estimate for Local Reliability Under Indirect Range Restriction

	α			ω			Test–retest		
Selection ratio	Low (.7)	Medium (.8)	High (.94)	Low (.7)	Medium (.8)	High (.94)	Low (.7)	Medium (.8)	High (.94)
1	−0.23	−0.16	−0.2	0	−0.06	−0.19	0	−0.01	0.01
.9	−0.28	−0.2	−0.21	0	−0.07	−0.19	−0.01	−0.02	0.01
.8	−0.33	−0.23	−0.22	−0.01	−0.09	−0.19	−0.02	−0.02	0.01
.7	−0.34	−0.29	−0.23	0.02	−0.11	−0.19	−0.01	−0.03	0.02
.6	−0.47	−0.37	−0.26	−0.05	−0.15	−0.2	−0.08	−0.04	0.03
.5	−0.68	−0.41	−0.28	−0.22	−0.14	−0.2	−0.12	−0.01	0.04
.4	−0.97	−0.51	−0.32	−0.5	−0.17	−0.2	−0.23	0.01	0.05
.3	−1.48	−0.72	−0.35	−1.28	−0.27	−0.17	−0.33	0.03	0.09
.2	−2.35	−1.08	−0.5	−3.78	−0.56	−0.22	−0.43	0.1	0.12
.1	−5.21	−2.54	−1.09	−11.78[a]	−4	−1.23	−0.58	0.15	0.32

[a]Exceeds the 10% underestimation threshold.

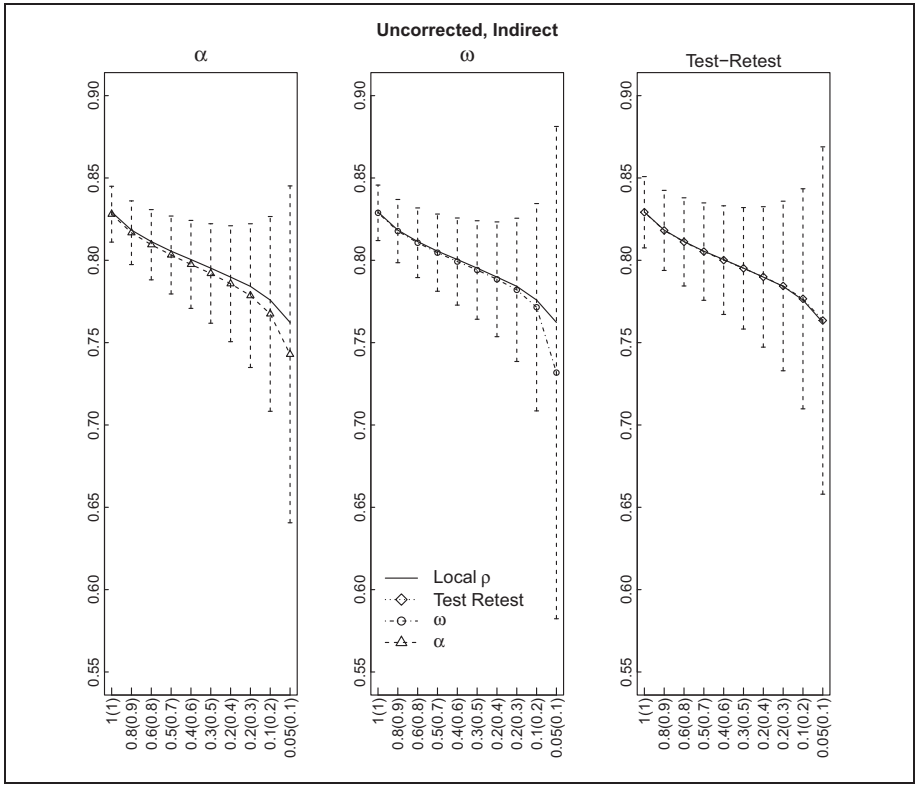

Figure 3. Uncorrected estimates of local reliability for indirect direct range restriction
Note: *y*-axis, reliability estimate; *x*-axis, restriction amount (in parentheses) and corresponding *U* value. Each estimate is banded with a 95% confidence interval.

Standard Errors as a Function of Sample Size Versus Instability

In the previous sections, we saw that standard errors increased as restriction increased. What is not known is whether this increase in standard errors is due solely to a decrease in sample size or whether range restriction increases the instability of estimates above and beyond the reduction in sample size. To test whether range restriction increased standard errors by itself, we increased the initial sample size from 200 to 500 and repeated the simulation. We then compared the standard errors of both sample size distributions for cases where the net sample sizes were equal. For example, a selection ratio of .4 with an original sample size of 500 yields the same net sample size (n = 200) as a selection ratio of 1 (*no selection*) and an original sample size of 200. If the standard errors were to differ, then we would conclude that range restriction increases standard errors beyond just the reduction in sample size. We also compared the obtained standard errors with the expected standard error,

$$se = \sqrt{\frac{2(1-\rho)^{2p}}{n(p-1)}}, \tag{10}$$

Table 4. Percentage of Bias of Each Corrected Average Estimate for Global Reliability Under Direct Range Restriction

	α			ω			Test–retest		
Selection ratio	Low (.7)	Med (.8)	High (.94)	Low (.7)	Med (.8)	High (.94)	Low (.7)	Med (.8)	High (.94)
1	−0.18	−0.14	−0.2	0.05	−0.04	−0.18	0.14	0.06	0.03
.9	−1.76	−1.3	−0.68	−1.52	−1.18	−0.67	0.54	0.55	0.33
.8	−2.2	−1.7	−1.03	−2.44	−1.59	−1.05	0.42	0.44	0.37
.7	−2.06	−1.61	−1.09	−4.73	−1.57	−1.16	0.05	0.07	0.15
.6	−1.42	−1.08	−0.75	−6.63	−1.55	−0.9	−0.39	−0.55	−0.32
.5	−0.29	−0.13	0.02	−3.02	−2.58	−0.21	−0.98	−1.31	−1.44
.4	1.43	1.32	1.15	4.02[a]	−1.97	0.61	−1.93	−2.45	−3.13
.3	3.92[a]	3.35[a]	2.48[a]	11.8[a]	3.05[a]	2.05[a]	−3.01	−3.95	−6.15
.2	7.61[a]	6.23[a]	3.76[a]	19.54[a]	9.32[a]	4.02[a]	−3.57	−6.43	−11.07[b]
.1	13.76[a]	10.6[a]	4.47[a]	27.47[a]	15.21[a]	4.72[a]	−6.8	−10.4[b]	−15.7[b]

[a]Exceeds the 2% overestimation threshold.
[b]Exceeds the 10% underestimation threshold.

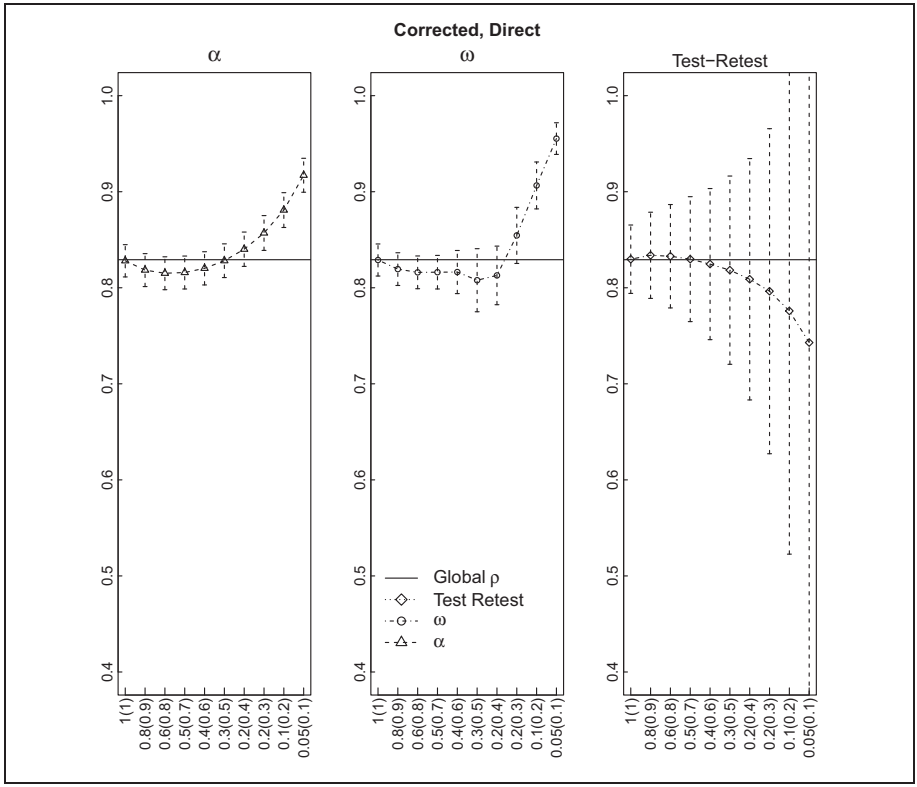

Figure 4. Corrected estimates of global reliability for direct range restriction
Note: *y*-axis, reliability estimate; *x*-axis, restriction amount (in parentheses) and corresponding *U* value. Each estimate is banded with a 95% confidence interval.

where p is the number of items, ρ is the sample estimate of reliability, and n is the sample size. Any deviation of the values obtained by Equation 10 we attribute to range restriction.

Table 6 compares these estimates. For simplicity, only uncorrected direct estimates are shown. Generally, range restriction inflated standard errors above and beyond the reduction in sample size, when compared with Equation 10. Notice that two values with the same sample size have different standard errors; the larger standard error typically belonging to the data set that was the most restricted. The only time standard errors did not exceed their expected values under range restriction was for ω; this is probably because the standard errors of ω are not monotonic as restriction increases but showed an inconsistent pattern. The range restriction increased standard errors beyond just the reduction in sample size.

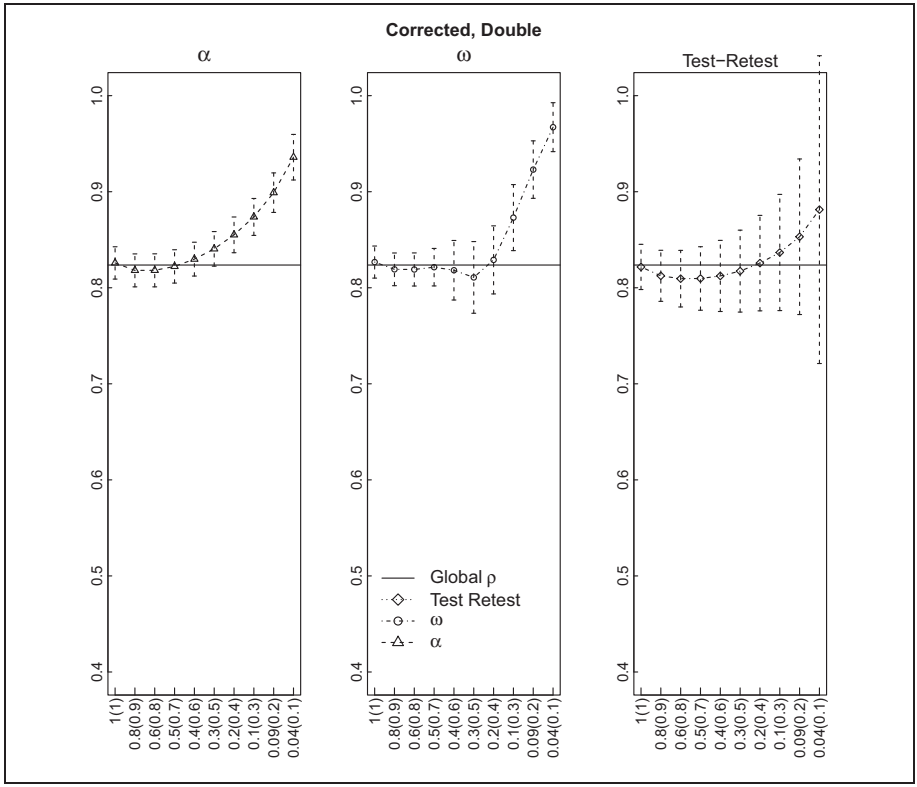

Figure 5. Corrected estimates of global reliability for double-hurdle range restriction
Note: *y*-axis, reliability estimate; *x*-axis, restriction amount (in parentheses) and corresponding *U* value. Each estimate is banded with a 95% confidence interval.

Discussion

It has long been known that both validity and reliability coefficients suffer under range restriction. To our knowledge, until now no one had systematically assessed the behavior of the most common reliability estimators under range restriction, probably because these estimators (i.e., α and ω) require item-level data. We have used IRT to create dichotomous item responses found in many cognitive tests, and investigated the standard errors of α, ω, and test–retest to understand both precision and bias in these estimators. Finally, we addressed how robust the estimators and their corrections are to violating the assumption of independence between *T* and *E*.

This study has demonstrated that under both direct and double-hurdle range restriction, α and ω are biased estimators of both local and global reliability parameters and probably should not be used, even under modest restriction. α and ω underestimated local reliability, likely because the assumption of independence was violated (Lord & Novick, 1968; Mendoza & Mumford, 1987). Additionally, α and ω *overestimated* global reliability when corrected. The reason for this overestimation is that the correction formula (Equation 3) assumes that $\sigma'_E = \sigma_E$, or that the residual

Table 5. Percentage of Bias of Each Corrected Estimator for Global Reliability Under Indirect Range Restriction

	α			ω			Test–retest		
Selection ratio	Low (.7)	Medium (.8)	High (.94)	Low (.7)	Medium (.8)	High (.94)	Low (.7)	Medium (.8)	High (.94)
1	−0.23	−0.16	−0.2	0	−0.06	−0.19	0	−0.01	0.01
.9	−0.3	−0.32	−0.34	−0.05	−0.2	−0.33	−0.05	−0.16	−0.13
.8	−0.33	−0.3	−0.39	−0.05	−0.17	−0.37	−0.06	−0.11	−0.18
.7	−0.27	−0.24	−0.31	0.05	−0.09	−0.27	0.01	−0.02	−0.09
.6	−0.26	−0.11	−0.08	0.09	0.06	−0.04	0.07	0.15	0.14
.5	−0.19	0.08	0.29	0.19	0.28	0.35	0.27	0.38	0.51
.4	−0.12	0.32	0.88	0.25	0.57	0.95	0.47	0.69	1.08
.3	0	0.62	1.49	0.14	0.94	1.57	0.89	1.13	1.68
.2	0.63	1.11	2.13[a]	−0.45	1.45	2.22[a]	2.01[a]	1.87	2.3[a]
.1	3.36[a]	3.27[a]	2.94[a]	−0.66	2.5[a]	2.93[a]	6.18[a]	4.64[a]	3.22[a]

[a]Exceeds the 2% overestimation threshold.

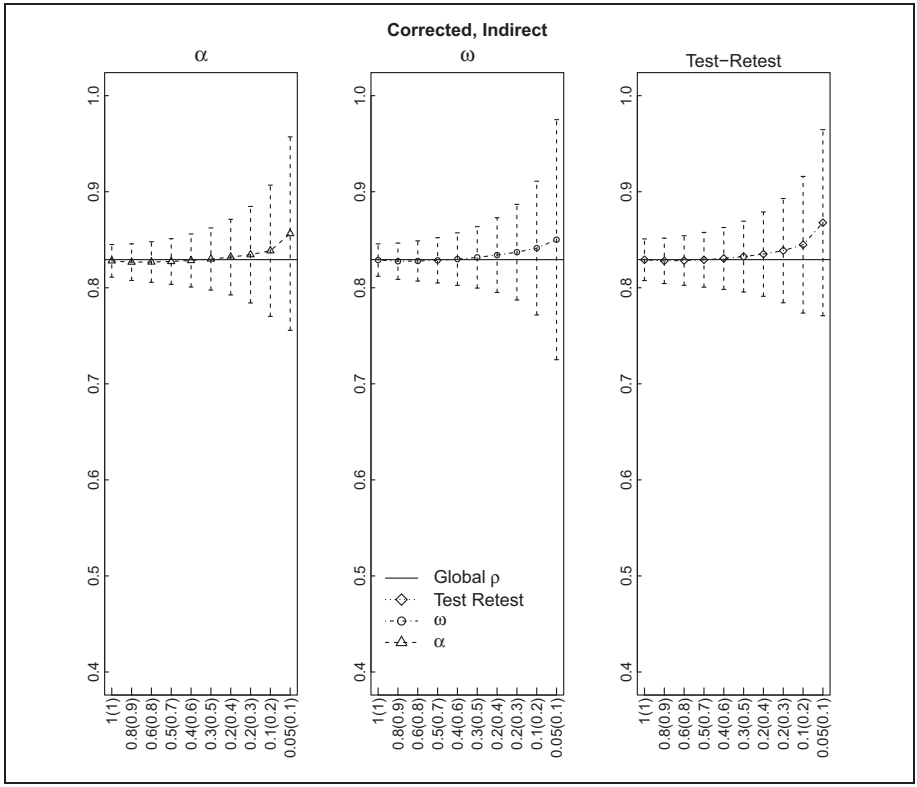

Figure 6. Corrected estimates of global reliability for indirect range restriction
Note: *y*-axis, reliability estimate; *x*-axis, restriction amount (in parentheses) and corresponding *U* value. Each estimate is banded with a 95% confidence interval.

variance is unaffected by selection (Lord & Novick, 1968). However, because X is a composite of T and E, the variance of E will necessarily be effected when selection is made on X. Lord and Novick (1968) note that if this assumption is not met (i.e., if $\sigma'_E \neq \sigma_E$), then the equation will overcorrect.

What is interesting to note is that coefficient α produced negative estimates under direct range restriction. These estimates in some situation were less than −1. The issue of negative reliability estimates (particularly using coefficient α) is not new (Krus & Helmstadter, 1993; Reinhardt, 1991). Reinhardt (1991) noted that when total test variance is small α can go negative. (See also the Appendix for an explanation.)

Another interesting finding of this study is that α frequently exceeded ω. α is supposed to be a lower bound to ω (McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005), yet this was not the case in this study under severe range restriction. There are two important things to consider when understanding why this might be the case. First, α and ω are equivalent under the τ-equivalent model. In the present study, the items were simulated to be τ-equivalent, so there should have been no advantage for ω. Second, ω is estimated from factor loadings. Under range restriction, the

Table 6. Percentage of Standard Errors as a Function of Sample Size and Selection Ratio for Direct Range Restriction and Mid (.8) Reliability

	α			ω			Test–retest		
Net sample size	Expected	200 sample size	500 sample size	Expected	200 sample size	500 sample size	Expected	200 sample size	500 sample size
200	0.018	0.017 (1)	0.079 (.4)	0.018	0.017 (1)	0.118 (.4)	0.018	0.022 (1)	0.049 (.4)
150	0.020	0.031 (.75)	0.117 (.3)	0.020	0.031 (.75)	0.045 (.3)	0.020	0.033 (.75)	0.061 (.3)
100	0.025	0.092 (.5)	0.185 (.2)	0.025	0.153 (.5)	0.020 (.2)	0.025	0.062 (.5)	0.079 (.2)
50	0.035	0.163 (.25)	0.413 (.1)	0.035	0.098 (.25)	0.029 (.1)	0.035	0.086 (.25)	0.122 (.1)

Note: Numbers in parentheses indicate selection ratio.

uniqueness estimates begin to go negative (i.e., a ''Heywood'' case). These uniqueness values were bounded between zero and one in this study, which may have altered the properties of ω.

It may be tempting to argue that because direct range restriction is less common than indirect (Thorndike, 1949), researchers need not worry about excessive bias. However, we have demonstrated that the effects of multiple hurdle range restriction mirror the effects of direct range restriction. If multiple hurdle selection is common, then these results do merit increased caution when using α or ω. Unfortunately, there has been little (if any) investigation about the frequency of multiple-hurdle procedures. However, there are reasons to expect that double-hurdle procedures are common (Roth et al., 2001; Schmidt & Hunter, 1998). First, institutions often must screen out a large number of applicants to reduce costs (Schmidt & Hunter, 1998). Setting a minimal cutoff allows institutions to filter through large amounts of applicants easily and at low cost. Additionally, Roth et al. (2001) also noted several large institutions that use a multiple hurdle procedure to ''weed out'' applicants, such as Toyota, the Federal Bureau of Investigation, the United States Employment Service, the Transportation Security Administration, and many universities. For example, medical schools may use MCAT/GPA scores to decide which students are interviewed (first hurdle) and then use an interview to decide final admission (second hurdle). Because the likelihood of multiple hurdle procedures is high, the results suggested in this article cannot be dismissed by simply assuming direct- and multiple-hurdle procedures are rare.

Another implication of this study is that τ-equivalence cannot be achieved under direct range restriction because the item variances and covariances will necessarily differ; as θ becomes more restricted, the range of *b* values exceeds the range of θ. In other words, when only the top performers remain, some items become sufficiently easy that their responses will produce no variability within that restricted sample, whereas moderately difficult items will produce some variability. Under such conditions, the item variances (and covariances) will differ, which indicates that the test is no longer τ-equivalent. Indeed, as reliability is a property of the scores in the sample rather than a property of the test (Feldt, 1965; Wilkinson, 1999), likewise τ-equivalence is also a property of the responses within a sample. Therefore, when the range of *b* parameters exceeds the range of θ, τ-equivalence is very difficult to achieve, as is an unbiased estimate of population reliability using ω or α.

Although the best estimator of reliability under direct and double-hurdle selection is test–retest, it unfortunately requires a second administration, which is costly and inconvenient. Future research may be directed at investigating alternative methods for estimating local and global reliability. These alternative methods must circumvent the assumption of independence to have success.

One alternative is to estimate reliability coefficients within the IRT framework. Because IRT estimates are invariant across samples, it is expected that range restriction will not bias reliability estimates. However, IRT is not a panacea for range restriction problems for two reasons. First, IRT estimates under range restriction will

Table 7. Recommended Estimates Depending on Type of Range Restriction (Direct vs. Indirect) and What Type of Estimate Is Desired (Local vs. Global)

Estimate Type	Direct Range Restriction	Indirect Range Restriction
Local	Test–retest	α, ω, test–retest
Global	Stauffer–Mendoza	Corrected α, ω, or test–retest

likely have large standard errors. This is because information functions are usually quite thin in the tails, which means that under selection (which typically occurs in the tails) standard errors will be large. Additionally, small sample sizes (which are common in a selected sample) will make the standard errors even larger. The second reason we hesitate to offer IRT as a general solution is because the problems identified in this article are test design issues, not statistical ones. When items are designed (selected) to discriminate among average performers (which is the typical test design), those items are often too easy for the above-average performers. No amount of statistical manipulation can address this issue. Instead, we recommend a tailored testing approach within the IRT framework. Such testing approaches will maximize information for the selected sample of interest.

Conclusion

Table 7 provides a summary of our recommendations based on direct versus indirect restriction and based on whether the researcher is estimating global or local reliability. When the sample is directly restricted, we recommend using test–retest to estimate the local reliability. The Stauffer–Mendoza correction can be used to estimate global reliability. However, it should only be used with large samples since it yields large standard errors. When the sample is indirectly restricted, α, ω, and test–retest yield fairly accurate estimates of local reliability with small standard errors. Consequently, we recommend using any of these estimates under indirect range restriction. On the other hand, if our interest is in global reliability, we recommend correcting these estimates using Equations 3 and 4.

Appendix

The total variance of a test can be expressed in terms of the item true, *T*, and error, *E*, components as follows:

$$\sigma_T^2 = 1'(\Sigma_{tt} + \Sigma_{ee} + 2\Sigma_{te})\mathbf{1},$$

where each Σ represent the covariance matrices of the items in terms of true and error scores, and **1** a vector of ones. Alternatively, we write the total variance in term of the (observed) item variances and covariances as follows:

$$\sigma_T^2 = \sum_i \sigma_{ii}^2 + \sum_{i,j} \sigma_{ij}.$$

Combining the two definitions we obtain that the total variance is given by the two sums

$$\sum_i \sigma_{ii} = \mathrm{tr}(\Sigma_{tt} + \Sigma_{ee} + 2\Sigma_{te}) \quad \text{and} \quad \sum_{i,j} \sigma_{ij} = 1'(\Sigma_{tt} + \Sigma_{ee} + 2\Sigma_{te})\mathbf{1} - \mathrm{tr}(\Sigma_{tt} + \Sigma_{ee} + 2\Sigma_{te})$$

or

$$\sum_{i,j} \sigma_{ij} = 1'(\Sigma_{tt} - D_{tt})\mathbf{1} + '(\Sigma_{ee} - D_{ee})\mathbf{1} + 2 \times 1'(\Sigma_{te} - D_{te})\mathbf{1},$$

where tr stands for trace and D is a diagonal matrix containing the diagonal elements of the corresponding covariance matrix. For KR-20 to be positive the sum of the item covariances must be smaller than the total variance, and ordinarily this is the case. Under severe range restriction, however, the sum of the covariance is often larger than the total variance.

Note that when selection is based on the test score (number correct), the matrix of item scores **I** becomes primarily a matrix of ones as we select only highly qualified applicants. (We refer to this condition as a decreasing selection ratio or as increasing the range restriction.) We observe less zeros and more ones as range restriction increases. Concurrently, the item true score representing the probability of getting an item correct given ability also increases as the range restriction increases. These item true scores approach one. Since we define the error matrix as the difference **E** = **I** − **T**, where **T** is a matrix containing the item true scores, the error becomes smaller as range restriction increases. Thus, the *item true scores increase* and the *item errors decrease* as range restriction increases. The process causes the variances to approach zero and the covariances between T and E to become negative. Mendoza and Mumford (1987) showed this trend at the total score level and we have explained above why it also occurs at the item level with dichotomous items. Furthermore, we have verified these results in our simulation, both by obtaining negative KR-20 estimates and by observing Σ_{TE} directly.

At heart is the relation between $1'(\Sigma_{te} - D_{te})\mathbf{1}$ and $1'\Sigma_{te}\mathbf{1}$. Note that $1'D_{te}\mathbf{1}$ is negative under severe range restriction, because it is the sum of the t_i and e_i covariances. Consequently,

$$1'(\Sigma_{te} - D_{te})\mathbf{1} = 1'\Sigma_{te}\mathbf{1} - 1'D_{te}\mathbf{1} = 1'\Sigma_{te}\mathbf{1} - (\text{Negative}) = 1'\Sigma_{te}\mathbf{1} + (\text{Positive}).$$

Furthermore, under severe range restriction

$$1'(\Sigma_{tt}\mathbf{1} - D_{tt})\mathbf{1} \approx 1'\Sigma_{tt}\mathbf{1}$$

and

$$1'(\Sigma_{ee}\mathbf{1} - D_{ee})\mathbf{1} \approx 1'\Sigma_{ee}\mathbf{1}$$

because the variances approach zero. Putting it all together we see that under severe range restriction, the sum of the item variances is larger than the sum of the total variance,

$$1'(\Sigma_{tt} - D_{tt})\mathbf{1} + 1'(\Sigma_{ee} - D_{ee})\mathbf{1} + 2 \times 1'(\Sigma_{te} - D_{te})\mathbf{1} \geq 1'(\Sigma_{tt} + \Sigma_{ee} + 2\Sigma_{te})\mathbf{1}.$$

It will not happen when we select on ability, but it will happen when we select on the total test score. Of course, better test designs would able to ameliorate the problem, but if we push the envelope it will eventually happen.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

1. Here, ω refers to the estimate of reliability (ω_T, to use the notation of Zinbarg et al., 2005) and not to the estimate of the general factor saturation of a test (ω_h).
2. This simulation assumes that the unrestricted variance of Y is known under double-hurdle selection, otherwise it could not be corrected.
3. A helpful reviewer noted that there is not a one-to-one mapping between discrimination parameters (a) and reliability. Instead, α maps into interitem correlations, which in turn maps into reliability.

References

Alexander, R. A., Carson, K. P., Alliger, G. M., & Cronshaw, S. F. (1989). Empirical distributions of range restricted SD_x in validity studies. *Journal of Applied Psychology, 74*, 253-258.

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. *Journal of Applied Psychology, 68*, 584-589.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296-322.

Cortina, J. M. (1993). What is coefficient α? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Cronbach, L. J. (1951). Coefficient α and the internal structure of tests. *Psychometrika, 16*, 297-334.

Dunbar, S. B., & Linn, R. L. (1991). Range restriction adjustments in the prediction of military job performance. In A. K. Wigdor & B. F. Green Jr. (Eds.), *Performance assessment for the workplace* (pp. 127-157). Washington, DC: National Academies Press.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30*, 357-370.

Gross, A. L., & Kagen, E. (1983). Not correcting for restriction of range can be advantageous. *Educational and Psychological Measurement, 43*, 389-396.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*, 523-531.

Kelley, T. L. (1921). The reliability of test scores. *Journal of Educational Research, 3*, 370-379.

Komaroff, E. (1996). *Coefficient alpha under simultaneous violations of essential tau-equivalence and uncorrelated errors* (Doctoral dissertation). University of Miami, FL.

Krus, D. J., & Helmstadter, G. C. (1993). The problem of negative reliabilities. *Educational and Psychological Measurement, 53*, 643-650.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and α factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*(1), 1-21.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational and Behavioral Statistics, 12*, 282-293.

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 200*, 1-66.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27*(1), 52-71.

Reinhardt, B. M. (1991). *Factors affecting coefficient alpha: A mini Monte Carlo study*. Paper presented at the meeting of the Southwest Educational Researchers Association, San Antonio, TX.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145-154.

Roth, P. L., Bobko, P., Switzer, F. S., III, & Dean, M. A. (2001). Prior selection causes biased estimates of standardized ethnic group differences: Simulation and analysis. *Personnel Psychology, 54*, 591-617.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology, 55*, 807-825.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*, 112-118.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529-540.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262-274.

Schmidt, F. L., Oh, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59,* 281-305.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Stauffer, J. M., & Mendoza, J. L. (2001). The proper sequence for correcting correlation coefficients for range restriction and unreliability. *Psychometrika, 66*, 63-68.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. Oxford, England: Wiley.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123-133.