Running Head: ESTIMATING RELIABILITY

Estimating Reliability in Primary Research

Michael T. Brannick

University of South Florida

Paper presented in S. Morris (Chair) *Advances in Meta-Analysis: New Approaches to Artifact Correction*. Symposium presented at the 20th annual conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA (April 2005).

Abstract

The current paper describes and illustrates three things: (a) sampling variance and confidence intervals for Cronbach’s Alpha, (b) the relative precision of reliability estimates from local studies and meta-analyses, and (c) how to blend the local and meta-analytic information to create an optimal local reliability estimate according to Bayesian principles. The paper is not about artifact corrections used to compute a meta-analysis. Rather it is about using information contained in a meta-analysis to improve local estimates of reliability. The improved estimates can result in better estimates and corrections for artifacts at the local level.

Estimating Reliability in Primary Research

Measurement experts routinely call for the estimation of the reliability of all measures (scores) used in a study based on that study's data (e.g., Thompson, 2003; Whittington, 1998). That is, primary researchers are asked to report estimates of the reliability of their measures based on their data. Despite such calls for reporting local estimates, many researchers fail to report any reliability estimates at all or else simply report estimates taken from test manuals or other literature reporting the development of the measure (e.g., Vacha-Haase, Ness, Nilsson, & Reetz, 1999, Yin & Fan, 2000). Measurement experts note that reliability estimates reported in test manuals or in articles reporting the development of a measure may not adequately represent the reliability of the data in any particular study because of the influence of the research context, including the variability of the trait in the population of interest and the context of measurement, including such factors as the purpose of measurement (e.g., selection vs. developmental feedback) and the testing conditions (e.g., noise, light, time of day, etc.).

On the other hand, such calls for local reliability estimates typically fail to mention of the importance of sampling error on the precision of the local study estimate (Hunter & Schmidt, 2004; for recent exceptions, see Cronbach & Shavelson, 2004; Vacha-Haase, Henson, & Caruso, 2002). With small samples, the local estimate of reliability will usually be much less precise than a comparable estimate taken from the test manual or from a meta-analysis. There may be a tradeoff between precision and applicability of primary study estimates and meta-analytic reliability estimates. That is, the local estimate may be more applicable than is the meta-analytic estimate (because of the influence of research context), but the local estimate may be less precise than is the meta-analytic estimate (because of sampling error). Clearly we would like to know the precision of the local estimate and to be able to articulate what the tradeoff may be. It is also possible to blend the local and meta-analytic estimates. The current paper therefore describes and illustrates three things:

1. sampling variance and confidence intervals for Cronbach's Alpha,
2. the relative precision of estimates from local studies and meta-analyses,
3. how to blend the local and meta-analytic information to create an optimal local estimate according to Bayesian principles (Lee, 1989; Brannick & Hall, 2003).

Confidence Intervals for Alpha

Cronbach's alpha appears to be the most commonly reported estimate of reliability in the psychological research literature (Hogan, Benjamin, & Brezinski, 2000). Because it is an intraclass correlation, its sampling distribution is awkward and confidence intervals have only recently become available for it. However, it is quite important to report the precision of the estimate of alpha (that is, its standard error or confidence interval, Iacobucci & Duhachek, 2003) so that researchers can understand the likely magnitude of error associated with the estimate.

An asymptotic (large sample) formula for the sampling variance of the function of Cronbach's Alpha $\sqrt{n}(\hat{\alpha}-\alpha)$ is shown by (Iacobucci & Dubachek, 2003, p. 480, Equation 2; van Zyl, Neudecker, & Nel, 2000, p. 276,Equations 20, 21):

$$Q=\left[\frac{2k^2}{(k-1)^2(j'Vj)^3}\right]\left[(j'Vj)(trV^2+tr^2V)-2(trV)(j'V^2j)\right] \qquad (1)$$

where *k* is the number of items that are added for form the composite whose reliability is indexed by alpha, ***j*** is a *k* x 1 column vector of ones, ***V*** is the covariance matrix of the items (that is, the sample estimate of the population covariance matrix), and *tr* is the trace function (the sum of the diagonal elements of a matrix). The asymptotic 95 percent confidence interval is given by (Iacobucci & Dubachek, 2003):

$$\hat{\alpha} \pm 1.96\left(\sqrt{\frac{Q}{n}}\right) \tag{2}$$

where *n* = (*N*-1). A small-scale simulation by van Zyl, Neudecker and Nel (2000) shows that the asymptotic estimate appears to yield reasonable results provided that N is a least 100. However, Yuan, Guarnaccia, and Hayslip (2003) recommended that bootstrap estimates be used to compute confidence intervals rather than asymptotic estimates. They based their recommendations on a comparison of methods using data from the Hopkins Symptom Checklist. Bootstrap estimates are computed by taking repeated samples of data with replacement from the study data to compute an empirical sampling distribution. The empirical sampling distribution is then inspected to see where the extremes of the distribution fall, that is, the bootstrap estimates allow for the calculation of an empirical confidence interval.

SAS programs that can be used to compute both asymptotic and bootstrap confidence intervals for alpha can be found at http://luna.cas.usf.edu/~mbrannic/software/softdir.htm. The programs contain sample data from students who completed a questionnaire composed of some IPIP extroversion items. Based on responses of 100 people to the ten items, the estimated alpha is .89, with asymptotic confidence interval (95 percent) of .85 to .92. The bootstrap confidence interval (95 percent) ranges from .85 to .91, so there appears to be good agreement between the two different confidence intervals estimated by the asymptotic and bootstrap methods in this case. Either method can be used to estimate the sampling variance of alpha for a local study.

Unfortunately, both the asymptotic sampling variance and bootstrap methods require information that is not typically presented in journal articles. The asymptotic method requires the covariance matrix for the items, and the bootstrap method requires the raw data. Both methods are of interest to primary researchers, but meta-analysts typically do not have access to the required data. It is possible, however, to assume compound symmetry for the covariance matrix (that is, to assume that the covariance or correlation among all items is the same). Under the assumption of compound symmetry, it is possible to solve for the common covariance and then to use the asymptotic formula to estimate the sampling variance of alpha. Such a procedure is analogous to using the reported estimates of the mean and standard deviation to estimate the reliability (KR-21) of tests composed of dichotomous items.

Under the assumption of compound symmetry, the expression for alpha becomes (van Zyl, Neudecker, & Nel, 2000, p. 272, Equation 3)

$$\alpha = \frac{k\rho}{1+\rho(k-1)} \tag{3}$$

where k is the number of items and ρ is the common element in the covariance matrix (i.e., the correlation of each item with all other items). A little algebra allows us to solve for the common element, thus:

$$\rho = \frac{\alpha}{k-\alpha(k-1)}\,. \tag{4}$$

For example, if alpha is .8 and there are 3 items, then the implied correlation matrix is

1	0.5714	0.5714
0.5714	1	0.5714
0.5714	0.5714	1

If the primary researcher reports alpha, the number of items, and the sample size, then the meta-analyst can find an approximate sampling variance and therefore confidence intervals for the alpha estimate.

Precision of Estimates

Although measurement experts routinely call for local estimates of reliability (and there is no real reason NOT to report them), measurement experts typically fail to note the relative precision of local and meta-analytic estimates of reliability. That is, they fail to note that the local estimates tend to contain much more sampling error than do meta-analytic estimates (see also Sawilowsky, 2000, for further aspects of the controversy about reporting local reliability estimates and their meaning). The second contribution of the current paper is to quantify the precision of the two estimates to allow an explicit comparison of the precision of the two different estimates.

There can be no stock answer to the question of the relative precision of the estimated reliability in a given sample versus the mean reliability of a meta-analysis. For the local study, the uncertainty of the reliability depends chiefly on the number of items, the magnitude of the covariances, and the number of people[check this]. As the number of items, the size of the covariance, and the size of the sample all increase, the local estimate will become precise. The meta-analytic result will become precise as the local samples become precise and also as the number of studies in the meta-analysis increases. In both cases, the sampling variance can be used to index the precision of the estimate.

Let's look at a single example. The data in the following table were copied from Thompson and Cook (2002). The study was a meta-analysis of reliability estimates for a survey assessing user satisfaction with library services across 43 different universities. Table 1 shows the alpha estimates for the total score (based on k=25 items), as well as the alpha estimates for one of the subscales (5 items), Information Access. The sample size for each sample is also reported. From the information given in Table 1, I computed an estimate of the common element (rho) based on the assumption of compound symmetry. I then calculated the estimated sampling variance for each study. For the Information Access subscale, the study sampling variance estimates ranged from .000144 to .001425. The mean sampling variance was .000593. The estimated population variance of the alpha coefficients (raw data) was.00148651. I divided that by 43 (the sample size) to find the variance of the mean (the squared standard error of the mean), which was .00003547. The variance estimates tell us about the precision of the estimates. If we compare the variances by forming ratios, we can get a single number that represents the relative precision of the estimates (relative efficiency in statistical terms). In this study, the mean of the meta-analysis was from about 4 to 40 times more precise than the individual study. The meta-analytic mean was about 17 times more precise than the average local study.

Another way to consider the same issue is to examine the width of confidence intervals (the width is the difference between the maximum and minimum value of the confidence interval). For the individual studies, the width of the confidence interval, which was computed by taking four times the standard error, ranged from .048 to .151; the mean width was .095. The width of the confidence interval for the mean of the study reliabilities was .024, so the confidence intervals from the these studies tend to be on average about four times wider than the confidence interval for the mean across studies. All of the confidence intervals tend to be rather small. Compare them to the confidence interval width for a correlation of .30 with N = 100, which has a value of .367, which is nearly four times larger than the average confidence width for these studies. The variance estimate for the correlation was computed by

$$\sigma_r^2 = \frac{(1-\rho^2)^2}{N-1} \text{ , with } \rho = .30 \text{ and N=100.} \qquad (5)$$

There are two main points to this exercise. First, it should be clear that the mean of the studies has less sampling error than do the individual studies. Second, even moderate alpha estimates can be rather precise.

Blending Local and Meta-Analytic Estimates

By allowing researchers to combine local and meta-analytic data, Bayesian estimates allow researchers with small samples to ‘borrow strength’ from the meta-analytic estimates in a statistically optimal way. The paper shows how to combine both overall or global estimates from a meta-analysis with a local estimate, and also to combine the output of a meta-analytic regression analysis with a local reliability estimate. For example, if size of company (or another continuous variable) has been shown to moderate the reliability of the measure, it is possible to calculate the estimated reliability for the current company from the meta-analytic regression and to combine that meta-analytic estimate with the local study data to yield a local ‘best’ estimate. Such a result is important in practical applications in which reliability influences the interpretation of the results. The approach described in this paper is based on the work of Brannick (2001) and Brannick and Hall (2003).

If we want to borrow strength from a meta-analysis to bolster our local study, we need estimates of the uncertainty of both the local estimate and for the meta-analytic estimate. For the local study, we will consider only sampling error as a source of uncertainty. The index we will use is the sampling variance estimate for alpha. Two different sources of uncertainty may apply to the meta-analysis, however. The first of these concerns the value of the mean of the meta-analysis. Because the meta-analysis is based on a finite number of observations (and our universe of generalization is typically infinite), the actual population mean cannot be known; it can only be estimated. The confidence interval for the meta-analytic mean quantifies the degree of uncertainty of this type. The variance associated with sampling error can be calculated in several ways in a meta-analysis. The simplest way is to calculate a sampling variance of the mean effect size is to do it just as you would for any raw data:

$$V_s = \frac{\sum (ES - \overline{ES})^2 \Big/ (k-1)}{k}, \qquad (6)$$

that is, the estimated population variance of the effect sizes divided by the number of studies. Most meta-analysts would use a weighted formula, but regardless of the computation, we are estimating the sampling variance of the mean effect size.

The second kind of uncertainty has to do with the distribution of infinite-sample reliabilities in our population of interest. If all of the studies of interest have the same underlying parameter, that is, all of the infinite-sample reliabilities are the same across studies, then there is no uncertainty about the value of the parameter from situation to situation (assuming, of course, that we could actually compute reliabilities without sampling error in them). Such a scenario is known as fixed-effects in meta-analysis. If the infinite sample reliability varies from context to context or sample to sample, then the underlying population has a distribution of reliabilities. Such a scenario is known as random-effects in meta-analysis. In the fixed effects case, the variance of the distribution of infinite-sample effect sizes is zero. In the random-effects case, the variance of the distribution of infinite-sample effect sizes is greater than zero. Some texts refer to the variance of the distribution of infinite-sample effect sizes as the random-effects variance component (REVC); others call it the variance of rho or tau-squared.

$$REVC = \sigma_{\rho}^{2} = \tau^{2} \qquad (7)$$

There are also several ways to calculate such a value. All of them depend upon knowing the sampling distribution of the effect size. In this case, we need to know the sampling variance of coefficient alpha, which was expressed as *Q/n* (see Equations 1 and 2).

To blend the meta-analysis mean and the local study estimate, we simply take a weighted average. The weights are the inverse of the respective uncertainties of the two quantities. The uncertainty from the local study comes from the sampling variance. The uncertainty from the meta-analysis comes from the REVC, that is, from Equation 7. Note that if we were trying to update the meta-analytic mean, the uncertainty for the meta-analysis would come from the sampling error of the mean, that is, from Equation 6. The blending of interest in this paper results in a revised estimate of the local parameter, not the global mean.

It is often the case that the meta-analysis employs moderators to explain some of the variability in effect sizes. In the case of reliability, moderators such as test length, delay between test and retest, and participant characteristics such as age can be used as moderators. The effect of employing such moderators in meta-analysis is two-fold. First, it allows the more accurate prediction of population values through consideration of the characteristics of the context, such as the nature of the participants and protocol of data collection. Second, it tends to reduce the size of the REVC, so that uncertainty about the values of infinite-sample effect sizes are reduced. Models that employ moderators but also allow for a residual REVC are known as mixed-effects models in meta-analysis, because the moderator part is treated as a fixed effect, but the residual variance in infinite-sample effect sizes is treated as random.

Empirical Bayes estimates can be employed in both cases, that is, the estimates can be computed both when there are no moderators and when there are moderators. Next, I present an illustration of each.

Random-Effects Model

Beretvas, Meyers, and Leite (2002) reported a meta-analysis of reliability estimates for the Marlowe-Crowne social desirability scale. Their method of analysis differed in several respects from what I would recommend (they first took the square root of internal consistency estimates, transformed by Fisher's r to z, and considered the sampling variance of internal consistency so transformed to be 1/(N-3)). For purposes of this paper, please pretend that they had not taken any transformation and had instead used the sampling error estimates provided here. Beretvas et al. (2002) found a mean internal consistency across studies of .726 (back-transformed to the original metric). The estimated REVC was .02633. Suppose that I collect data for my local study and find that my estimated alpha is .80 and my estimated sampling variance is .01. Then, according to Bayesian principles, I can take a weighted average of the two estimates, where the weights are the inverse of the respective variances, thus:

$$ES_{Post} = \frac{V_{\Pr ior}^{-1} ES_{\Pr ior} + V_{Likelihood}^{-1} ES_{Likelihood}}{V_{\Pr ior}^{-1} + V_{Likelihood}^{-1}} \qquad (8)$$

The uncertainty of the result is the inverse of the sum of the weights in the previous step, that is,

$$V_{Post} = \frac{1}{V_{\Pr ior}^{-1} + V_{Likelihood}^{-1}} . \qquad (9)$$

In our example,

$$ES_{Post} = \frac{37.98(.726) + 100(.80)}{37.98 + 100} = .78$$

and

$$V_{Post} = \frac{1}{37.98 + 100} = .007247 .$$

The standard error of the estimate would be sqrt(1/137.98) = .085. In this particular example, we would have an initial reliability estimate of .80 with a confidence interval ranging from .6 to 1.0 to a revised reliability estimate of .78 with a confidence interval from .61 to .95.

Mixed Effects Model

In the same article, Beretvas et al. (2002) computed a regression equation linking study characteristics (moderators) to the obtained reliability estimates. The intercept for the study was .93, and the statistically significant regression coefficients were .30, -.22, .01, and .01 for the moderators Age Range, Proportion of Men, Mean Number of Items, and Number of Items, respectively (number of items appears twice, but this is a peculiarity of their data collection and need not concern us for the present application).

Of further interest was the REVC between studies, which was .01987 (compare to our earlier value of .02633). The weight associated with this value is 50.33 (up from 37.98). Assuming that we knew the Beretvas et al. (2002) coding system for the regression and that we had the required information on our local sample, we could plug the study characteristics into their regression equation to find a predicted value of reliability for our study. Suppose that we have done so and found that value to be .85. Again suppose that our local value is .80 and the estimated sampling variance is .01. Then our revised value would be:

$$ES_{Post} = \frac{50.33(.85)+100(.80)}{50.33+100} = .82$$

and

$$V_{Post} = \frac{1}{50.33+100} = .006652$$

So we would start with an estimate of .8 with a confidence interval from .6 to 1, and end up with a revised estimate of .82 with a confidence interval from .66 to .98.

Applications

I see three primary applications of the revised local estimates of reliability. These are the same as the application of the original estimates, namely (a) communication of the magnitude of measurement error, (b) communication of the magnitude of error of individual scores (standard error of measurement), and (c) correction of effect size data for unreliability (i.e., the disattenuation for reliability formula). In line with the purpose of this symposium, I suppose it could be argued that some sort of adjusted reliabilities might be preferable to unadjusted reliabilities for purposes the of meta-analysis of validities. The main purpose of the meta-analysis of reliabilities, however, is to understand the reasons for the variance in the observed reliabilities.

Remarks on the Model

It may seem strange to average the result of the local study with the result of a meta-analysis. Whether it is reasonable to take such an average depends primarily on the exchangeability or relevance of the meta-analysis to the local study. If the studies contained within the meta-analysis appear similar to the local study in the essential elements, then taking the average appears reasonable. On the other hand, if the studies in the meta-analysis do not appear relevant to the local study, then taking such an average is not reasonable. However, similar constraints ought to apply to the meta-analysis. In other words, if you believe it is reasonable to take the average of the studies in the meta-analysis, and the local study could reasonably be drawn from the meta-analysis, then you should find it reasonable to take the average of the local and meta-analytic result.

I personally do not think that most meta-analyses in industrial and organizational psychology take a sensible approach to averaging effect sizes, that is, most meta-analyses average many different things rather than one independent variable and one dependent variable. Consider the difference between a meta-analysis of the benefits of aspirin for preventing a second heart attack and a meta-analysis of all the effect sizes in the *Journal of Applied Psychology*. One of the things I would hope to see as we move forward is online databases of study effect sizes and perhaps even raw data, so that practitioners could sample those studies that are most relevant to their situation, in effect completing a smaller meta-analysis that is most relevant to their situation.

In my opinion, the REVC computed for most tests to date are large enough so that it is reasonable to expect that the test validity is not constant, but rather varies to some extent, although this could be due to mixing different tests and criteria into the same meta-analysis as much as to differences in context or other specific moderators. Because we have only one or two random-effects (or mixed-effects) meta-analyses of reliability data so far, so we don't really know how large the REVCs will tend to be for reliability data. I hope to analyze some existing datasets soon to check on this.

The equations that I used for the empirical Bayes estimates are based on the assumption of the normal distribution for both the prior and the likelihood. We know that alpha is not normally distributed, and that the sampling distribution of alpha only approaches the normal as the sample size increases. My view on this is that all quantitative models in psychology can be shown to be wrong if only the proper data are gathered in sufficient number and detail. However, the question should not be whether the models are absolutely correct (they are not), but rather whether they are sufficiently beneficial to justify their use. Research is needed to determine whether the procedure I advocated is reasonable with sample sizes typically encountered in practice.

Table 1.

Data for the LibQual+ Meta Analysis

Alpha for Information Access	Alpha for Total Score	N
.8038	.9567	232
.7906	.9541	251
.7377	.9554	760
.6629	.9305	408
.7778	.9473	266
.7198	.9330	775
.7323	.9501	265
.7050	.9275	412
.7023	.9361	224
.7828	.9505	679
.7493	.9483	412
.7640	.9519	689
.7645	.9413	219
.7421	.9424	653
.7774	.9561	303
.7681	.9442	429
.8069	.9517	275
.7597	.9526	289
.7151	.9435	184
.7540	.9533	286
.7619	.9477	588
.7936	.9570	206
.6483	.9275	572
.7592	.9535	363
.7945	.9548	274
.7336	.9466	690
.7575	.9573	441
.7863	.9500	223
.7114	.9430	379
.7546	.9493	869
.7544	.9435	274
.7447	.9399	305
.7177	.9256	954
.8133	.9578	180
.8136	.9543	807
.7274	.9447	462
.7746	.9487	395
.7359	.9448	406
.7433	.9283	210
.7436	.9375	938
.7230	.9575	174
.7360	.9488	272
.8395	.9710	168

References

Beretvas, S. N., Meyers, J. L., Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 62*, 570-589.

Brannick, M. T. (2001). Implications of empirical Bayes meta-analysis for test validation. *Journal of Applied Psychology, 86*, 468-480.

Brannick, M. T., & Hall. S. M. (2003). Validity generalization from a Bayesian perspective. In K. Murphy (Ed.) *Validity generalization: A critical review* (pp. 339-364). Mahwah, NJ: Lawrence Erlbaum.

Cronbach, L J., & Shavelson, R. J. (2004). My current thoughts on coefficient apha and successor procedures. *Educational and Psychological Measurement*, *64*, 391-418.

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.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology*, *13*, 478-487.

Lee, P. M. (1989). *Bayesian statistics: An introduction*. New York: Halsted Press.

Sawilowsky, S.S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's "reliability generalization" method and some *EPM* editorial policies. *Educational and Psychological Measurement, 60*, 157-173.

Thompson, B (Ed.) (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Thompson, B., & Cook, C. (2002). Stability of the reliability of LibQUAL+-super(TM) scores: A reliability generalization meta-analy6sis study. *Educational and Psychological Measurement, 62*, 735-743.

Vacha-Haase, T., Henson, R. K., Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62*, 562-569.

Vacha-Haase, T., Ness, C.M., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education, 67*, 335-341.

van Zyl, J. M., Neudecker, H. & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65*, 271-280.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58*, 21-37.

Yin, P., & Fan, X. (2000). Assessing the reliability of the Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*, 201-233.

Yuan, K., Guarnaccia, C. A., & Hayslip, B. (2003). A study of the distribution of the sample coefficient alpha with the Hopkins Symptom Checklist: Bootstrap versus asymptotics. *Educational and Psychological Measurement*, *63*, 5-23.