Journal of Organizational Behavior

Journal of Organizational Behavior, J. Organiz. Behav. **36**, 3–15 (2015)
Published online 13 October 2014 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/job.1960

Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning

DOUGLAS G. BONETT[1]* AND THOMAS A. WRIGHT[2]

[1]*Psychology Department, University of California, Santa Cruz, California, U.S.A.*
[2]*Gabelli School of Business, Fordham University, Bronx, New York, U.S.A.*

Summary Cronbach's alpha is one of the most widely used measures of reliability in the social and organizational sciences. Current practice is to report the sample value of Cronbach's alpha reliability, but a confidence interval for the population reliability value also should be reported. The traditional confidence interval for the population value of Cronbach's alpha makes an unnecessarily restrictive assumption that the multiple measurements have equal variances and equal covariances. We propose a confidence interval that does not require equal variances or equal covariances. The results of a simulation study demonstrated that the proposed method performed better than alternative methods. We also present some sample size formulas that approximate the sample size requirements for desired power or desired confidence interval precision. R functions are provided that can be used to implement the proposed confidence interval and sample size methods. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: Cronbach's alpha; confidence interval; sample size planning

Cronbach's alpha reliability (Cronbach, 1951) is one of the most widely used measures of reliability in the social and organizational sciences. Cronbach's alpha reliability describes the reliability of a sum (or average) of q measurements where the q measurements may represent q raters, occasions, alternative forms, or questionnaire/test items. When the measurements represent multiple questionnaire/test items, which is the most common application, Cronbach's alpha is referred to as a measure of "internal consistency" reliability. If the measurements are "parallel" (see, e.g., McDonald, 1999), they will have equal variances and equal covariances. If the measurements are "essentially tau-equivalent" (see, e.g., McDonald, 1999), they will have equal covariances but will not necessarily have equal variances. It can be shown (see, e.g., McDonald, 1999) that Cronbach's alpha correctly describes the reliability of the sum or average of q measurements that satisfy the parallel assumption or the less restrictive essentially tau-equivalent assumption. If the measurements are "congeneric" measurements (Joreskog, 1971), they can be represented by a one-factor model with uncorrelated measurement errors. The congeneric assumption is more realistic than the tau-equivalence assumption because the measurements are not required to have equal variances and the covariances among the measurements are not required to be equal.

As an example, suppose we have a five-item questionnaire ($q = 5$) to measure employee work ethic with each item scored on a 1–7 Likert scale. In this example, the parallel measurement assumption requires all five items to have equal variances and all $5(4)/2 = 10$ covariances among the five items are required to be equal. A tau-equivalent assumption does not require the five items to have equal variances but does require the 10 covariances to be equal. With a congeneric assumption, the variances of the five items can be unequal and the 10 covariances can be unequal.

Sijtsma (2009) has criticized the widespread use of Cronbach's alpha because it understates the reliability of the sum or average of the q measurements if the essentially tau-equivalent assumption is not satisfied. However, this criticism may be too harsh because the degree of understatement will be small in typical applications where the measurements are well approximated by a one-factor model and the factor loadings for the measurements are not

*Correspondence to: Douglas G. Bonett, Psychology Department, University of California, Santa Cruz, California 95064, U.S.A. E-mail: dgbonett@ucsc.edu

Received 02 August 2014, Accepted 29 August 2014

highly dissimilar. Many q-item congeneric measurements have standardized factor loadings that are not too dissimilar (e.g., ranging from about .5 to .8) because items with small loadings or items that load on a second factor are typically discarded during scale development and Cronbach's alpha will only slightly understate the reliability of the scale in these situations. Furthermore, the tau-equivalent assumption is realistic in applications where the q measurements represent alternative forms of a placement test or the ratings of properly trained raters. The essentially tau-equivalent assumption will be violated in applications where the measurements are not all measured using the same metric, for example, if some items are scored agree/disagree while other items are scored 1 to 7. In these applications, it may be possible to transform congeneric measures into approximate essentially tau-equivalent measurements by simply rescaling some of the measurements.

It is a common but inappropriate practice to report only the sample value of Cronbach's alpha. The sample value of Cronbach's alpha contains sampling error of unknown direction and unknown magnitude. A confidence interval for the population value of Cronbach's alpha, denoted here as ρ_q, should be reported along with the sample value. For example, if Cronbach's alpha was estimated for our five-item work ethic questionnaire in a very small sample, the sample value of Cronbach's alpha might look impressive, but a 95 percent confidence interval for ρ_q could have a lower limit that indicates very poor reliability. Furthermore, the reliability of the measurements could vary across subpopulations of males and females, different age or ethnic groups, or different testing environments. For these reasons, reliability should be examined under different demographic and testing conditions and a confidence interval for the difference in population Cronbach's alpha values could be reported for all interesting pairs of demographic or testing conditions. For example, Cronbach's alpha for the five-item work ethic questionnaire could be assessed in one sample of hourly workers and in a second sample of salaried workers.

Although the reliability of a measurement is informative in and of itself, it is important to have some idea about the value of ρ_q in any study that uses a sum or average of q measurements as a response variable or predictor variable in a statistical analysis. The unreliability of the response variable reduces the power and precision of inferential statistical methods. Furthermore, unreliability of a predictor variable in a simple linear regression model attenuates a slope estimate, unreliability of a response variable attenuates a standardized mean difference, and a bivariate correlation is attenuated by the unreliability of each variable. All of the above statements refer to the effect of the population reliability value and not the sample reliability value that is reported in the vast majority of social science and organizational studies. This is another reason why it is important to report a confidence interval for ρ_q rather than just the sample value of ρ_q.

Some researchers worry that the sample value of Cronbach's alpha for a response variable or a predictor variable in a statistical analysis might be unacceptably small (we have both heard of numerous reports where manuscripts were rejected simply because the sample value of Cronbach's alpha was below .7). However, there is no universal minimally acceptable reliability value. An acceptable reliability value depends on the type of application, and furthermore, the focus should be on the population reliability value and not on the sample reliability value.

In real life organizational applications where a rating or test score will be used to make an important selection or placement decision for a job applicant or a current employee, one could argue that a reliability value of .95 or higher is desirable (Nunnally & Bernstein, 1994, p. 265). However, in more typical research applications where the focus is on estimating the size of population effect sizes (e.g., population correlations, population slopes, and population standardized mean differences), information regarding the response variable reliability or predictor variable reliability can assist the researcher in a more accurate interpretation of the effect size results. In these situations, much smaller reliabilities can be tolerated as long as the effect size results are interpreted accordingly. For instance, the population standardized mean difference will be attenuated by the square-root population reliability value of the response variable (see Hedges & Olkin, 1980, p. 135). With the information provided by a confidence interval for ρ_q, researchers will be better able to assess the extent to which the endpoints of a confidence interval for a standardized mean difference (see Bonett, 2008) have been attenuated due to the unreliability of the response variable.

Consider a study that compares the mean work ethic in subpopulations of male and female employees, and suppose a 95 percent confidence interval for a population standardized mean difference (female minus male) is [.42, .54]. Suppose also that a 95 percent confidence interval for Cronbach's alpha reliability of the work ethic

DOI: 10.1002/job

questionnaire is [.68, .85]. Assuming ρ_q is at most .85, the plausible range for the population standardized mean difference will be at least [$.42/\sqrt{.85} = .46$, $.54/\sqrt{.85} = .59$], and assuming ρ_q is at least .68, the plausible range for the population standardized mean difference will be at most [$.42/\sqrt{.68} = .51$, $.54/\sqrt{.68} = .65$]. In this example, a response variable (i.e., work ethic) reliability as low as .65 or as high as .85 would lead to essentially the same conclusions even though current practice would characterize a reliability value of .65 as "unacceptable" and a reliability value of .85 as "excellent".

In the sections that follow we propose a new confidence interval for ρ_q that does not make the traditional assumptions of parallel measurements, and we explain how this confidence interval can be used in hypothesis testing applications. We then report the results of a simulation study that demonstrates that the proposed confidence interval performs better than two competing methods. Next, we develop some sample size planning formulas for approximating the sample size needed to obtain a confidence interval with desired precision or to perform a hypothesis test with desired power. R functions to implement the recommended confidence interval and sample size formulas are given in the Appendix. We conclude with some best-practice recommendations.

Interval Estimation

The classic confidence interval for ρ_q (Feldt, 1965), which has been implemented in SPSS, assumes parallel measurements. Since ρ_q is an appropriate measure of reliability under the less restrictive assumption of essentially tau-equivalent measurements, a confidence interval for ρ_q that requires parallel measurements is unnecessarily restrictive. van Zyl, Neudecker, and Nel (2000) derived an approximate variance for an estimator of ρ_q under the non-restrictive assumption of a general covariance structure of the measurements (i.e., the variances can be equal or unequal and the covariances can be equal or unequal). We use their result to derive the following approximate $100(1-\alpha)\%$ confidence interval for ρ_q

$$1-exp\left[ln\left(1-\hat{\rho}_q\right)-ln[n/(n-1)]\pm z_{\alpha/2}\sqrt{var\left(\hat{\rho}_q\right)/\left(1-\hat{\rho}_q\right)^2}\right] \quad (1)$$

where $\hat{\rho}_q = [q/(q-1)][1-tr(\mathrm{V})/j'Vj]$, $ln[n/(n-1)]$ is a bias adjustment proposed by Bonett (2010), j is a $k \times 1$ vector of ones, and

$$var\left(\hat{\rho}_q\right) = \left[\frac{2q^2}{(q-1)^2\left(j'Vj\right)^3}\right]\left[\left(j'Vj\right)\left(trV^2+tr^2V\right)-2(trV)\left(j'V^2j\right)\right]/(n-3) \quad (2)$$

Equation (2) is a minor modification (i.e., $n-3$ replacing n) of the variance formula derived by van Zyl et al. (2000) and $\hat{\rho}_q$ is the sample value of ρ_q. Unlike the classic confidence interval for ρ_q, Equation (1) is appropriate for both essentially tau-equivalent and congeneric measurements with the understanding that ρ_q will understate the reliability of the sum or average of congeneric measurements. As noted above, the degree of understatement will be very minor in multi-item questionnaires where the items have been factor analyzed with problematic items removed or modified. The degree of understatement should also be very small in applications where the q measurements represent q alternate forms of a placement test or the ratings of q properly trained raters. The classic confidence interval and Equation (1) both assume that the measurements can be approximated by a multivariate normal distribution.

To illustrate the computation of Equation (1) using the R function in the Appendix (Function 1), consider a study where a random sample of $n = 150$ employees each receive $q = 3$ peer evaluations of their job performance. To compute Equation (1), the researcher must compute the 3×3 covariance matrix for the three peer evaluations (using any of the popular statistical packages). First, copy the Function 1 commands at the R prompt, then assign the variances and covariances to the covariance matrix V as shown below.

DOI: 10.1002/job

```
> V = matrix(c(1.1, .82, .75, .82, 1.3, .77, .75, .77, 1.2), nrow=3, ncol=3)
```

To obtain a 95 percent confidence interval for the population reliability, enter the following command

```
> CIreliability(.05, 3, 150, V)
```

which will display the 95 percent lower limit for ρ_3 (.80), the sample reliability value (.85), and the 95 percent upper limit for ρ_3 (.89) for this example.

For the special case of parallel measurements, the Feldt confidence interval or the following approximate confidence interval of Bonett (2010) can be used

$$1 - exp\left[ln\left(1-\hat{\rho}_q\right) - ln\{n/(n-1)\} \pm z_{\alpha/2}\sqrt{2q/[(q-1)(n-2)]}\right] \tag{3}$$

and $2q\left(1-\hat{\rho}_q\right)^2/[(q-1)(n-2)]$ can be used as an estimate of $var\left(\hat{\rho}_q\right)$. We recommend the routine use of Equation (1) when reporting the results of a study where Cronbach's alpha has been computed, but we will exploit the computational simplicity of Equation (3) and $var\left(\hat{\rho}_q\right) = 2q\left(1-\hat{\rho}_q\right)^2/[(q-1)(n-2)]$ for parallel measurements in our derivation of sample size requirement formulas.

As noted above, the reliability of a sum or average can vary across demographic groups or testing conditions. In some applications, it will be interesting to check for gender, ethnicity, employee tenure, or age differences in Cronbach's alpha. In other applications, it could be useful to determine if the reliability differs when data are collected under different testing conditions, such as online or traditional paper and pencil methods. Let ρ_{q1} and ρ_{q2} denote Cronbach's alpha reliability for two different demographic subpopulations or two different testing conditions. Approximate lower and upper $100(1-\alpha)\%$ interval estimates for $\rho_{q1} - \rho_{q2}$ may be expressed as (see Bonett, 2010)

$$L = \hat{\rho}_{q1} - \hat{\rho}_{q2} - \sqrt{\left(\hat{\rho}_{q1} - L_1\right)^2 + \left(\hat{\rho}_{q2} - U_2\right)^2} \tag{4a}$$

$$U = \hat{\rho}_{q1} - \hat{\rho}_{q2} + \sqrt{\left(\hat{\rho}_{q1} - U_1\right)^2 + \left(\hat{\rho}_{q2} - L_2\right)^2} \tag{4b}$$

where L_i and U_i are the lower and upper limits, respectively, of Equation (1) computed from group i and $\hat{\rho}_{qi}$ is an estimate of ρ_q computed from group i.

To illustrate the computation of Equations (4a) and (4b) using the R function in the Appendix (Function 2), consider a study where the reliability of an eight-item leadership questionnaire was estimated from a sample of male employees and a sample of female employees. The sample Cronbach's alpha reliability for male employees is .91 with a 95 percent confidence interval of [.88, .93], and the sample Cronbach's alpha reliability for female employees is .86 with a 95 percent confidence interval of [.82, .89]. First, copy the Function 2 commands at the R prompt and then enter the following command

```
> CIrelDiff(.91, .88, .94, .86, .82, .89)
```

which will display the 95 percent lower limit for $\rho_{81} - \rho_{82}$ (.008), the difference in sample reliabilities (.05), and the 95 percent upper limit for $\rho_{81} - \rho_{82}$ (.094) for this example.

Hypothesis Testing

A test of H_0: $\rho_q = h$, where h is some value specified by the researcher, can be obtained from Equation (1) using the following three-decision rule: if the lower limit is greater than h, then reject H_0 and accept $\rho_q > h$; if the upper limit is

DOI: 10.1002/job

less than h, then reject H_0 and accept $\rho_q < h$; otherwise, the results are inconclusive. In the above example, the 95 percent confidence interval for ρ_8 among male employees is [.89, .94]. If we wanted to test H_0: $\rho_8 = .80$, we would reject H_0 and accept $\rho_8 > .80$ because the lower limit is greater than .80. The probability of a directional error (i.e., accepting $\rho_q > h$ when $\rho_q < h$ or accepting $\rho_q < h$ when $\rho_q > h$) is equal to $\alpha/2$ (Jones & Tukey, 2000).

A test of H_0: $\rho_{q1} = \rho_{q2}$ can be obtained from Equations (4a) and (4b) using the following three-decision rule: if the lower limit is greater than 0, then reject H_0 and accept $\rho_{q1} > \rho_{q2}$; if the upper limit is less than 0, then reject H_0 and accept $\rho_{q1} < \rho_{q2}$; otherwise, the results are inconclusive. The probability of a directional error (i.e., accepting $\rho_{q1} > \rho_{q2}$ when $\rho_{q1} < \rho_{q2}$ or accepting $\rho_{q1} < \rho_{q2}$ when $\rho_{q1} > \rho_{q2}$) is equal to $\alpha/2$. In the above example where internal consistency reliability of the eight-item leadership scale for male and female employees was compared, the 95 percent confidence interval for the difference (male minus female) was [.008, .094]. Since the lower limit is greater than 0, we can reject H_0: $\rho_{81} = \rho_{82}$ and accept $\rho_{81} > \rho_{82}$. This result suggests that the leadership questionnaire is more reliable for male employees than female employees.

A test of H_0: $|\rho_{q1} - \rho_{q2}| > h$ (called an "equivalence test"), where h represents a small or unimportant difference in reliability values, also can be obtained from Equations (4a) and (4b) using the following rule: if the lower $100(1-\alpha)\%$ limit is greater than $-h$ *and* the upper $100(1-\alpha)\%$ limit is less than h, then accept $|\rho_{q1} - \rho_{q2}| < h$; if the lower $100(1-\alpha)\%$ limit is greater than h *or* if the upper $100(1-\alpha)\%$ limit is less than $-h$, then accept $|\rho_{q1} - \rho_{q2}| > h$; otherwise, the results are inconclusive. The probability of falsely accepting $|\rho_{q1} - \rho_{q2}| < h$ is equal to $\alpha/2$ (Wellek, 2010). In the two-group leadership reliability study, suppose we believe that $h = .04$ represents a small and unimportant difference in reliability values. The 95 percent confidence interval for $\rho_{81} - \rho_{82}$ is [.008, .094] which straddles the .04 value and so the results are inconclusive. Ninety percent confidence intervals are typically used in equivalence testing (Wellek, 2010). We recomputed the confidence interval for $\rho_{81} - \rho_{82}$ using 90 percent confidence and obtained [.014, .086] which also gives an inconclusive result.

Note that all of the above hypothesis tests are within the class of "informative" tests described by Bonett and Wright (2007, 2009) and do not have the limitations of "non-informative" tests that have been the focus of criticisms leveled at traditional significance testing approaches. Note also that our test of H_0: $|\rho_{q1} - \rho_{q2}| > h$ is a much simpler alternative to the Bayesian approach proposed by Kruschke, Aguinis, and Joo (2012).

Simulation Study

Two competing approaches to Equation (1) have been proposed that do not require the strong assumption of parallel measurements. Duhachek and Iacobucci (2004) proposed a Wald confidence interval $\hat{\rho}_q \pm z_{\alpha/2}\sqrt{var\left(\hat{\rho}_q\right)}$, where $var\left(\hat{\rho}_q\right)$ is given by van Zyl et al. (2000). The Wald interval is attractive because of its computational simplicity. Kistner and Muller (2004) proposed a computationally intensive confidence interval for ρ_q with a general covariance structure and they conducted a simulation study which showed that their method performed remarkably well even with sample sizes as small as $n = 10$. Bonett (2010) conjectured that a confidence interval similar to Equation (1) should perform better than the Duhachek–Iacobucci method and possibly as well as the Kistner–Muller method. We conducted a simulation study to compare the performance of Equation (1) with the two other methods using the exact same sample sizes and covariance structures used by Kistner and Muller (2004). The results, based on 100 000 Monte Carlo replications, are shown in Table 1. Following Kistner and Muller (2004), we simulated data from a multivariate normal distribution.

It can be seen in Table 1 that Equation (1) performs substantially better than the Duhachek–Iacobucci method and slightly better than the computationally intensive Kistner–Muller confidence interval. The results for the Kistner–Muller confidence interval in Table 1 are taken directly from the results reported by Kistner and Muller (2004). As noted by Bonett (2010), all three confidence intervals are expected to be anti-conservative (i.e., the actual

DOI: 10.1002/job

Table 1. Coverage probabilities for three 95 percent confidence interval methods ($q = 4$).

ρ	n	Method 1	Method 2	Method 3
.2	10	.9345	.9102	.9518
	50	.9470	.9428	.9499
	100	.9485	.9466	.9499
	200	.9485	.9487	.9502
.8	10	.9330	.9064	.9510
	50	.9465	.9423	.9502
	100	.9485	.9461	.9501
	200	.9495	.9481	.9500

Note: Method 1 is the computationally intensive method proposed by Kistner and Muller (2004), Method 2 is the Wald confidence interval proposed by Duhachek and Iacobucci (2004), and Method 3 is the confidence interval proposed here (Equation (1)). The covariance matrix represents congeneric measurements with variances of [1 2 3 4] and an AR(1) correlation structure with parameter ρ.

coverage probability is less than the nominal value) with leptokurtic (i.e., peaked and long tailed) distributions and conservative (i.e., the actual coverage is greater than the nominal value) with platykurtic (i.e., flat and short tailed) distributions.

Sample Size Requirements for Desired Precision

The width of Equations (2), (4a), and (4b) depends on the sample size with larger sample sizes giving narrower intervals. A narrow confidence interval for ρ_q and $\rho_{q1} - \rho_{q2}$ is desirable because narrower intervals are more informative. In hypothesis testing applications, more powerful tests are obtained with larger sample sizes.

In studies where the goal is to report a confidence interval for ρ_q or $\rho_{q1} - \rho_{q2}$, researchers should use a sample size that will provide the desired level of confidence and a desired confidence interval width. In studies where the goal is to test a hypothesis regarding the value of ρ_q or $\rho_{q1} - \rho_{q2}$, researchers should use a sample size that will provide the desired level of power for some specified α level and effect size.

We now give some simple sample size formulas for achieving desired confidence interval width or desired power. These sample size formulas are approximations that are accurate if the measurements are parallel and will slightly overstate the required sample size if the measurements are essentially tau-equivalent. We could easily derive the sample size requirements for congeneric measurements, but those sample size formulas would then require the researcher to specify planning values for all q variances and all $q(q-1)/2$ covariances. Unless these planning values happen to be very similar to the unknown population variances and covariances, the congeneric sample size approximations could be less accurate than the approximations from the simple formulas we propose. Our sample size formulas only require the researcher to specify a planning value for Cronbach's alpha reliability.

Let $\widetilde{\rho}_q$ denote a planning value for Cronbach's alpha reliability. This planning value could be obtained from a pilot study, a literature review, or expert opinion. The required sample size to obtain a $100(1-\alpha)\%$ confidence interval for ρ_q with desired width (w) can be obtained by first computing a preliminary sample size approximation

$$n_0 = [8q/(q-1)]\left(1-\widetilde{\rho}_q\right)^2\left(z_{\alpha/2}/w\right)^2 + 2 \tag{5}$$

where $z_{\alpha/2}$ is a critical two-sided z-value (e.g., $z_{\alpha/2} = 1.96$ for 95 percent confidence). Next, compute Equation (3) or Feldt's confidence interval using the preliminary sample size from Equation (5) and with $\widetilde{\rho}_q$ replacing $\hat{\rho}_q$. Compute the width of this interval (denoted as w_0) and, following Bonett and Wright (2000), compute the following more accurate sample size approximation using the following equation

DOI: 10.1002/job

$$n = (n_0 - 2)(w_0/w)^2 + 2 \tag{6}$$

Suppose we want to obtain a 95 percent confidence interval for Cronbach's alpha reliability of a new seven-item job satisfaction scale that has a width of about .1. A pilot study suggests that ρ_7 will be about .8. Copy the commands for Function 3 (see Appendix) at the R prompt and then enter the command

```
> sizeCIrel(.05, 7, .1, .8)
```

which will display a sample size requirement of 147.

The required sample size per group to obtain a $100(1-\alpha)\%$ confidence interval for $\rho_{q1} - \rho_{q2}$ with desired width can be obtained by first computing a preliminary per group sample size approximation

$$n_0 = [8q/(q-1)]\left[\left(1-\widetilde{\rho}_{q1}\right)^2 + \left(1-\widetilde{\rho}_{q2}\right)^2\right](z_{\alpha/2}/w)^2 + 2 \tag{7}$$

where $\widetilde{\rho}_{qi}$ is a planning value of ρ_{qi}. Next, compute Equation (3) or Feldt's confidence interval for each reliability using the preliminary sample size from Equation (7) and with $\widetilde{\rho}_{qi}$ replacing $\hat{\rho}_{qi}$. Finally, compute Equations (4a) and (4b), compute its width (w_0), and then compute a more accurate per group sample size approximation using Equation (6).

Consider a study where we want to compare the reliabilities for the average of $q = 4$ peer performance reviews for minority and non-minority employees. We want a 95 percent confidence interval for $\rho_{41} - \rho_{42}$ that has a width of about .15. From previous research, we set $\widetilde{\rho}_{41} = .72$ and $\widetilde{\rho}_{42} = .85$. Copy the commands for Function 4 (see Appendix) at the R prompt and then enter the command

```
> sizeCIrel2(.05, 4, .15, .72, .85)
```

which will display a sample size requirement of 189 per group.

Sample Size Requirements for Desired Power

The required sample size to test H_0: $\rho_q = h$ for a given α value and with desired power is approximately

$$n = [2q/(q-1)](z_{\alpha/2} + z_\beta)^2/\left(\widetilde{\rho}_q^* - h^*\right)^2 + 2 \tag{8}$$

where $\widetilde{\rho}_q^* = ln(1-\widetilde{\rho}_q)$, $h^* = ln(1-h)$, z_β is a critical one-sided z-value, and $\beta = 1 -$ power (see Bonett, 2002).

To illustrate the use of Function 5 (see Appendix), consider a reliability study of our five-item ($q = 5$) work ethic questionnaire where we will test H_0: $\rho_5 = .65$ at $\alpha = .05$ with $\widetilde{\rho}_5 = .80$. Copy the Function 5 commands at the R prompt. The sample size required to test H_0: $\rho_5 = .65$ at $\alpha = .05$ with power of .9 is obtained by entering the following command

```
> sizePOWrel(.05, 5, .9, .8, .65)
```

which will display a sample size requirement of 86.

The required sample size per group to test H_0: $\rho_{q1} = \rho_{q2}$ for a given α value and with desired power is approximately

$$n = [4q/(q-1)](z_{\alpha/2} + z_\beta)^2/\left(\widetilde{\rho}_{q1}^* - \widetilde{\rho}_{q2}^*\right)^2 + 2 \tag{9}$$

where $\widetilde{\rho}_{q1}^* = ln(1-\widetilde{\rho}_{q1})$ and $\widetilde{\rho}_{q2}^* = ln(1-\widetilde{\rho}_{q2})$ (see Bonett, 2003).

DOI: 10.1002/job

To illustrate the use of Function 6 (see Appendix), consider a study where we want to compare the reliabilities of a nine-item ($q = 9$) self-regulation scale for male and female employees and we want to test H_0: $\rho_{91} = \rho_{92}$ at $\alpha = .05$ with $\widetilde{\rho}_{91} = .70$ and $\widetilde{\rho}_{92} = .80$. Copy the Function 6 commands at the R prompt. The sample size required to test H_0: $\rho_{91} = \rho_{92}$ at $\alpha = .05$ with power of .85 is obtained by entering the following command

```
> sizePOWrel2(.05, 9, .85, .7, .8)
```

which will display a sample size requirement of 248 per group.

The sample size requirements in the above examples might be too costly for some researchers, especially in studies where participants must be remunerated or the measurement process is time-consuming. In these situations, reliability estimates from two or more studies can be combined as described by Bonett (2010) to obtain more precise confidence intervals or more powerful tests. Reliability estimates from two or more studies, where participants differ in terms of interesting demographic characteristics, can be compared using Equations (4a) and (4b). If reliability differences are detected across certain demographic groups (male vs. female, blue collar vs. white collar, minority vs. non-minority, etc.) or testing conditions (phone interview vs. online survey, group testing vs. individual testing, and timed vs. untimed), then future studies should strive to obtain accurate reliability estimates within each demographic or testing condition.

Best-practice Recommendations

The sample value of Cronbach's alpha should be accompanied by a confidence interval for its population value. Equation (1) is the recommended interval estimation approach in applications where the measurements have an approximate normal distribution. Equation (1) is expected to have a coverage probability that is less than $1 - \alpha$ when all measurements are leptokurtic (i.e., excess kurtosis > 0) and a coverage probability that is greater than $1 - \alpha$ when the measurements are platykurtic (i.e., excess kurtosis < 0). Thus, Equation (1) will be useful in applications where the normality assumption has been violated if the measurements are platykurtic or mildly leptokurtic.

Data transformation (e.g., log, square-root and reciprocal) of all q measurements may help reduce leptokurtosis. If the measurements are highly leptokurtic and data transformations are ineffective or inappropriate, $var\left(\hat{\rho}_q\right)$ in Equation (1) could be replaced with the asymptotically distribution-free (ADF) variance estimate proposed by Yuan, Guarnaccia, and Hayslip (2003). Alternatively, a bootstrap confidence interval for ρ_q (Padilla, Divers, & Newton, 2012) could be used. However, unlike Equation (1) which performs properly under normality with sample sizes as small as 10, the ADF and bootstrap methods require much larger sample sizes before they will perform properly. Equation (1) should also perform properly with q dichotomous measurements as long as the item difficulties are within the range of about .2 to .8 (Bonett, 2010).

If the measurements of interest have been obtained for $i \geq 2$ demographic subgroups or testing conditions, Equation (1) could be used to obtain a confidence interval for each ρ_{qi} and Equations (4a) and (4b) could be used for any interesting pairwise comparison. These results could provide important reliability information that might qualify the results of certain statistical analyses. For instance, if predictive validity correlation coefficients differ based on employee ethnicity, this difference could be due to differential reliability in the outcome variable or the predictor variable. Unless the study was specially designed with adequate subgroup sample sizes to assess these individual reliabilities, the confidence interval for ρ_{qi} could be wide. However, the subgroup reliability estimates should be reported so that future research can combine these estimates with subgroup estimates from other studies using the meta-analysis methods described by Bonett (2010). In addition to reporting a confidence interval for the population value of Cronbach's alpha, we recommend that correlations and variances of the q measurements be reported in a table, appendix, or online source which will enable other researchers to compute Equation (1) and also combine or compare reliability results from multiple studies.

DOI: 10.1002/job

Although Equations 5–9 are not difficult to compute by hand, most researchers will want to perform these sample size computations for several different input values and then choose a sample size that represents the best compromise between the cost of the sample and the benefits of the precision or power. The R functions provided in the Appendix will assist researchers in the rapid computation of these sample size approximations.

Organizational researchers would never consider reporting only sample means, sample correlations, or sample regression coefficients without also reporting appropriate inferential results (i.e., hypothesis test results or confidence intervals). Likewise, a sample value of Cronbach's alpha should be accompanied by a confidence interval. Many years ago, Leonard Feldt, the illustrious psychometrician, expressed concern about the (then) widespread practice of reporting only sample reliability coefficients without supplemental inferential results (Feldt, 1965). Sadly, the problem of reporting only the sample value of Cronbach's alpha continues to this day despite Feldt's compelling arguments and the requirement of the *American Psychological Association* to report confidence intervals for all important parameters. We close with the recommendation that journal editors require sample values of Cronbach's alpha be supplemented with an appropriate confidence interval.

Author biographies

Douglas G. Bonett is the Director of the Center for Statistical Analysis in the Social Sciences at UC Santa Cruz. He received his PhD from the University of California, Los Angeles. He has taught a wide range of applied statistics courses and has received several research and teaching awards. His current research focuses on interval estimation methods and meta-analysis problems.

Thomas A. Wright is the Felix E. Larkin Distinguished Professor in Management at Fordham University's Gabelli School of Business. He received his PhD from the University of California, Berkeley. The highlight of his professional career has been publishing a number of articles on ethics with his father, Vincent P. Wright.

References

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 235–240.

Bonett, D. G. (2003). Sample size requirements for comparing two alpha reliability coefficients. *Applied Psychological Measurement*, *27*, 235–240.

Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods*, *13*, 99–109.

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, *15*, 368–385.

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. *Psychometrika*, *65*, 23–28.

Bonett, D. G., & Wright, T. A. (2007). Comments and recommendations regarding the hypothesis testing controversy. *Journal of Organizational Behavior*, *28*, 647–659.

Bonett, D. G., & Wright, T. A. (2009). Confidence intervals in supply chain and operations research. *The Journal of Supply Chain Management*, *45*, 26–33.

Cronbach, L. J. (1951). Coefficient alpha and the interval structure of tests. *Psychometrika*, *16*, 297–334.

Duhachek, A., & Iacobucci, D. (2004). Alpha standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*, 792–808.

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

Hedges, L. V., & Olkin, I. (1980). *Statistical methods for meta-analysis*. San Diego: Academic Press.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, *5*, 411–414.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–133.

Kistner, E. O., & Muller, K. E. (2004). Exact distribution of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika*, *69*, 459–474.

DOI: 10.1002/job

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*, 722–752.
McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: LEA.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd edn). New York: McGraw-Hill.
Padilla, M. A., Divers, J., & Newton, M. (2012). Coefficient alpha bootstrap confidence interval under nonnormality. *Applied Psychological Measurement*, *36*, 331–348.
Sijtsma, K. (2009). Reliability: Beyond theory and into practice. *Psychometrika*, *74*, 169–173.
van Zyl, J., Neudecker, H. M., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*, 271–280.
Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd edn). Boca Raton: Chapman & Hall.
Yuan, K.-H., Guarnaccia, C. A., & Hayslip, B. (2003). A study of the distribution of sample coefficient α with the Hopkins symptom checklist: Bootstrap versus asymptotics. *Educational and Psychological Measurement*, *63*, 5–23.

Appendix R Functions

Function 1: Confidence interval for Cronbach's alpha coefficient

```
CIreliability <- function(alpha, q, n, V) {
 # Computes a confidence interval for Cronbach' s alpha
 # reliability without assuming parallel measurements
 # Args:
 # alpha: alpha value for 1-alpha confidence
 # q:     number of measurements (items, raters, etc.)
 # n:     sample size
 # V:     sample qxq covariance matrix of q measurements
 # Returns:
 #        lower limit, estimated alpha, upper limit
 z <- qnorm(1 - alpha/2)
 b <- log(n/(n - 1))
 j <- matrix(rep(1, q), nrow=q, ncol=1)
 a0 <- t(j)%*%V%*%j
 t1 <- sum(diag(V))
 t2 <- sum(diag(V%*%V))
 a1 <- a0^3
 a2 <- a0*(t2 + t1^2)
 a3 <- 2*t1*t(j)%*%(V%*%V)%*%j
 a4 <- 2*q^2/(a1*(q - 1)^2)
 r <- (q/(q - 1))*(1 - t1/a0)
 var <- a4*(a2 - a3)/(n - 3)
 LL <- 1 - exp(log(1 - r) - b + z*sqrt(var/(1 - r)^2))
 UL <- 1 - exp(log(1 - r) - b - z*sqrt(var/(1 - r)^2))
 out <- c(LL, r, UL)
 return(out)
}
```


DOI: 10.1002/job

Function 2: Confidence interval for the difference of two Cronbach's alpha coefficients

```
CIrelDiff <- function(r1, LL1, UL1, r2, LL2, UL2){
 # Computes a confidence interval for a difference
 # of two Cronbach alpha reliabilities
 # Args:
 #   r1:    sample reliability in group 1
 #   LL1:   group 1 lower limit
 #   UL1:   group 1 upper limit
 #   r2:    sample reliability in group 2
 #   LL2:   group 2 lower limit
 #   UL2:   group 2 upper limit
 #   Returns:
 #          lower limit, estimated difference, upper limit
 d <- r1 - r2
 LL <- d - sqrt((r1 - LL1)^2 + (UL2 - r2)^2)
 UL <- d + sqrt((UL1 - r1)^2 + (r2 - LL2)^2)
 out <- c(LL, d, UL)
 return(out)
}
```

Function 3: Sample size to estimate a Cronbach's alpha reliability with desired precision

```
sizeCIrel <- function(alpha, q, w, r){
 # Computes sample size required to estimate a Cronbach alpha
 # reliability with desired precision
 # Arguments:
 #   alpha:   alpha value for 1-alpha confidence
 #   q:       number of measurements
 #   w:       desired CI width
 #   r:       reliability planning value
 # Returns:
 #           required sample size
 z <- qnorm(1 - alpha/2)
 n0 <- ceiling((8*q/(q - 1))*(1 - r)^2*(z/w)^2 + 2)
 b <- log(n0/(n0 - 1))
 LL <- 1 - exp(log(1 - r) - b + z*sqrt(2*q/((q - 1)*(n0 - 2))))
 UL <- 1 - exp(log(1 - r) - b - z*sqrt(2*q/((q - 1)*(n0 - 2))))
 w0 <- UL - LL
 n <- ceiling((n0 - 2)*(w0/w)^2 + 2)
 return (n)
}
```


DOI: 10.1002/job

Function 4: Sample size to estimate a difference in two Cronbach alpha reliabilities

```
sizeCIrel2 <- function(alpha, q, w, r1, r2){
 # Computes sample size per group required to estimate a difference
 # of two Cronbach alpha reliabilities with desired precision
 # Arguments:
 # alpha: alpha value for 1-alpha confidence
 #   q:    number of measurements
 #   w:    desired CI width
 #   r1:   reliability planning value
 #   r2:   reliability planning value
 # Returns:
 #         required sample size per group
 z <- qnorm(1 - alpha/2)
 n0 <- ceiling((8*q/(q - 1))*((1 - r1)^2 + (1 - r2)^2)*(z/w)^2 + 2)
 b <- log(n0/(n0 - 1))
 LL1 <- 1 - exp(log(1 - r1) - b + z*sqrt(2*q/((q - 1)*(n0 - 2))))
 UL1 <- 1 - exp(log(1 - r1) - b - z*sqrt(2*q/((q - 1)*(n0 - 2))))
 LL2 <- 1 - exp(log(1 - r2) - b + z*sqrt(2*q/((q - 1)*(n0 - 2))))
 UL2 <- 1 - exp(log(1 - r2) - b - z*sqrt(2*q/((q - 1)*(n0 - 2))))
 LL <- r1 - r2 - sqrt((r1 - LL1)^2 + (UL2 - r2)^2)
 UL <- r1 - r2 + sqrt((UL1 - r1)^2 + (r2 - LL2)^2)
 w0 <- UL - LL
 n <- ceiling((n0-2)*(w0/w)^2 + 2)
 return (n)
}
```

Function 5: Sample size to test a Cronbach's alpha reliability with desired power

```
sizePOWrel <- function(alpha, q, pow, r, h){
 # Computes sample size required to test a Cronbach
 # alpha reliability with desired power
 # Arguments:
 #   alpha: alpha value for test
 #  q:    number of measurements
 #   r:   reliability planning value
 #  h:    null hypothesis value of reliability
 # Returns:
 #        required sample size
 za <- qnorm(1 - alpha/2)
 zb <- qnorm(pow)
 e <- (1 - r)/(1 - h)
 n <- ceiling((2*q/(q - 1))*(za + zb)^2/log(e)^2 + 2)
 return(n)
}
```


DOI: 10.1002/job

Function 6: Sample size to test equality of two Cronbach's alpha reliabilities with desired power

```
sizePOWrel2 <function(alpha, q, pow, r1, r2){
 # Computes sample size required per group to test equality
 # of two Cronbach alpha reliabilities with desired power
 # Arguments:
 #   alpha: alpha value for test
 #   q:   number of measurements
 #   r1:   group 1 reliability planning value
 #   r2:   group 2 reliability planning value
 # Returns:
 #         required sample size per group
 za <- qnorm(1 - alpha/2)
 zb <- qnorm(pow)
 e <- (1 - r1)/(1 - r2)
 n <- ceiling((4*q/(q - 1))*(za + zb)^2/log(e)^2 + 2)
 return(n)
}
```


DOI: 10.1002/job