Behavior Research Methods
2010, 42 (3), 863-870
doi:10.3758/BRM.42.3.863

Synthesizer 1.0: A varying-coefficient meta-analytic tool

Zlatan Krizan
Iowa State University, Ames, Iowa

Meta-analysis has become an indispensable tool for reaching accurate and representative conclusions about phenomena of interest within a research literature. However, in order for meta-analytic computations to provide accurate estimates of population parameters (e.g., a population correlation), underlying statistical models need to be both efficient and unbiased. Current fixed-effect (i.e., constant-coefficient) models that assume a common effect for all research results perform poorly under conditions of effect size heterogeneity, whereas current random-effects (i.e., random-coefficient) models require unrealistic assumptions about random sampling of observed effect sizes from a normally distributed superpopulation. This article describes a free statistical software tool that employs a varying-coefficient model recently proposed by Bonett (2008, 2009). The software (Synthesizer 1.0) employs procedures that do not require effect homogeneity or random sampling of effect sizes from a normal distribution. It may be used to meta-analyze correlations, alpha reliabilities, and standardized mean differences. The Synthesizer tool for Microsoft Excel 2007 may be downloaded from the author at www.psychology.iastate.edu/~zkrizan/Synthesizer.htm or as a supplement to the article at http://brm.psychonomic-journals.org/content/supplemental.

Around the turn of the 20th century, communicable infectious diseases were a great threat to human populations. It was only with the advent of vaccination protocols that many of such modern "plagues" were successfully controlled—indeed, today's children are routinely vaccinated against diseases such as measles and smallpox. Karl Pearson, one of the fathers of modern statistics, lived during this period. Given the controversies that existed at the time, he too was interested in the efficacy of newly developed inoculation procedures for combating vulnerability to such infections. In what can probably be called the first meta-analysis, Pearson (1904) collected correlation coefficients to examine whether inoculation against smallpox predicted survival. Furthermore, he *quantitatively* aggregated these correlations, yielding an unweighted average correlation of .63, a truly massive effect size with enormous real-world significance. Although it took a hiatus for a large part of the 20th century, due to seminal work of pioneers such as Bob Rosenthal (1976) and Gene Glass (1976), meta-analysis became a critical tool for reaching accurate conclusions about empirical findings and resolving sticky questions about whether and when particular effects manifest. Today, meta-analysis is one of the most popular methods for conducting secondary research and literature reviews—a search of the PsycINFO database revealed that 630 publications with the term "meta-analysis" in the abstract were published in 2008, but only 11 were published in 1980.

Although there is no denying that quantitative synthesis of empirical findings is both prevalent and desirable (see, e.g., Cooper & Hedges, 1994), there is much less agreement about *how* to best statistically aggregate research results. Answering this question involves everything from fairly high-level considerations (e.g., how to conceptualize the research question, how to define relevant research findings, how to determine what qualifies as an appropriate test of relevant hypotheses) to lower level considerations of how to choose a relevant metric and what statistical models to use for aggregation of the metrics (e.g., Cooper, Hedges, & Valentine, 2009). Perhaps the most contentious issues revolve around the last question: What statistical procedures are the most appropriate when aggregating effect sizes? In this article, I briefly review the two dominant statistical models employed in meta-analyses, emphasize their limitations, and describe a new, no-cost statistical software tool that employs novel statistical techniques (Bonett, 2008, 2009) and overcomes some of the limitations of traditionally used meta-analytic techniques.

STATISTICAL MODELS IN META-ANALYSIS

In order for a researcher to infer "true" values of a particular effect size (e.g., how strong is the link between smoking and lung cancer in an actual population) based on inherently limited empirical data, he or she must specify a "scheme" (generally a mathematical function) for how a given empirical finding relates to the actual state of

Z. Krizan, zkrizan@iastate.edu

the world. For example, a common scheme is to assume that a given effect size observed in a single study is a function of the "true" population effect size, and some amount of error inherent in examining only a limited sample of that population. Given such a scheme, the meta-analyst attempts to infer the actual population values based on the limited set of empirical data that is available. Below, I briefly describe the two most common models employed in meta-analyses, highlight their assumptions, expose how the issue of heterogeneity is handled in each, and argue for the value of an alternative approach.

Constant-Coefficient Models

Perhaps the most prevalent statistical model in meta-analyses is what is commonly referred to as a *fixed-effect* or *common-effect* model (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009). Such a model is most accurately called a *constant-coefficient* model (Bonett, 2010), given that all studies included in the meta-analysis are assumed to reflect a single, constant, common effect size. In other words, there is only one "true" value of the relationship in question (e.g., the link between smoking and lung cancer is the same for everyone across all settings and periods of time), and the purpose of the meta-analytic technique is to infer the best estimate of this common effect size. Bonett (2010) provides the following mathematical expression for a constant-coefficient model:

$$\hat{\rho}_j = \rho + \varepsilon_j. \quad (1)$$

In this expression, the expected value of sample estimates ($\hat{\rho}_j$) is assumed to equal an unknown constant (i.e., fixed-effect) ρ. The unobservable errors ε_j are assumed to be independent and randomly distributed. The model assumes stratified random sampling such that a random sample is obtained from each study population (Bonett, 2010, p. 6). In simple terms, the effect size from a given study is a linear combination of the true effect size and random error (as in classic reliability theory; see Spearman, 1910). Constant-coefficient models underlie popular techniques proposed by Hedges and Olkin (1985) and Hedges and Vevea (1998) for aggregation of correlations and standardized mean differences (see also Rodriguez & Maeda, 2006; Shadish & Haddock, 1994), and have been used in hundreds of meta-analyses. Given the assumption that the only source of error in estimates is sampling error of participants into studies, fixed-effect models use a weighted average as an estimate of the population effect size, which affords much greater weight to estimates coming from larger samples (e.g., via the inverse variance adjustment; Hedges & Olkin, 1985).

Thus, for methods based on constant-coefficient models, inference applies only to the *m* study populations represented in the *m* studies ("conditional inference"; see Hedges & Vevea, 1998). The population estimate of the effect size (based on the weighted average) will be a meaningful parameter to estimate *only if* the population effect sizes are approximately equal across the *m* study populations (i.e., there are no substantial moderating factors). However, when population effect sizes are not identical, weighted estimates can be severely biased, especially when sample sizes are unequal across studies (see Bonett, 2008). Given that the assumption of equal population effect sizes (i.e., a common effect) is basically never true in practice (Hunter & Schmidt, 2000, 2004), the National Research Council (1992) has recommended that the use of classic fixed-effect methods be restricted.

Random-Coefficient Models

The other prevalent statistical model employed in meta-analyses is the random-effects model (e.g., Hunter & Schmidt, 2004). Such *random-coefficient* models do not have the unrealistic assumption of effect size homogeneity in the population—in other words, they allow for multiple effect sizes in the population (e.g., the link between smoking and lung cancer might be greater among those who have a genetic predisposition to cancer). However, they assume that studies that are meta-analyzed have been *randomly* sampled from a clearly defined superpopulation that follows a normal distribution. Bonett (2010) provides the following mathematical expression for a random-coefficient model:

$$\hat{\rho}_j = \rho_j + \varepsilon_j. \quad (2)$$

In this expression, sample estimates $\hat{\rho}_j$ are independent and unobservable random variables that conform to a normal distribution (defined by a grand mean ρ and unknown variance τ^2). The ρ_j sample values are thought to be orthogonal from the unobservable errors ε_j. The model assumes two-stage cluster sampling such that a random sample is obtained from different study populations (each with a mean of ρ_j) that have themselves been randomly sampled from a superpopulation of studies. The parameters ρ and τ^2 are the unknown mean and variance population parameters in this superpopulation of studies (Bonett, 2010, pp. 8–9). For example, the critical parameters might involve the average (ρ) of correlational links between smoking and lung cancer across groups with different genetic predispositions and health histories ($\rho_1, \rho_2, \ldots, \rho_N$), and the dispersion of these links across this superpopulation (τ^2). In simple terms, the effect size from a given study is a linear combination of the true effect size from a relevant population (that is randomly sampled from a normally distributed superpopulation of effect sizes with a given mean and variance) and random error, the two components being orthogonal. Random-coefficient models underlie popular techniques proposed by Hunter and Schmidt (2000, 2004) and Hedges and Vevea (1998), and have been frequently utilized.

Random-coefficient models have been recommended because they do not require the unrealistic assumption of effect size homogeneity in the superpopulation. However, the meta-analyst cannot make a valid statistical inference about the superpopulation mean effect size and variance unless the different study populations have been *randomly* sampled from the superpopulation. Unfortunately, this is basically never justified in practice (Hedges & Vevea, 1998). Studies in a given research literature are intentionally designed to be similar or dissimilar to previous studies

(e.g., to replicate or extend prior findings), and will thus never conform to the random-sampling assumption. Additionally, an estimate of the variance of population effect sizes (τ^2) is a critical value for accurate construction of confidence intervals in random-coefficient models. However, this value is extremely difficult to estimate with any precision when nonnormality, leptokurtosis, or a small number of studies is involved (e.g., Bonett, 2010; Miller, 1986). In sum, random-coefficient models also have serious limitations; as noted by Schulze (2004, p. 41), the random-sampling assumption might be the critical feature limiting the applicability of random-effects models (see also Shuster, 2010).

Assessing Effect Size Heterogeneity

As elucidated above, both constant-coefficient and random-coefficient models have serious limitations when being used to estimate the overall population effect size. Moreover, meta-analysts are seldom interested in only a single estimate of an overall population effect (e.g., Borenstein et al., 2009). Rather, the critical questions often involve various tests of moderating factors (e.g., is the link between smoking and lung cancer larger among those individuals who smoke more?). At the core of such questions lies assessment of effect size heterogeneity—namely, an examination of variability in effect sizes, and an estimation of the degree to which this variability is substantial and the extent to which it reflects the presence of moderating factors.

In constant-coefficient models, the running assumption is one of no heterogeneity, because one common population effect is expected. A typical way of empirically examining this assumption is to compute the Q statistic (Cochran, 1954), which is the sum of the squared deviations of each study's effect estimate from the overall effect estimate, weighting the contribution of each study by its inverse variance. Usually, if this value is statistically significant at the .05 level, one concludes that heterogeneity is present and a random-coefficient model is adopted, whereas lack of significance is taken as evidence that a common-effect view is justified and the analyst proceeds with the estimation using the constant-coefficient model (e.g., Field, 2001). However, each of these conclusions is inappropriate. First, a significant Q test does not indicate a meaningfully large dispersion, because a significant p value could reflect a minor amount of dispersion from fairly precise studies (see Bonett & Wright, 2007; see also Borenstein et al., 2009, p. 113). Furthermore, even if substantial dispersion exists, this does not warrant adoption of random-coefficient models given that they replace the constant-coefficient assumption with an additional and equally unrealistic assumption that the m studies have been randomly selected from a well-defined and normally distributed superpopulation of study populations (Schulze, 2004). Second, a nonsignificant Q test does not indicate a lack of true heterogeneity since it often suffers from low power (e.g., due to a small number of studies), which is why most meta-analysts agree that one may proceed with more focused tests of moderation regardless of the result of this test (see Borenstein et al., 2009). Other heterogeneity indices based on the Q statistic involve similar ambiguities. For example, the I^2 statistic reflects the proportion of between-study variation relative to within-study variation; that is, it estimates the proportion of *real* variance that is not due to sampling error (see Higgins, Thompson, Deeks, & Altman, 2003). Again, however, the I^2 is mute with respect to whether the "real" dispersion is substantial enough to be of practical or theoretical interest (cf. Borenstein et al., 2009, p. 119). As advised below, heterogeneity is easily accessed by examining the range in study estimates (or their dispersion), and performing focused tests of theoretically relevant subgroup differences (e.g., confidence intervals of differences in correlations). This approach should be preferred because it yields information about the magnitude of the difference and its precision without relying on rather arbitrary decisions about whether heterogeneity is present or not.

Introducing a Varying-Coefficient Model

Given the important limitations of constant- and random-coefficient models, on many occasions it would be advisable to employ a varying-coefficient model (e.g., Bonett, 2008, 2009, 2010), which does not assume a common effect or random sampling of study populations from a normally distributed, well-defined superpopulation of studies. Bonett (2010) provided the following expression for a varying-coefficient model:

$$\hat{\rho}_j = \rho_j + \varepsilon_j. \tag{3}$$

In this expression, ρ_j is the expected value of sample estimates ($\hat{\rho}_j$). The sample values are unknown constants (i.e., fixed effects) that need not be equal. The model assumes stratified random sampling such that a random sample is obtained from each study population (Bonett, 2010, p. 8). When effect sizes in the population are not too disparate, an unweighted average would be a meaningful parameter to estimate. Although intuition suggests that weighted averages will have a smaller mean squared error, this is not necessarily true for a typical meta-analysis in which population effects are not identical and sample sizes show variability across studies (see Bonett, 2008). If strong moderating factors are hypothesized, certain linear contrasts of ρ_j would be of more interest. Varying-coefficient models have been used for meta-analytic confidence intervals of Pearson correlations and their differences (Bonett, 2008), meta-analytic confidence intervals of standardized and unstandardized mean differences (Bonett, 2009), and meta-analytic confidence intervals of internal consistency estimates (Bonett, 2010). Additional information about varying-coefficient models can be found in Judge, Griffiths, Hill, Lutkepohl, and Lee (1985, chap. 19).

Using Monte Carlo methods, Bonett (2008, 2009, 2010) showed that the confidence interval based on the varying-coefficient model dramatically outperformed traditional techniques that employ constant- and random-coefficient models. In a case of Pearson correlations involving 50,000

trials, the varying-coefficient 95% confidence intervals had an average coverage probability that ranged from .949 to .952 across all the conditions, with minimum coverage never falling below .943 (Bonett, 2008, Table 1). In contrast, the true coverage probability for traditional fixed-effect models could be as small as 0 in some cases. In the case of random-effects models, the coverage probabilities could be much less than the specified .95 value and the average interval width was consistently higher than in the variable-coefficient model. Similar results were obtained when estimating unstandardized and standardized mean differences (Bonett, 2009) or alpha reliabilities (Bonett, 2010). These results confirmed previous statements that confidence intervals based on fixed-effect models are too narrow (e.g., Hunter & Schmidt, 2004), while suggesting that confidence intervals based on random-effects models may be unnecessarily wide.

SYNTHESIZER 1.0

Given the prevalence and importance of meta-analyses mentioned earlier (cf. Cooper et al., 2009), it is important to maximize researchers' opportunity to perform quantitative syntheses and provide the largest possible flexibility in relevant methods. Multiple meta-analytic software packages already exist, and can be found at www.psychwiki.com/wiki/Meta-analysis. Some are free to researchers and are relatively simple (e.g., Meta-Analysis Calculator, by Larry C. Lions), others offer a slew of additional statistical analysis features (e.g., the R Project), whereas more comprehensive programs offer considerable flexibility in converting among effect sizes and graphing results—although with a hefty price tag (e.g., Comprehensive Meta-Analysis; Borenstein & Rothstein, 2006). However, all of these packages rely only on the constant- or random-coefficient models described earlier and are thus limited in their applicability. Furthermore, most of the packages that are free to researchers do not allow for easy estimation of population reliabilities, tests of moderation, or construction of confidence intervals regarding differences or linear contrasts of effect sizes. In order to accommodate these needs, I have developed a free program for quantitative synthesis of effect sizes with the name *Synthesizer 1.0*. The employed aggregation computations rely on the varying-coefficient model described above and developed in Bonett (2008, 2009, 2010). This program employs Microsoft Excel 2007 and requires only basic familiarity with its functions. Moreover, it allows for reliability generalization via aggregation of internal consistency estimates, and construction of confidence intervals about correlation or reliability differences. Although it does not have all the features of expensive analysis suites, it is accessible to researchers, provides flexibility of customization, and uses cutting-edge computations. Below, I briefly describe the basic functions of Synthesizer 1.0 and how to use them, followed by an illustrative example. Note that this is the first version of the software, and the program is thus likely to undergo revisions and expansions of its capabilities.

Basic Features of Synthesizer 1.0

In this section I describe only the basic features of the software, with the aim to inform potential analysts about whether they might benefit from using the program. The program file and the user's guide containing precise operational instructions and all of the relevant mathematical formulas (Krizan, 2009) can be found at the program Web site, www.psychology.iastate.edu/~zkrizan/Synthesizer.htm. (A copy is also available in the Psychonomic Society's supplemental archive.) Once the program file has been opened in Excel, the user will see the main sheet, which will include the program title, logo, and a quick guide to the program's functions. The functions are organized within different analysis modules (i.e., sheets) accessible via tabs on the bottom right of the screen. The purpose of different modules is described in the quick guide, as is the basic approach for testing moderation. Below, I briefly outline the purpose of each module and the information necessary to use it.

Correlation Aggregator (*r*)

This module should be used to aggregate Pearson correlations (*r*) from independent samples. It contains columns for entry of study references, subgroups within studies, sample, and study identification numbers. The two critical columns in the middle are "sample size" and "*r*," because they require the analyst to enter the correlation coefficient and the sample size on which that coefficient is based. Given this information (and providing that the user copies the standard error formula to all relevant rows), the module will automatically calculate the variance for each sample. To the left of this critical information, one can specify moderator variables in different columns and enter codes for the moderator variables. The data can then be sorted according to these codes, and separate estimates for each subgroup can be obtained by removing data from the remaining subgroups (by copying the sheet, original data can easily be preserved). Below the data-entry sheet, on the bottom section of the screen, the user can see the basic descriptive statistics regarding the included data: number of studies, observed dispersion of effect sizes, and estimated variance of the population effect size estimate. Of critical interest will be the estimated population correlation and its 95% confidence interval, which appears immediately to the left of the descriptive statistics (see Figure 1). Following Bonett (2008), the population correlation is estimated using the unweighted average of sample estimates from *m* studies:

$$\bar{\rho} = m^{-1}\sum_{i=1}^{m}\hat{\rho}_i. \quad (4)$$

The parameter variance is estimated as follows across studies with sample sizes *n*:

$$\mathrm{var}(\bar{\rho}) = m^{-2}\sum_{i=1}^{m}\left(1-\hat{\rho}_i^2\right)^2 \Big/ \left(n_i - 3\right). \quad (5)$$

In order to normalize the sampling distribution of the correlations, Fisher's (1921) correction ($\tanh^{-1}$) is applied both to the mean estimate and to the computation of the

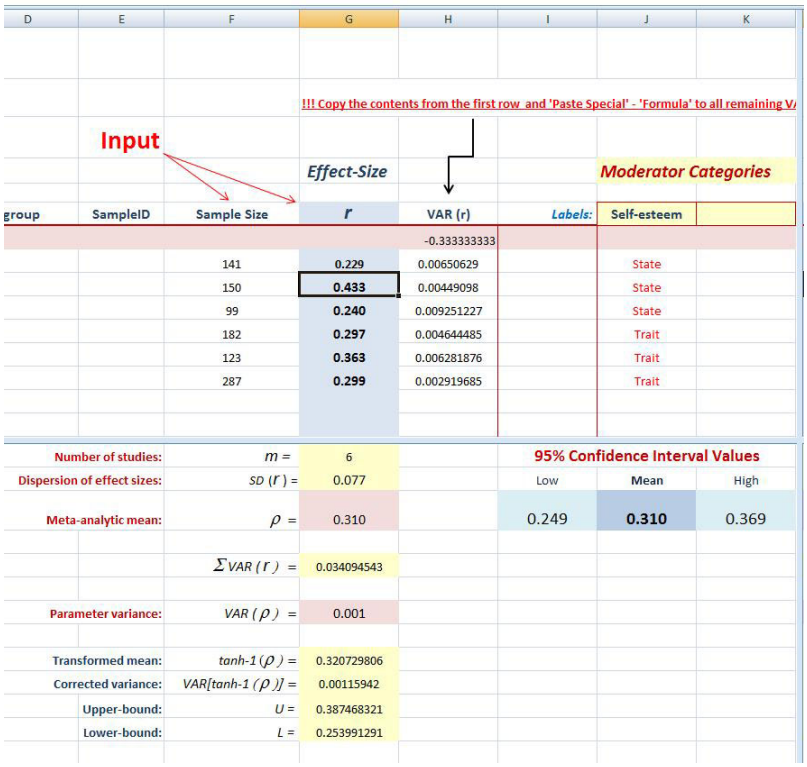

Figure 1. Aggregating correlations in Synthesizer 1.0.

parameter variance [in the latter case, this is performed by dividing the variance estimate by $(1 - \bar{\rho}^2)^2$; see Bonett, 2008]. These values are then used to yield the 95% confidence interval around the population mean for m studies (where $\tanh^{-1}$ represents Fisher's transformation):

$$\tanh^{-1}(\bar{\rho}) \pm z_{\alpha/2}\left\{\text{var}(\bar{\rho}) \Big/ \left(1-\bar{\rho}^2\right)^2\right\}^{-1/2}. \qquad (6)$$

Finally, the interval boundaries are back-transformed [$\tanh(x) = \{\exp(2x) - 1\}/\{\exp(2x) + 1\}$] in order for the confidence interval to be interpreted.

Testing moderation. Very often, the analyst is interested in examining how effect sizes might differ across various subpopulations, rather than in arriving at a single estimate that might be broadly accurate but of little practical or theoretical use. Synthesizer 1.0 allows for estimation of the difference between two sets of correlations, and the construction of the associated 95% confidence intervals. Again, this approach is preferred because it does not depend on significance tests that might suffer from low power and that do not convey the magnitude of the difference, nor its precision. This module is accessed via the [r − Contrasts] tab. On this sheet, the only required input is the estimated mean and respective upper- and lower-bound values from the confidence interval that was estimated for a given group of studies per the procedures described above. Thus, after estimating the confidence interval for a given subgroup (see above), one can then input confidence interval values for each respective subgroup on this sheet. Of critical interest, then, is the estimated difference in population correlations and its 95% confidence interval (see Figure 2). Following Bonett (2008), upper-bound (U) and lower-bound (L) values from confidence intervals around subpopulation meta-analytic estimates are employed to yield the upper- and lower-bound values for the confidence interval around the difference between subpopulation correlations (denoted by A and B subscripts below):

$$L = \bar{\rho}_A - \bar{\rho}_B - \left\{\left(\bar{\rho}_A - L_A\right)^2 + \left(U_B - \bar{\rho}_B\right)^2\right\}^{1/2} \qquad (7)$$

$$U = \bar{\rho}_A - \bar{\rho}_B + \left\{\left(U_A - \bar{\rho}_A\right)^2 + \left(\bar{\rho}_B - L_B\right)^2\right\}^{1/2}. \qquad (8)$$

Mean Difference Aggregator (*d*)

This module should be used to aggregate standardized mean differences (i.e., Cohen's d) from independent samples. Besides standard columns for study descriptors, it requires input of sample sizes for each condition (labeled A and B), and their respective means and standard deviations. Given this information and following Bonett (2009), the module will automatically correct the effect size for bias in small samples (b; cf. Hedges, 1981), yielding the following meta-analytic mean across the m studies:

$$\bar{\delta} = m^{-1}\sum_{i=1}^{m} b_i\hat{\delta}_i. \qquad (9)$$

Following Bonett (2009), the variance of sample means is computed based on variances (σ) and degrees of freedom (df) for each group:

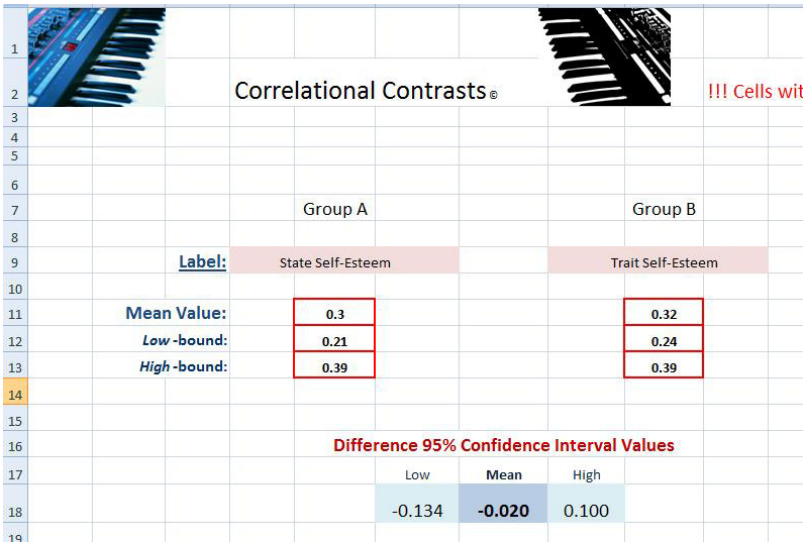

Figure 2. Testing differences between correlations in Synthesizer 1.0.

$$\text{var}\left(\hat{\delta}_i\right) = \left\{\hat{\delta}_i^2\left(\hat{\sigma}_{i1}^4 / df_{i1} + \hat{\sigma}_{i2}^4 / df_{i2}\right) / 8\hat{\sigma}_i^4 + \left(\hat{\sigma}_{i1}^2 / df_{i1} + \hat{\sigma}_{i2}^2 / df_{i2}\right) / \hat{\sigma}_i^2\right\}. \quad (10)$$

Below the data-entry sheet, on the bottom section of the screen, the user can see the basic descriptive statistics regarding the included data: number of studies, observed dispersion of effect sizes, and estimated variance of the population effect size estimate. Columns for coding moderating factors are also included. Of critical interest will be the estimated population mean difference and its 95% confidence interval, which appears immediately to the left of the descriptive statistics and is computed as follows:

$$\bar{\delta} \pm z_{\alpha/2}\left\{m^{-2}\sum_{i=1}^{m} b_i^2 \,\text{var}\left(\hat{\delta}_i\right)\right\}^{1/2}. \quad (11)$$

Testing moderation. Again, often the interest is in examining how effect sizes might differ across various subpopulations. Synthesizer 1.0 allows for estimation of *differences* in standardized mean differences, and construction of the associated 95% confidence intervals. Specifically, the program allows for testing any linear contrast of means. Again, this approach is preferred because it does not depend on significance tests that might suffer from low power and also conveys the magnitude of the difference in addition to its precision. This module is accessed via the [*d* − Contrasts] tab. On this sheet, the only required input is the estimated mean, associated estimate variance (both of which the program can calculate), and the contrast weight (*c*) for the comparison groups (e.g., "−1" for men and "+1" for women). Thus, after estimating the confidence interval for a given subgroup (see above), one can then input the information for each comparison group. Of critical interest is the estimated difference in population mean differences and its 95% confidence interval (see Bonett, 2009):

$$\sum_{i=1}^{m} c_i\hat{\delta}_i \pm z_{\alpha/2}\left\{\text{var}\left(\sum_{i=1}^{m} c_i\hat{\delta}_i\right)\right\}^{1/2}. \quad (12)$$

Reliability Generalizator (Alpha)

An important feature of Synthesizer 1.0 is reliability generalization, a module accessible via the [Alpha] tab. This module, in addition to basic study information, requires input of sample size (*n*), the internal consistency estimate (i.e., Cronbach's alpha), and the number of test items (*q*). Following Bonett (2010), the meta-analytic estimate of the reliability is computed in the same manner as the one described for correlations:

$$\bar{\rho} = m^{-1}\sum_{j=1}^{m}\hat{\rho}_j. \quad (13)$$

The variance of a given reliability estimate is approximated by

$$\text{var}\left(\hat{\rho}_j\right) \approx 2q\left(1-\hat{\rho}_j\right)^2 / \left\{(q-1)(n-2)\right\}. \quad (14)$$

Logarithmic transformations are then employed to stabilize the variance of relevant parameters (see Bonett, 2010), yielding the following 95% confidence interval regarding the estimated population reliability:

$$1 - \exp\left(\ln(1-\bar{\rho}) - b \pm z_{\alpha/2}\left[\text{var}\left\{\ln(1-\bar{\rho})\right\}\right]^{1/2}\right). \quad (15)$$

In this expression, $b = \ln\{\bar{n}/(\bar{n}-1)\}$ is an approximate bias adjustment, with $\bar{n}$ as the harmonic mean sample size for the *m* studies.

Given that so much of psychology relies on scales and measures that must be reliable, this feature should be important for achieving a more firm sense regarding reliability of novel or seldomly used scales. This feature is especially useful because reliability estimates from small samples can be very imprecise.

Testing moderation. An even more important issue might involve questions about how reliability of a given measure varies across different populations or administration settings. One appealing feature of Synthesizer 1.0 is the ability to test differences in reliability across different groups of studies. This function is accessed via

the [r − Contrasts] tab (the same module used for testing differences in correlations). The only required input is the estimated mean reliability and respective upper- and lower-bound values from the confidence interval that was estimated for a given group of studies per the procedures described earlier. Of critical interest is the estimated difference in reliabilities and its 95% confidence interval (which are computed in the same manner as the ones for correlations; see above). I hope that this feature will motivate researchers to perform more systematic reliability generalization of their measures.

AN ILLUSTRATIVE EXAMPLE

In order to illustrate the application of the software, an example based on correlations is presented below. The presented analysis examines the association between grandiose narcissism (as measured by the Narcissistic Personality Inventory; Raskin & Hall, 1979) and self-esteem (across several unpublished data sets; Krizan, 2010). Furthermore, the analysis examines whether the association between narcissism and self-esteem depends on whether the latter is measured as a state variable (i.e., current, momentary self-esteem) or a trait variable (chronic self-worth, as in Rosenberg, 1965). Six of these correlation values and their sample sizes were entered in the Correlation Aggregator module of Synthesizer 1.0 (see Figure 1). To the right, under Moderator Categories, it has been indicated whether the self-esteem measure reflects a state or a trait.

Variance estimates for each sample mean are displayed to the right of correlations. The untransformed mean (.310) and its variance (.001) are shown toward the bottom of column G. Further below, one can see transformed estimates of the mean, variance, and associated upper- and lower-bound values of the confidence interval. These values are then back-transformed in order to be interpreted—shown to the right, under the heading "95% Confidence Interval Values." It is easily seen that the overall correlation between narcissism and self-esteem is estimated to be a moderate one, hovering around .31, with .25 and .37 as the lower and upper boundaries of the population value.

The other question of interest was whether the link between narcissism and self-esteem depends on whether the latter is measured as a state or a trait variable. To examine this question, population estimates for the link between narcissism and self-esteem were computed *separately* for studies that measured the latter as a trait rather than a state (according to the same procedure illustrated above). This yielded .30 [.21, .39] and .32 [.24, .39] as 95% confidence intervals for correlations with state and trait self-esteem, respectively. To estimate the magnitude of this difference, confidence interval values for each group were entered in the Correlational Contrasts module (see Figure 2). As can be seen, the estimated population difference is minuscule (−.02) and, more importantly, could go in either direction (given the boundaries of −.13 and .10). Given that the interval includes 0, we can state that the difference between correlations is nonsignificant. In sum, narcissism predicts both trait and state self-esteem in roughly equal magnitude, although the latter measures tend to be less stable and may have reduced the correlations.

CONCLUSIONS

The purpose of the present article was to introduce a new meta-analytic software tool that is free to researchers and uses a varying-coefficient model for confidence interval estimation in quantitative syntheses proposed by Bonett (2008, 2009, 2010). Given the limited applicability of traditional fixed-effect and random-effects models (cf. Hedges & Vevea, 1998), this tool is recommended for routine use. Although meta-analysis is primarily thought of as a tool for literature reviews, quantitative syntheses of research can also be used to combine the results from multiple studies in a single research report. In today's publishing environment, where multiple studies per report are the rule rather than the exception, most subfields of psychology would benefit from aggregating the research results across studies within a given report. This will lead to the most accurate estimation of parameters, and should promote cumulative progress. Moreover, one can perform focused tests comparing the results of a newly conducted study with studies already in the literature (see Bonett, 2009). In today's environment, where mathematical models for meta-analysis can often approach "bewildering complexity" (Shapiro, 1994), the computations used in Synthesizer 1.0 are straightforward and relatively simple. Remember that Pearson (1904) reported an unweighted average correlation between smallpox inoculation and mortality rates, which is exactly what is proposed by Bonett (2008).

Although refining statistical models for meta-analyses is important, the conclusions based on meta-analytic methods ultimately depend on the quality and representativeness of relevant research studies. As others have noted (e.g., Cooper & Hedges, 1994), meta-analytic results may be heavily constrained by characteristics of employed samples, measures, or experimental procedures. In this vein, researchers should remember that the external validity of meta-analytic conclusions does not depend only on the appropriateness of statistical methods used. Rather, meta-analytic findings should always be understood in the light of the quality of available evidence. Such awareness is critical both for accurate characterization of meta-analytic results and for planning further research.

AUTHOR NOTE

Thanks to Doug Bonett and Jeff Miller for their helpful comments regarding the manuscript. All correspondence should be addressed to Z. Krizan, Department of Psychology, W112 Lagomarcino Hall, Iowa State University, Ames, IA 50011 (e-mail: zkrizan@iastate.edu).

REFERENCES

BONETT, D. G. (2008). Meta-analytic interval estimation for Pearson correlations. *Psychological Methods*, **13**, 173-189.

BONETT, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*, **14**, 225-238.

BONETT, D. G. (2010). *Varying coefficient meta-analytic methods for alpha reliability*. Manuscript submitted for publication.

BONETT, D. G., & WRIGHT, T. A. (2007). Comments and recommenda-

tions regarding the hypothesis testing controversy. *Journal of Organizational Behavior*, **28**, 647-659.
BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. T., & ROTHSTEIN, H. R. (2009). *Introduction to meta-analysis*. New York: Wiley.
BORENSTEIN, M., & ROTHSTEIN, H. (2006). Comprehensive Meta-Analysis: A computer program for research synthesis (Version 2). Englewood, NJ: Biostat.
COCHRAN, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, **10**, 101-129.
COOPER, H., & HEDGES, L. V. (EDS.) (1994). *Handbook of research synthesis*. New York: Russell Sage.
COOPER, H., HEDGES, L. V., & VALENTINE, J. C. (EDSS.) (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage.
FIELD, A. P. (2001). Meta-analysis of correlation coefficients: Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, **6**, 161-180.
FISHER, R. A. (1921). On the "probable error" of the coefficient of correlation deduced from a small sample. *Metron*, **1**, 1-32.
GLASS, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, **5**, 3-8.
HEDGES, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**, 107-128.
HEDGES, L. V., & OLKIN, I. (1985). *Statistical methods for meta-analysis*. San Diego: Academic Press.
HEDGES, L. V., & VEVEA, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, **3**, 486-504.
HIGGINS, J., THOMPSON, S. G., DEEKS, J. J., & ALTMAN, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, **327**, 557-560.
HUNTER, J. E., & SCHMIDT, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative knowledge in psychology. *International Journal of Selection & Assessment*, **8**, 275-292.
HUNTER, J. E., & SCHMIDT, F. L. (2004). *Methods of meta-analysis: Correcting errors and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.
JUDGE, G. G., GRIFFITHS, W. E., HILL, R. C., LUTKEPOHL, H., & LEE, T.-C. (1985). *The theory and practice of econometrics* (2nd ed.). New York: Wiley.
KRIZAN, Z. (2009). *Synthesizer 1.0: User's guide*. Ames, IA: Iowa State University.
KRIZAN, Z. (2010). [Narcissism and self-evaluation]. Unpublished raw data.
MILLER, R. G. (1986). *Beyond ANOVA: Basics of applied statistics*. New York: Wiley.
NATIONAL RESEARCH COUNCIL (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academies Press.
PEARSON, K. (1904). Report on certain enteric fever inoculation statistics. *Behavioral Medicine Journal*, **5**, 1243-1246.
RASKIN, R. N., & HALL, C. S. (1979). A Narcissistic Personality Inventory. *Psychological Reports*, **45**, 590.
RODRIGUEZ, M. C., & MAEDA, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, **11**, 306-322.
ROSENBERG, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
ROSENTHAL, R. (1976). *Experimenter effects in behavioral research*. New York: Wiley.
SCHULZE, R. (2004). *Meta-analysis: A comparison of approaches*. Toronto: Hogrefe & Huber.
SHADISH, W. R., & HADDOCK, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 261-281). New York: Russell Sage.
SHAPIRO, S. (1994). Point/counterpoint: Meta-analysis of observational studies. Meta-analysis/shmeta-analysis. *American Journal of Epidemiology*, **140**, 771-778.
SHUSTER, J. J. (2010). Empirical vs. natural weighting in random effects meta-analysis. *Statistics in Medicine*, **29**, 1259-1265.
SPEARMAN, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, **3**, 271-295.

SUPPLEMENTAL MATERIALS

The Synthesizer 1.0 tool and user manual may be downloaded from http://brm.psychonomic-journals.org/content/supplemental.

(Manuscript received September 9, 2009;
revision accepted for publication March 20, 2010.)