Running Head: RELIABILITY GENERALIZATION

# The Replicability of Reliability Generalization

Michael C. Rodriguez and Taeho Jung

University of Minnesota

Paper presented at the AERA annual meeting, April 2003, Chicago, IL.

Preparing for publication consideration.

# The Replicability of Reliability Generalization

Meta-analysis of reliability coefficients has recently increased dramatically under the label reliability generalization (RG), although the quantitative synthesis of reliability coefficients has been reported in the literature since at least the 1970s. The RG term was coined by Vacha-Haase (1998), who synthesized reliability coefficients obtained from published literature on the Bem Sex Role Inventory. Since then, *Educational and Psychological Measurement* (EPM) published a special issue on RG in April 2000 and a second RG special issue in August 2002. In addition, the book *Score Reliability: Contemporary Thinking on Reliability Issues* (Thomson, 2002) was released through Sage Publications, which contains reprints from EPM, as well as several new manuscripts.

Little methodological work has been conducted on the RG practices employed to date. The purpose of this paper is to present the findings of an attempt to replicate three RG studies and to recompute the syntheses employing a statistically and theoretically grounded approach. The three studies replicated here include:

1. Capraro, Capraro, & Henson’s (2001) synthesis of reliability coefficients from the Mathematics Anxiety Rating Scale;
2. Caruso’s (2000) synthesis of reliability coefficients from the NEO personality scales; and
3. Henson, Kogan, & Vacha-Haase’s (2001) synthesis of reliability coefficients from the Teacher Efficacy Scale and related instruments.

The attempt to replicate these three studies uncovered several important obstacles to the completion of successful RG studies. These obstacles appear to have been overlooked or ignored by the RG authors. Our purpose is to bring basic principles of meta-analysis to the forefront of this work. We hope to present the findings in an amicable way, being instructive by providing solutions to the obstacles that were uncovered.

### *Theoretical Framework*

The framework employed in these replications was based directly upon the purposes of RG studies as set out by Vacha-Haase (1998): to characterize “(a) the typical reliability of scores for a given test across studies, (b) the amount of variability in reliability coefficients for given measures, and (c) the sources of variability in reliability coefficients across studies” (p. 6). In addition, RG authors stress the idea that reliability is not a function of the test (or measurement instrument), per se, but of the set of scores that results from a particular administration of a test. Thus, reliability coefficients are sample specific and should be reported based on the data resulting from a given study. We certainly agreed and paid particular attention to the presence of reliability coefficients reported in primary studies that were in fact clearly based on the data resulting from the study in which it was reported.

Each RG study research team quantitatively synthesized the reliability coefficients reported in their respective primary studies. In addition, a number of study-level characteristics that were potential moderators of anticipated variation in reliability coefficients were coded for inclusion in the analyses. In each case, the moderators were employed to explain variation in reliability characteristics under a general linear model approach.

The current replications rely on the meta-analysis framework of Hedges (1992, 1994) who employed a test of homogeneity of effects and a weighting scheme to minimize the variance of the estimates based on the sampling variance (variance of the sampling distribution of the effect size). We also rely heavily on Hakstian and Whalen (1976) who provided the statistical testing framework for combining *k* coefficient alphas based on the sampling distribution derivations of Kristof (1963) and Feldt (1965).

Our primary emphasis here is to produce replications of the three RG studies. At the same time, we argue that each RG study should have been grounded in the principles of study effect synthesis and sampling distribution theory. When possible, we replicated the procedures employed by the RG authors for each of the three studies reviewed here and then reanalyzed each data set for comparison purposes. We expect to demonstrate the efficiency in statistical terms of the sampling distribution approach to synthesizing coefficient alpha.

## Methods

The three RG studies included in this replication were selected to be representative of the initial surge of RG studies in 2000 and 2001 and that appeared to present a variety of instruments (personality assessment, teacher efficacy, and mathematics anxiety). The articles listed in the reference section of each RG study were retrieved. From the reference list provided by Capraro, Capraro, and Henson (2001), 16 studies were listed and retrieved (although the authors reported to find 17 in the text). Caruso (2000) included 37 synthesized articles in his reference list as he reported in the text, all of which were retrieved. Finally, Henson, Kogan, and Vacha-Haase (2001) included 49 articles in their reference list of synthesized work, although reported to have located 52 in the text. Those listed in the reference list were retrieved.

The articles were coded, based on the coding schemes presented by the RG authors. The second author coded all studies while two other researchers independently coded one half of each of the articles employed in each RG study. Numerous disagreements in coding resulted—all disagreements were documented and reviewed by all three researchers to achieve consensus. Six months later, the first author reviewed each article when final coding decisions were made. Few additional changes were made at that time.

The sources of disagreements provided opportunities to uncover obstacles in completing RG studies. These disagreements provided additional bases for our analysis. We meticulously recorded and considered all results from these activities.

The same procedure was employed for each RG study. We first accounted for the number of studies and reliability coefficients obtained and compared these results to those reported by the RG authors. We then compared summary statistics given our findings, and attempted, when possible, to replicate the statistical analyses as employed by the RG authors. As described below, we disagreed with the procedures employed in all three RG studies.

We followed the generalized notation employed in the *Handbook of Research Synthesis* (Cooper & Hedges, Eds., 1994), where each of the *k* study effects are noted as $T_i$, with a weighted mean $\overline{T}_\bullet$. In our case, the study effect is the transformed sample coefficient alpha: $T_i = (1 - r_{\alpha i})^{1/3}$, based on the sampling distribution work of Hakstian and Whalen (1976). The weighted mean study effect is $\overline{T}_\bullet = \frac{\sum w_i T_i}{\sum w_i}$, where the weights are the reciprocal of the variance of each study effect (due to sampling subjects within studies) plus the random-effects variance component (due to sampling studies from the universe of studies), $w_i = \frac{1}{v_i + \sigma^2}$. By weighting in this way, we obtain the mean point estimate with the appropriate sampling variance for generalizations beyond the studies included in the meta-analysis, a random-effects model. The variance of study effect *i* is $v_i = \frac{18 J_i (n_i - 1)(1 - r_{\alpha i})^{2/3}}{(J_i - 1)(9n_i - 11)^2}$ (as derived by Hakstian & Whalen, 1976) given $r_\alpha$ (alpha), *J* items, and a sample size of *n*. The variance of the mean is $v_\bullet = \frac{1}{\sum w_i}$ with a standard error of the mean $\sqrt{v_\bullet}$ .

In the event that test-retest coefficients were synthesized, these coefficients were transformed using Fisher's normalizing and variance stabilizing *Z*-transformation. As noted by Rosenthal (1994), for any correlation *r*, $Z_r = \frac{1}{2}\log_e\left[\frac{1+r}{1-r}\right]$. The variance of $Z_i$ is $v_i = \frac{1}{(n_i - 3)}$ where *n* is the within-study sample size. Similarly, a random-effects variance component would be added to the conditional variance for computing the appropriate weights, as above.

When necessary, a maximum-likelihood estimate of the random-effects variance component $\left(\sigma^2\right)$ was computed with HLM 5.0 (Raudenbush, Bryk, & Congdon, 2000) and added to the conditional variance estimate for computing effect weights. Under fixed-effects models, this random-effects variance component is set to zero.

There is one remaining methodological issue that is not easily addressed. It regards the functional relationship between the magnitude of coefficient alpha and score variability. Caruso (2000) noted that reliability coefficients, like correlations, are dependent on variance, and chose to correct each coefficient for restriction of range, using standardization data from NEO manuals for the reference variance. We relied on Lord and Novick (1968), who presented the precise nature of the impact of group heterogeneity on reliability in the following statement of equivalence of the resulting reliability $\left(\rho^2_{\tilde{X}\tilde{T}}\right)$ given the original reliability $\left(\rho^2_{XT}\right)$, the original observed score variance $\left(\sigma^2_X\right)$, and the more or less heterogeneous group observed score variance $\left(\sigma^2_{\tilde{X}}\right)$, which Caruso referred to as the reference variance: $\rho^2_{\tilde{X}\tilde{T}} = 1 - \frac{\sigma^2_X}{\sigma^2_{\tilde{X}}}\left(1 - \rho^2_{XT}\right)$.

As long as there is constant error variance, a reduction in observed variance will reduce score reliability. The assumption of constant error variance is not tenable if the standard errors of measurement vary as a function of true-score level *and* the distribution of true scores is different in the two groups (Allen & Yen, 1979). It is not clear whether the appropriate procedure is to correct reliability estimates for group heterogeneity directly, given the required assumptions, or to include group variance as a covariate for any analyses. We completed this adjustment for the replication of Caruso (2000).

We used these methods to conduct the reliability generalization when possible and reported differences resulting from the use of these methods versus those used by the RG authors. We now present a summary of each RG study and the results of our replication.

## Three RG Studies & Replications

### *Capraro, Capraro, and Henson RG Summary*

Capraro, Capraro, and Henson (2001) synthesized reliability coefficients from studies employing the Mathematics Anxiety Rating Scale (MARS, Richardson & Suinn, 1972), an instrument that assesses anxiety about one’s performance in mathematics. The original form of the MARS contained 98 items, consisting of short statements describing real-world and academic situations thought to provoke mathematics anxiety. Responses were based on a 5-point scale. The MARS has subsequently been revised to include a form for children in elementary school and also for adolescents. There have also been revisions to develop short forms of the test with one-fourth the number of items.

There were two goals for the meta-analysis: (a) to characterize the typical score reliability and (b) to investigate study characteristics that may explain variation in coefficients; two goals common to RG studies. The authors located 67 unique articles that used the MARS. Of those, 17

(25%) reported reliability information. Several studies contained more than one reliability coefficient (typically for subsamples). The coding scheme for primary studies included study characteristics, such as number of items used, number of points on the rating scale, sample size, a nominal age variable, type of reliability coefficient computed, and the standard deviation of scores.

No adjustments or transformations were made to the coefficients. The 28 alpha coefficients yielded a mean of .915 (*SD* = .08); the seven test-retest coefficients yielded a mean of .841 (*SD* = .07). They also reported a total mean for all 35 coefficients which was not particularly meaningful since each reliability type captures a different primary source of measurement error and are not comparable.

Four regressions were completed to assess the explanatory power of the study level characteristics, including 10 variables. The four models differed by their inclusion of reliability coefficient type and score standard deviation; all other 8 variables were included in each. The associated $R^2$s including only alpha coefficients were .33 without score *SD* (*n*=28) and .58 with *SD* (*n*=15) and for alpha and test-retest combined were .40 without *SD* (*n*=35) and .64 with *SD* (*n*=18). The primary problem with comparing these models is the reduction in sample size due to the availability of information in the primary studies. In addition, there are sample size issues regarding the available degrees of freedom to evaluate 10 variables, particularly among the two smaller models.

Capraro et al. summarized their results by indicating that several study characteristics were related to score reliability variability. They suggested that adult samples (approximately 11% of the samples) yielded less reliable scores. They also reported that test length was positively related to reliability coefficients except in model 3 where test-retest coefficients from the longest tests were included and the three test-retest coefficients were lower than most. This point is a clear indication of the instability of the estimates, largely due to sample size and nonnormally distributed variables. Since reliability type was included in the model, it accounted for the smaller coefficients; the confounding of reliability type and test length confused their interpretation.

Finally, as expected, score standard deviation was an important explanatory variable, increasing the amount of variance explained by about 25% in both models. “This finding highlights the potential impact of total score variance on reliability estimates” (pp. 383-384). “In sum, measurement error in MARS scores appears to increase in adult samples and perhaps in other homogenous age groups” (p. 384).

The authors concluded by noting the usefulness of computing reliability estimates for the data reported in a given study, rather than continue the practice of reporting coefficients from elsewhere, whether from the test manual or other studies.

### *Capraro, Capraro, & Henson Replication Results*

Although the authors reported to find 17 primary studies, only 16 were denoted in the reference list. From the 16 recovered primary studies, we located 15 unique articles. Rounds and Hendel (1980) was based on their AERA paper presented in 1979, both of which were included in the Capraro et al. RG study. From the 15 studies, 23 reliability coefficients were obtained, which did not compare well with the 35 located by the RG authors. Four primary studies included references to reliabilities reported elsewhere, but did not include reliabilities for their own samples, including Rounds and Hendel (1979, 1980), Suinn (1989), and Wilson (1997). Suinn (1989) actually reported results from the same study in two articles, one including the full sample (1988) and one including only Hispanic children (1989); however, they reported the reliability for the full sample in both studies. D'Ailly (1992) reported score reliabilities for two subscales, but the RG study did not include subscale analysis. It is not clear how these coefficients were handled—we believe they should have been excluded, which reduced the number of coefficients to 21, of which 7 were test-retest coefficients and 14 were alpha coefficients.

Alexander (1998) included two coefficients for the same sample, test-retest and alpha. We chose to include only the test-retest coefficient and excluded alpha from the replication. Richardson and Suinn (1972) also reported both types of coefficients, the test-retest coefficient for a subsample (*n* = 35) of a larger sample (*n* = 397). Although we believe only one should have been included, we decided to include both since the samples were possibly significantly different (one was only 9% of the other and we already had a dramatically smaller number of coefficients than that in the original RG study).

The first notable difference from the original RG summary statistics was the difference in sample size. However, we reported the same number of test-retest coefficients with identical results (*M*=.841, *SD*=.07). For the results of the alpha coefficients, our mean was higher (.946 compared to .915) with significantly lower variation (*SD* of .037 compared to .083). The lower variation can also be seen in the limited range, where our minimum value was .88 compared to .55 in the RG study. We were not sure about the location of the .55 coefficient. Perhaps it was obtained from Wilson (1997) who reported a coefficient alpha of .55 for a test called the Alleviating Statistics Anxiety Assessment in their study including the MARS, but did not result from the MARS.

The regression analyses no longer seemed advisable, with only 21 observations and 10 explanatory variables. Even worse, three of the four models completed by the RG authors would now have fewer than 15 observations. It would be difficult to compare $R^2$ for these models because the number of variables was close to the number of observations, essentially providing enough information to uniquely identify each case, thus yielding spuriously high $R^2$s.

Several problems existed with the regressions completed by Capraro et al. The primary issue concerned multicolinearity and the confounding of variables included in the analyses. The authors noted that the number of items used in MARS varied a great deal across studies. As we reviewed this variation, we found that in fact, the number of items employed by primary authors ranged from 98 (as initially designed) to 22 (a shortened form for children). Upon further review, we noted that the number of items was directly correlated with the score standard deviation ($r$ = .98), as expected (since the standard deviation is a direct product of the number of items or points on the scale). In addition, the forms used with college students were the longer forms, 80 items or more; the forms used with children included the shortest forms, 26 items or fewer. The shortest forms with 22 items were also those forms that included a 4-point scale (Chiu, 1990), where all others included the standard 5-point scale. Of the 13 cases with standard deviation information and both coefficient alpha and test-retest reliability, the two test-retest reliability correlations were from college age students employing the 98 item form with the lowest standard deviations and yielded the lowest reliability coefficients. All of these variables were included simultaneously, although highly covarying and in some cases, completely confounded. The data for the 21 obtained coefficients are displayed in Table 1 in order by type of coefficient and number of items employed in the scale.

We reviewed the Spearman correlations among the coded variables. Spearman correlations were employed because of non-normal distributions among most variables; we typically see non-normal distributions for reliability coefficients, since they typically tend to be high, with a ceiling of 1.0 and few lower values. We also included an equated *SD* by dividing the standard deviation by the number of items and making a further adjustment for the forms with 4-point scales rather than 5-point scales. The rank-order correlation between reliability and the standard deviation was .24, but with the equated *SD* was .80, much larger as it should be. At the same time, the correlation between *SD* and the number of items was .95, while the correlation with the equated *SD* was -.06.

Unfortunately, we were not able to conduct a replication of the regression analyses completed by Capraro et al. We did complete a synthesis of alpha coefficients based on the *T*-transformation described earlier. Through the transformation and synthesis of the 14 alpha coefficients, we obtained a mean value of .978 with a standard error of .002. For 11 of the coefficients we were able to make an adjustment for group heterogeneity and number of items in the scale (Spearman-Brown adjustment). We employed equated *SD*s (per-item *SD*s adjusted for number of scale points employed) in the adjustment for group heterogeneity, with the largest *SD* from the 98-item scale (.67 was the reference *SD*) where the *SD*s ranged from .43 to .86. The mean transformed coefficient alpha was .981 with a *SE* of .002. Table 9 contains the results of these analyses.

Finally, we examined the heterogeneity of each set of coefficients given the model employed in Table 2. The complete set of 14 coefficient alphas had a heterogeneity value of $Q$ = 1412, distributed as $\chi^2$ with 14-1 degrees of freedom (highly heterogeneous). The 11 coefficients with variance information had a heterogeneity value of $Q$ = 1305, a similar result. The 11 coefficients that were adjusted for number of items in the scale and the equated *SD* had a heterogeneity value of $Q$ = 242, a significant reduction in heterogeneity (81% explained variance); however, a significant amount of heterogeneity remained. This change in variation can be seen in Figures 1 and 2. Random-effects analysis explained the heterogeneity by encompassing the additional uncertainty due to sampling studies from the hypothetical population of studies that could have been published with reliability information (for coefficient alpha, *M*=.98, *SE*=.011, *Q*=10, n.s.).

### *Caruso RG Summary*

Caruso's (2000) meta-analysis of reliability coefficients involved the NEO Personality Inventory (Costa & McCrae, 1985), an instrument which assesses the domains of the five-factor personality model (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness). Three forms of the instrument exist, including the original NEO-PI (a 180 item instrument assessing 5 domains and 6 facets in the first three domains), a revised version (NEO-PI-R, a 240 item instrument with 48 items and 6 facets per domain), and a short form named the NEO Five-Factor Inventory (NEO-FFI, a 60 item instrument with 12 items per domain).

Consistent with the RG literature, Caruso presented two primary purposes for the meta-analysis: (a) characterize the typical reliability of scores in terms of its distribution, and (b) identify sample characteristics that may explain variation in reliability across studies. To accomplish these purposes, Caruso identified 244 empirical studies (after preliminary screening for relevance), of which 37 apparently reported reliabilities for each domain. He also noted problems with the other studies, where 57% did not report reliability coefficients and 19% reported reliability coefficients from the NEO manuals or other studies. This left him with 15% of the available empirical literature. In addition, several primary studies reported multiple reliability coefficients from multiple samples (e.g., by gender), which provided 51 separate reliability coefficients.

Several study-level characteristics were coded and employed in the analysis to explain variation in reliability coefficients, including the referent (self or other); the instrument version; language of administration; whether the sample was of students, general population, clinical, or some combination; gender composition; type of reliability coefficient (test-retest, alpha); sample size; and mean age. Reliability coefficients were transformed to Fisher's *Z* prior to statistical analysis.

Caruso also noted the relationship between reliability and score variation, suggesting that reliability coefficients are essentially correlations, and chose to correct coefficients for range restriction using sample variance estimates (when available) and reference variances from standardization samples as reported by Costa and McCrae (1985, 1992). No indication was given regarding the adjustment to coefficients for which no sample variance was available. In the manuals providing the reference variances, there are variances available for each form, by gender, and for the self-referent forms and other-referent forms. In addition, there are college sample estimates available for the self-referent form.  It is unclear as to whether or not the various estimates of variance were used with respect to the form and sample in a particular study. Complicating these adjustments is the fact that many primary authors used modified forms, deleting items for various reasons. Scales with fewer items yield scores with smaller variances. An adjustment based on a standard variance for a given number of items would be inappropriate for scales with different numbers of items. However, adjustments due to different numbers of items were not made to the reliability coefficient. Caruso did not address this problem.

Caruso reported descriptive statistics for each of the NEO domains, based on 51 coefficients obtained from 37 studies (Table 6, first column). He noted that the Neuroticism scale yielded the scores with the highest reliability (*M*=.88, *SD*=.07) and lowest variability (which is likely partly a function of a ceiling effect). He then reported bivariate correlations between each domain reliability coefficient, sample size, and mean age. The $r^2$ values ranged between .00 and .06.

Finally, Caruso completed seven ANOVAs to evaluate differences among the nominal study characteristics and the reliabilities for each of the five domains. Of the 35 completed ANOVAs, Caruso reported 12 differences significant at $p < .01$ and three significant at the .05 level. One study characteristic, version of NEO used, yielded consistent significant differences ($p < .001$) for all five domains. This was likely a function of the number of items in each scale, which led Caruso to correct coefficients using the Spearman-Brown correction formula. This correction resulted in significant differences in NEO version for four of the five domains. Caruso also noted surprise that the type of coefficient (test-retest versus alpha) was not significant in all cases (only for the Agreeableness domain); however, there were only four test-retest coefficients, so although the test-retest coefficients were lower in all cases, the differences were not statistically significant.

Caruso noted limitations. First, he suggested that the data points were not entirely independent since several studies reported multiple reliability coefficients. However, the limitation of independence was also manifest in the testing of five domains that are correlated as though they too were independent; Costa, McCrae, and Dye (1991) reported intercorrelations as high as –.49 for N and C and .43 for O and E. Caruso reported that no weighting was done, suggesting that a secondary analysis produced similar results. He also addressed concerns about the “file drawer” problem, the situation where less than significant results fail to appear in the literature and that a large proportion (85%) of the authors of NEO empirical studies failed to report reliability.

Finally, Caruso provided a section in his discussion of results which was titled "Situations Resulting in Inadequate Score Reliability," a statement with causal connotations. He noted that among the five domains, Agreeableness scores had low score reliability when the referent was self (the person responding about themselves, 45 of 51 coefficients), when administered in English (41 coefficients), when administered to students (17) or clinical subjects (7), and when based on the NEO-FFI version (20).

*Caruso Replication Results*

All of the 37 recovered primary studies we located were unique studies. From the 37 studies, 53 reliability coefficients were obtained, which is more than the 51 coefficients reported by Caruso. At the same time, one article reported in the RG study (Ball, Tennen, Poling, Kranzler, & Rounsaville, 1997) did not report a reliability coefficient for the study data, but cited the reliability from Costa and McCrae's (1992) original work. The 53 coefficients included 4 test-retest correlations (matching Caruso's count) and 49 alpha coefficients (two more than reported by Caruso).

We are not sure how the correction for range restriction was accomplished for 23 of the reliability coefficients for which no score variance was reported. We believe that these cases were not corrected, since Caruso reported that the sample reliabilities were "equalized with respect to their variability where possible" (p. 241). In addition, variance information was reported under various conditions. For several studies, standard deviations were reported for subgroups where reliabilities were computed for the entire group (to solve this problem we computed total sums of squares from subgroup means and standard deviations, to obtain total group variance). Standard deviations were occasionally reported in *T*-scale metric, which is a standardized scale with mean of 50 and a standard deviation of 10. The *T*-score standard deviations are not comparable to raw score standard deviations, which were also reported, nor are they adjustable in the same way. In some cases, reported standard deviations were based on item scores while others were based on total scores (we converted item-level score standard deviations to total-score values by multiplying the item-level values by the number of items in the subscale). Finally, in some cases, the standard deviation was based on a different sample size than the reliability coefficients. In these cases, we used the sample size from the reliability coefficient as the relevant sample size for the study.

Several issues arose in the computation of summary statistics and subsequent interpretations. At the time of the preparation of this paper, the NEO-PI technical manual was not available, so the *SD* employed to adjust alpha for the NEO-PI was the median *SD* as reported by the primary study authors. The NEO-PI-R and NEO-FFI reliability adjustments were made based on *SD*s retrieved from the technical manual. Many of the reliability coefficients reported by primary authors were reduced in magnitude when adjusted for group heterogeneity because the standardization standard deviation was smaller than the standard deviation from the sample employed in primary study. This created a question about the appropriateness of interpreting the “typical” or mean reliability coefficient. Perhaps the standardization sample was unnecessarily restricted in range and samples from primary studies were more representative in their heterogeneity. In all, of the 10 NEO-PI coefficients that were adjusted, half were reduced because of the adjustment (because the median *SD* value was employed for the adjustment); of the 9 NEO-PI-R adjusted coefficients, 4 to 8 were reduced (given the subscale); of the 11 NEO-FFI adjusted coefficients, 9-11 were reduced.

Overall, the results were similar to those reported by Caruso (Table 6, second column). Nearly all mean and median reliabilities were within .02, with one difference of .04, with not much greater departure when examining only those alpha coefficients with score variability information. We caution readers of Caruso and this replication regarding the interpretation of these summary statistics, because of the choice of adjustment *SD* – which as described above, led to a reduction in most reliability coefficients. However, when we conducted the *T*-transformed weighted meta-analysis, we found much higher estimates of mean reliabilities for each domain (Table 7), ranging from .04 to .15 higher.

Table 7 also contains results from alpha coefficients adjusted for the number of items in the scale, employing the Spearman-Brown prophecy formula. This formula provided an adjustment for coefficient alpha under the assumption that items were either randomly eliminated or items equivalent in quality to the existing items were added to the scale. Since there is a functional relationship between the number of items in a scale and the magnitude of reliability, this was an appropriate adjustment to make comparisons more equivalent. Not only were the three versions of the NEO constructed with different numbers of items, but three primary authors modified the forms to include a different number of items from the original forms.

Caruso then conducted ANOVAs to assess variability based on several study-level characteristics. It appears that the individual ANOVAs were not well considered. Tables 4 to 6 contain alpha coefficients from each primary study with codes for four of the study-level characteristics by NEO form. Based on review of coding by NEO form, the ANOVAs were influenced by the lack of variability, overlap, and confounding. For example, the NEO-PI-R was only conducted on mixed genders with self-referent forms. All clinical samples were tested with the NEO-PI.

*Henson, Kogan, and Vacha-Haase Original RG Study*

Henson, Kogan, and Vacha-Haase (2001) synthesized reliability coefficients from the Teacher Efficacy Scale and related instruments. The authors introduced their project by reviewing the role of teacher efficacy in recent research, as well as the various measures employed by researchers. Several instruments were included in this synthesis.

The Teacher Efficacy Scale (TES, Gibson & Dembo, 1984) appears in 30 item and 16 item forms, typically employing a 6-point agreement scale with statements about how a teacher can personally affect the schooling outcomes of children (personal teaching efficacy, PTE) and how teachers can generally impact learning given the external factors children bring to the classroom (general teaching efficacy, GTE).

The Science Teaching Efficacy Belief Instrument (STEBI, Riggs & Enochs, 1990) was created by modifying items from the TES to address the context of the elementary science classroom, with additional items written to create a larger pool. Through factor analysis, 25 items were chosen to represent personal science teaching efficacy beliefs (PSTE), and science teaching outcome expectancy (STOE), on a 5-point agreement scale. A preservice version has also been created with items modified to the preservice context, but was not differentiated from the initial inservice version in the Henson et al. analysis.

Finally, two locus of control instruments were included. The Locus of Control instrument (TLC, Rose & Medway, 1981) includes 28 forced-choice items, half of which cover student success outcomes (I+) and half cover student failure outcomes (I–). One point is assigned to each scale if the teacher makes an internal attribution in either case. The Responsibility for Student Achievement instrument (RSA, Guskey, 1980) includes 30 items, half regarding positive classroom situations and half regarding negative situations. Respondents assign some proportion of 100% to four probable causes, including teaching ability, teaching effort, task difficulty, and luck, resulting in mean weightings for self-responsibility (teaching ability and teaching effort) to positive events (R+) and negative events (R–).

The authors located 52 articles that contained sufficient information to be included in the meta-analysis, although only 49 were denoted in the reference list. The number of articles that were empirical studies involving one of the instruments out of over 600 articles initially identified was not provided. Of the 52 articles retained, the authors reported to locate 86 reliability coefficients (all internal consistency estimates), although their summary statistics reported an *n* of 94. These included coefficients for the PTE (25), GTE (21), PSTE (13), STOE (11), I+ (7), I– (7), RSA+ (5), and RSA– (5). In addition, 15 study level characteristics were coded, but analyses were only completed on eight, including teacher experience, teaching level (elementary, mixed), teaching area (regular education, other), gender composition of sample, sample size, number of items, mean score, and score standard deviation.

To complete the analyses, reliability estimates were computed by Henson et al. for the TLC subscales. Since the scores on the TLC are based on dichotomous items, they recorded score means and standard deviations and the number of items and computed KR-21. They also noted that total score variance is related to internal consistency reliability estimates, but did not employ a correction to the coefficients.

Henson et al. reported descriptive statistics for all subscales (Table 10, column one) and illustrated variation in reliability coefficients with boxplots. The smallest mean reliability was found with GTE (*M*=.70, *SD*=.07, *n*=21); the largest was found with PSTE (*M*=.89, *SD*=.05, *n*=13). The only other analyses completed included bivariate correlations between reliability coefficients for each subscale with the study-level characteristics.

Results of the correlational analyses were varied, depending on the subscale evaluated. The primary problem with interpreting the consistency of results is that only one variable, sample size, was reported with sufficient consistency to be correlated with each of the 8 subscales. The correlations with sample size ranged from -.50 (GTE, *n*=21) to .93 (STOE, *n*=11). Teacher experience was correlated with 6 subscales, ranging from .12 (PSTE, *n*=12) to .99 (I–, *n*=7). Gender composition also correlated at a wide range, from –.70 (PSTE, *n*=7) to .99 (RSA–, *n*=3). As can be seen, however, samples sizes are quite small.

The authors noted the correlations between score variance and alpha coefficients, which were computed on 5 of the 8 subscales. These correlations ranged from .68 (PSTE, *n*=5) to .99 (I–, *n*=6). The authors noted that all correlations between the number of items and test score reliability were positive, except for the correlation with PSTE (*n*=13). They suggested that shorter forms may yield more reliable scores. All else equal, this is not plausible. This is the primary limitation of comparing bivariate relationships only. We investigated such anomalous results more carefully in the replication.

### *Replication Results*

We obtained the 49 articles denoted in the reference list by the RG authors, although they stated within the text that they had 52 articles. Of the 49 recovered primary studies, we located 47 unique articles. Guskey (1988) previously reported the same study in a 1987 paper presented at the annual meeting of the AERA. Riggs and Enochs (1990) also previously reported their study in a 1989 paper presented at the annual meeting of the National Association for Research in Science Teaching. Although all four were included in the Henson et al. RG study, we chose to include the published articles only. At the same time, both of these primary studies presented additional problems. Guskey (1988) reported reliabilities for a pilot sample, with no additional information about the pilot sample. He reported full information about the sample employed in the full study, but did not compute reliabilities on that sample. Riggs and Enochs (1990) did not report score standard deviations in their published article, we obtained them from the paper presented earlier. In addition, the score standard deviations were estimated from a longer form than what was used to estimate the reliability coefficients. We adjusted the score standard deviation to accommodate a smaller form (in the spirit of trying to retain as much data as possible).

Of the 47 unique studies, 19 studies either did not report reliabilities at all or cited reliabilities from other sources, not the data in hand. We present these based on the instrument employed by the primary authors.

We located 8 unique studies employing the RSA (excluding the duplicated Guskey, 1987 paper). Guskey (1981a) and Pratt (1985) did not mention reliability at all. Benninga (1981) and Guskey (1981b, 1984, 1987) cited reliabilities reported by Guskey (1980), although the RG authors did not include the 1980 source. Mehan (1981) employed an instrument with a similar name, but of different origin: The Responsibility for Student Achievement Questionnaire (RSAQ; Stallings, Needels, Stayrook, 1979). Mehan did not report reliability information. This left one unique study employing the RSA, Guskey (1988). Henson et al. reported summary statistics for 5 RSA reliability coefficients, although they included 9 studies employing the RSA in the reference list.

We located 6 unique studies employing the STEBI (excluding the duplicated Riggs and Enochs, 1990 paper). Wenner (1995) cited reliabilities from Riggs and Enochs (1990). This left 5 unique studies employing both subscales of the STEBI with reliability coefficients reported for their samples. Henson et al. reported summary statistics for 13 STEBI reliability coefficients (13 for PSTE, 11 for STOE).

We located 5 unique studies employing the TLC. Greenwood (1990), Marcinkiewi (1994), and Parkay (1988) cited reliabilities reported by Rose and Medway (1981). Payne (1991) employed a different instrument with a similar name: the Locus of Control Scale for Teachers (Sadowski, 1982), which indicates the degree of internal versus external control, with no separate scores for positive or negative events. This left one unique study employing the TLC with reliabilities reported for the study sample by Rose and Medway (1981), the authors of the TLC. However, Greenwood (1990), Parkay (1988), and Rose and Medway (1981) provided score means and standard deviations, which were used by Henson et al. to compute KR-21 (appropriate for dichotomously scored items assuming item difficulties are equal). We noted that Rose and Medway (1981) also reported coefficient alpha estimates of .71 (I+) and .81 (I-) which were higher than the KR-21 estimates of .66 and .79 for the same subscales using reported means and standard deviations (demonstrating the fact that KR-21 underestimates reliability). Henson et al. reported summary statistics for 7 TLC reliability coefficients. We were able to compute three such coefficients.

Finally, we located 28 unique studies employing the TES. Tracz (1986) did not include any reliability information. Anderson (1988), Grafton (1987), Kim (1998), Landrum (1992), Poole (1989), and Paese (1991) all cited reliabilities from Gibson and Dembo (1984). Although most primary authors employed altered or modified forms, some authors used instruments that were not the same as the TES designed by Gibson and Dembo (1984). Hagen (1998) employed a different Teacher Efficacy Scale, designed by Emmer and Hickman (1991). Meijer (1988) used the Dutch Teacher Efficacy Scale, for which 6 of the 11 items were similar to those used in the

Gibson and Dembo short form (we retained this one). Muman (1995) used a different form for the GTE developed by Perlin (1981); information for this scale was not retained.

Other minor problems were uncovered, for which we unfortunately made compromises to retain as much data as possible. Coladarci (1992) reported standard deviations for a different sample size than that used to estimate reliability. Hoy (1990) reported standard deviations for student teachers and a control group with different sample sizes for the two subscales (PTE and GTE), with no indication whether the reliabilities were computed employing the full sample or student teachers only; we assumed the full sample was used. Podell (1993) reported standard deviation information for standardized scores, so that the standard deviation was 1; we did not use this standard deviation information. This left 20 unique studies employing the TES, resulting in 20 coefficients for PTE and 17 coefficients for GTE. The RG authors reported 25 coefficients (25 for PTE and 21 for GTE).

Summary statistics for the reliabilities included in this replication were compared to the RG results. The means and standard deviations were remarkably similar for the TES and STEBI (Table 10, column two). However, when compared to the results of a random-effects meta-analysis when coefficient alpha was adjusted for group heterogeneity and the number of items employed in a study, results were different (see Table 11). For the PTE, the original RG study yielded a 95% confidence interval of (.76, .80) while the random-effects replication yielded (.82, .87); for the GTE, the original RG yielded (.67, .73) while the random-effects replication yielded (.73, .82). These confidence intervals suggested slight but meaningful differences.

Henson et al. computed correlations on unadjusted reliability coefficients with study level variables. In the replication, there was only one unique study reporting reliability for the RSA. Similarly, only 3 coefficients were available for the TLC correlations and only 5 coefficients were available for the STEBI, which would have yielded highly unstable correlations. For the STEBI studies, teaching level and area were constant for the 5 studies, all but one study included inservice teachers; the number of items used on the PSTE scales was constant. These variables would not be useful to study variation in reliability.

The study characteristics for those studies including the TES were complex. Among these studies, all regular education teachers were also preservice teachers. There were only two inservice teacher studies, both of which were elementary education and non-regular education teacher samples, where both the PTE tests were 12 items and both GTE tests were 8 items. For the studies with mixed-levels of education teachers, all were preservice teachers. Because of this confounding, bivariate correlations would be misleading and not informative.

Aside from the inability to replicate the correlations from the original RG study, we were perplexed by the varying sample sizes available for various correlations. Particularly troubling was the sample size available for the correlations with score variance. Although the means of the score variance were similar between the original and replication study (.36 and .37 respectively for PTE and .53 and .49 respectively for GTE), the standard deviations of sample variances were quite different (.05 and .11 respectively for PTE and .12 and .17 respectively for GTE). We were not sure what the source of difference was in the number of variances reported by Henson et al. As we reviewed our own coding, we observed that seven studies reported score standard deviations for total scores, which we converted to item-level values. In one case, to obtain the total group variance, we computed sums of squares from subgroup summary data to compute total variance for which a total group reliability coefficient was reported (Soodak, 1994). Because of the dependence of reliability on score variability, it was important to capture and employ as many score variances associated with reliability coefficients as possible.

The interpretation of this analysis is further complicated because of the varied number of items primary authors employed in each scale. The mean number of items for the PTE was 10 (ranging from 6 to 14) while the mean number of items for the GTE was 7 (ranging from 4 to 10). The Gibson and Dembo (1984) scales included 9 and 7 items respectively. The point is that the functional relationship between the number of items and score reliability cannot be ignored. So, the correlations, although providing some information about the nature of relationships with score reliability, were problematic to interpret because multiple variables simultaneously impact reliability – even the correlation between score reliability and variability is not particularly useful when each case employed a different number of items.

## Discussion

We were not able to replicate the three RG studies. For the most part, many of the intermediate results were similar, including most counts and summary statistics of coded variables for each RG study. Results from the replications differed meaningfully when correcting reliability coefficients for group heterogeneity (range-restriction) and number of items, and employing the normalizing *T*-transformation in a weighted least-squares random-effects analysis. However, the primary differences were based in the number of usable reliability coefficients. Unfortunately, in all three RG studies, we located primary studies where the only reliability coefficient reported was that obtained from previous reports or data not included in the present study – a salient item on the agenda of the RG synthesist, referred to as reliability induction (Vacha-Hasse, Kogan, & Thompson, 2000). We also located primary studies that reported *no* reliability coefficient, either cited from elsewhere or for the data in hand. In addition, for two of the RG studies, we located primary studies that were included twice, once as a conference paper presentation and again as a published journal article.

There were other methodological issues that presented significant barriers to the completion of the replications. In particular, the reporting of score standard deviations was haphazard and often missing in primary studies. This was significant because of the functional relationship between the magnitude of a reliability coefficient and score variance (Feldt & Brennan, 1989). In many cases, standard deviations were not reported. In others, they were reported for subgroups while the reliability coefficient was reported for the whole sample. In addition, standard deviations were reported in either item-level scales or score-level scales. Each presented a unique problem to which we reported solutions.

Another significant problem was that some primary study authors made minor to substantial alterations to the original instrument. These alterations included the addition or subtraction of items and the modification of the response scale (e.g., from a 5-point scale to a 4-point or 6-point scale). These modifications certainly resulted in a functional impact on the reliability coefficient. For example, the relationship between the magnitude of the reliability coefficient and the number of items is well known (the basis for the Spearman-Brown prophecy formula). When attempting to put all standard deviations on the same scale, the number of score points on the scale clearly impacts the magnitude of the standard deviation.

We have learned a great deal about the practice of RG through replication of published RG studies. We have learned less about the generalizability of reliability coefficients because of the low level of methodological rigor employed to date. Several barriers to successful completion of an RG study were uncovered in this replication with suggested solutions, where available. These were provided as a guide to future RG researchers and to meta-analysts who attempt to synthesize psychometric studies.

Score reliability is an important test score characteristic because it affects every statistic that employs its corresponding scores. Reliability is an indicator of the degree to which a set of scores contains random error – and as we know, random error attenuates all resulting statistics (e.g., correlations, regression coefficients, t-tests, etc.). In clinical settings, reliability coefficients provide us with an indication of the precision of a given score (i.e., we use reliability to obtain the standard error of measurement, employed to interpret individual test scores). The role of reliability is clearly important. The more we understand the characteristics of score reliability, the better our ability to use it appropriately.

One might argue that generalizing from RG studies is a dangerous venture simply because so few primary authors report reliability information for the scores they employ. In our case, less than 15% of the authors of empirical studies employing the MARS and NEO reported relevant reliability information; for the TES and related scales, this may have been as little as 5%. Based on the results of this study, the generalizability of RG studies is further limited because of the inability to replicate results.

# Reference

Capraro, M. M., Capraro, R. M., & Henson, R. K. (2001). Measurement error of scores on the mathematics anxiety rating scale across studies. *Educational and Psychological Measurement, 61,* 373-386.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60*, 236-254.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

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

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education and Macmillan Publishing.

Hakstian, A. R. & Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika, 41*, 219-231.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics, 17,* 279-296.

Hedges, L. V. (1994). Statistical considerations. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 29-38). New York: Russell Sage Foundation.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61*, 404-420.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika, 28*, 221-238.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company.

Maeda, Y. & Rodriguez, M. C. (2002, April). *A Theoretical and Statistical Framework for the Meta-Analysis of Coefficient Alpha*. A paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Marascuilo, L. A. (1966). Large-sample multiple comparisons. *Psychological Bulletin, 65*, 280-290.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2000). *HLM* (Version 5.0) [Computer software]. Chicago, IL: Scientific Software International.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp.231-244). New York: Russell Sage Foundation.

Thompson, B. (Ed.). (2003). *Score Reliability: Contemporary Thinking on Reliability Issues*. Thousand Oaks, CA: Sage Publications.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509-522.

Table 1

*MARS Data Recovered from Primary Studies during Replication*

| *Study ID* | *Coefficient* | *Type* | *SD* | *Size* | *Items* | *Points* | *Children* | *College* | *Adults* |
|---|---|---|---|---|---|---|---|---|---|
| ch90.5 | .90 | 1 | 7.53 | 104 | 22 | 4 | 1 | 0 | 0 |
| ch90.1 | .90 | 1 | 10.02 | 40 | 22 | 4 | 1 | 0 | 0 |
| ch90.2 | .92 | 1 | 10.64 | 144 | 22 | 4 | 1 | 0 | 0 |
| ch90.3 | .92 | 1 | 11.39 | 171 | 22 | 4 | 1 | 0 | 0 |
| ch90.4 | .93 | 1 | 11.65 | 103 | 22 | 4 | 1 | 0 | 0 |
| pp82 | .98 | 1 | 20.55 | 170 | 24 | 5 | 0 | 0 | 1 |
| ste88 | .88 | 1 | 12.93 | 1086 | 26 | 5 | 1 | 0 | 0 |
| sa95 | .96 | 1 | 20.72 | 154 | 26 | 5 | 1 | 0 | 0 |
| B95 | .98 | 1 | 50.29 | 173 | 80 | 5 | 0 | 1 | 0 |
| dgg83.2 | .99 | 1 | - | 209 | 98 | 5 | 0 | 1 | 0 |
| dgg83.3 | .97 | 1 | - | 550 | 98 | 5 | 0 | 1 | 0 |
| mt86 | .98 | 1 | - | 138 | 98 | 5 | 0 | 1 | 0 |
| dgg83.1 | .96 | 1 | 58.77 | 767 | 98 | 5 | 0 | 1 | 0 |
| Rs72.2 | .97 | 1 | 65.29 | 397 | 98 | 5 | 0 | 1 | 0 |
| am98 | .86 | 2 | - | 62 | 25 | 5 | 0 | 1 | 0 |
| B89 | .72 | 2 | - | 50 | 75 | 5 | 1 | 0 | 0 |
| dgg83.4 | .87 | 2 | - | 155 | 98 | 5 | 0 | 1 | 0 |
| dgg83.5 | .95 | 2 | - | 30 | 98 | 5 | 0 | 1 | 0 |
| dgg83.6 | .86 | 2 | - | 125 | 98 | 5 | 0 | 1 | 0 |
| Rs72.1 | .85 | 2 | 51.26 | 35 | 98 | 5 | 0 | 1 | 0 |
| sens72 | .78 | 2 | 55.50 | 119 | 98 | 5 | 0 | 1 | 0 |

*Note*. Type 1 = alpha coefficient, type 2 = test-retest coefficient. *Study ID* is based on authors' last initials followed by year and coefficient number from that study.

Table 2

*Comparison of Original MARS RG results and Replication Results*

| | *n* | *Mean* | *SE* | *Q* |
|---|---|---|---|---|
| Original RG study of alpha coefficients | 28 | .915 | .0157 | |
| Recovered coefficients (unweighted) | 14 | .946 | .0099 | |
| Recovered coefficients (*T*-transformed, weighted) | 14 | .978 | .0023 | 2739* |
| Random-effects (*T*-transformed, weighted) | 14 | .957 | .0273 | 13 |
| Coefficients with *SD* information (unweighted) | 11 | .936 | .0106 | 1305* |
| Coefficients with *SD* (*T*-transformed, weighted) | 11 | .981 | .0022 | 242* |
| Random-effects (*T*-transformed, weighted) | 11 | .983 | .0109 | 10 |
| Original RG study of test-retest coefficients | 7 | .841 | .0276 | |
| Recovered coefficients (unweighted) | 7 | .841 | .0276 | |
| Recovered coefficients (*Z*-transformed, weighted) | 7 | .847 | .0424 | 21* |
| Random-effects (*Z*-transformed, weighted) | 7 | .851 | .0933 | 7 |

*Significant heterogeneity, *p*<0.01.

Table 3

*NEO-PI Alpha Coefficients Recovered from Primary Studies during Replication*

| Study ID | N | E | O | A | C | Sample | Gender | Language | Referent |
|---|---|---|---|---|---|---|---|---|---|
| sasmgs94 | .89 | .88 | .83 | .70 | .84 | 1 | 0 | 0 | 1 |
| h96 | .62 | .60 | .72 | .77 | .84 | 1 | 0 | 1 | 1 |
| pmc92 | .91 | .87 | .86 | .78 | .87 | 1 | 0 | 1 | 1 |
| sb92 | .84 | .71 | .66 | .78 | .87 | 1 | 0 | 1 | 1 |
| tvs95.1 | .91 | .87 | .90 | .70 | .86 | 1 | 1 | 1 | 1 |
| tvs95.2 | .88 | .90 | .88 | .79 | .88 | 1 | 2 | 1 | 1 |
| ras91 | .87 | .83 | .63 | .68 | .84 | 2 | 0 | 0 | 1 |
| sasmgs94 | .88 | .83 | .86 | .65 | .83 | 2 | 0 | 0 | 1 |
| cm88.1 | .93 | .87 | .89 | .76 | .86 | 2 | 0 | 1 | 1 |
| cm88.2 | .94 | .88 | .91 | .88 | .91 | 2 | 0 | 1 | 0 |
| paph95.2 | .91 | .91 | .88 | .76 | .80 | 2 | 1 | 0 | 1 |
| paph95.4 | .94 | .93 | .91 | .62 | .86 | 2 | 1 | 0 | 1 |
| paph95.1 | .93 | .91 | .90 | .73 | .80 | 2 | 2 | 0 | 1 |
| paph95.3 | .94 | .95 | .91 | .78 | .89 | 2 | 2 | 0 | 1 |
| fwspmc91.1 | .93 | .88 | .89 | .72 | .87 | 3 | 1 | 1 | 1 |
| fwspmc91.2 | .94 | .83 | .91 | .75 | .79 | 3 | 2 | 1 | 1 |

*Note*. Sample 1=student, 2=general, 3=clinical. Gender 0=both, 1=female only, 2=male only. Referent 0=other, 1=self. Language 0=not English, 1=English.

Table 4

*NEO-PI-R Data Recovered from Primary Studies during Replication*

| Study ID | N | E | O | A | C | Sample | Gender | Language | Referent |
|---|---|---|---|---|---|---|---|---|---|
| b97 | .92 | .88 | .88 | .88 | .91 | 0 | 0 | 0 | 1 |
| b96.1 | .92 | .86 | .89 | .87 | .90 | 1 | 0 | 0 | 1 |
| lb93 | .91 | .90 | .80 | .85 | .93 | 1 | 0 | 0 | 1 |
| kca96 | .89 | .83 | .80 | .78 | .89 | 1 | 0 | 1 | 1 |
| p94 | .94 | .92 | .86 | .95 | .96 | 1 | 0 | 1 | 1 |
| sl95 | .92 | .89 | .85 | .88 | .89 | 1 | 0 | 1 | 1 |
| oa94 | .93 | .88 | .88 | .84 | .90 | 2 | 0 | 0 | 1 |
| pc97.1 | .92 | .84 | .83 | .80 | .89 | 2 | 0 | 0 | 1 |
| pc97.2 | .92 | .87 | .82 | .82 | .92 | 2 | 0 | 0 | 1 |
| pc97.3 | .89 | .87 | .84 | .73 | .89 | 2 | 0 | 0 | 1 |
| cm95 | .92 | .89 | .89 | .87 | .91 | 2 | 0 | 1 | 1 |
| cmd91 | .92 | .89 | .87 | .86 | .90 | 2 | 0 | 1 | 1 |
| pc97.4 | .93 | .88 | .82 | .81 | .87 | 2 | 0 | 1 | 1 |
| pc97.5 | .88 | .91 | .85 | .86 | .89 | 2 | 0 | 1 | 1 |

*Note*. Sample 0=other, 1=student, 2=general, 3=clinical. Gender 0=both, 1=female only, 2=male only. Referent 0=other, 1=self. Language 0=not English, 1=English.

Table 5

*NEO-FFI Data Recovered from Primary Studies during Replication*

| Study ID | N | E | O | A | C | Sample | Gender | Language | Referent |
|---|---|---|---|---|---|---|---|---|---|
| crkg97.1 | .67 | .54 | .26 | .52 | .72 | - | - | 1 | - |
| bhlp96 | .86 | .82 | .69 | .70 | .85 | 0 | 0 | 1 | 1 |
| fmcb97.1 | .91 | .73 | .73 | .77 | .85 | 0 | 0 | 1 | 1 |
| fmcb97.2 | .89 | .81 | .78 | .84 | .89 | 0 | 0 | 1 | 0 |
| mn96 | .75 | .84 | .83 | .75 | .74 | 1 | - | 1 | 1 |
| b96.2 | .85 | .81 | .70 | .61 | .81 | 1 | 0 | 0 | 1 |
| crkg97.2 | .72 | .68 | .71 | .66 | .80 | 1 | 0 | 1 | 1 |
| gs96 | .85 | .80 | .74 | .77 | .84 | 1 | 0 | 1 | 1 |
| h92 | .87 | .82 | .76 | .79 | .85 | 1 | 0 | 1 | 1 |
| pb92 | .88 | .80 | .70 | .79 | .83 | 1 | 0 | 1 | 1 |
| hf94 | .87 | .84 | .73 | .75 | .81 | 1 | 2 | 1 | 1 |
| ras97.2 | .85 | .79 | .62 | .78 | .86 | 2 | 0 | 0 | 0 |
| ras97.1 | .85 | .80 | .63 | .69 | .82 | 2 | 0 | 0 | 1 |
| a95 | .89 | .79 | .76 | .74 | .84 | 2 | 0 | 1 | 1 |
| hfm94 | .82 | .78 | .56 | .71 | .85 | 2 | 0 | 1 | 1 |
| mwvkh94.2 | .78 | .72 | .56 | .68 | .82 | 2 | 1 | 1 | 1 |
| mwvkh94.1 | .83 | .74 | .62 | .70 | .82 | 2 | 1 | 1 | 1 |
| ts95.1 | .86 | .75 | .76 | .72 | .84 | 2 | 1 | 1 | 1 |
| ts95.2 | .85 | .79 | .72 | .70 | .83 | 2 | 2 | 1 | 1 |

*Note*. Sample 0=other, 1=student, 2=general, 3=clinical. Gender 0=both, 1=female only, 2=male only. Referent 0=other, 1=self. Language 0=not English, 1=English.

Table 6

*Summary Unweighted Statistics for NEO Domain Reliability Coefficients (Z-transformed)*

| *NEO domain* | *Original Caruso RG results (n=51)* | *All coefficients in replication (n=53)* | *Alpha coefficients with SD info (n=29)* |
|---|---|---|---|
| Neuroticism | .88 (.010) | .88 (.036) | .87 (.050) |
| Extraversion | .83 (.013) | .83 (.038) | .83 (.058) |
| Openness | .79 (.018) | .80 (.042) | .79 (.056) |
| Agreeableness | .75 (.014) | .75 (.036) | .73 (.051) |
| Conscientiousness | .83 (.011) | .85 (.026) | .83 (.036) |

*Note.* Standard errors in parentheses.

Table 7

*Summary Weighted Statistics for NEO Domain Reliability Coefficients (T-Transformed)*

| | *Unadjusted T (n=48)* | | *Variance & item adjusted T (n=29)* | |
|---|---|---|---|---|
| *NEO domain* | *Fixed-effects* | *Random-effects* | *Fixed-effects* | *Random-effects* |
| Neuroticism | .89 (.002) | .89 (.011) | .93 (.003) | .92 (.012) |
| Extraversion | .85 (.002) | .84 (.011) | .90 (.003) | .89 (.013) |
| Openness | .81 (.002) | .81 (.014) | .88 (.003) | .87 (.012) |
| Agreeableness | .78 (.002) | .78 (.011) | .89 (.003) | .86 (.013) |
| Conscientiousness | .86 (.002) | .86 (.008) | .93 (.003) | .92 (.010) |

*Note.* Standard errors in parentheses.

Table 8

*TES Data Recovered from Primary Studies during Replication*

| | PTE | | | GTE | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Study ID | Alpha | SD | Items | Alpha | SD | Items | Exp | Level | Area | Gender | Size |
| smr88 | .79 | | 14 | .64 | | 10 | 0 | | | 75 | 435 |
| hw90 | .84 | .50 | 12 | .72 | .61 | 8 | 0 | 1 | 0 | 90 | 191 |
| wh90 | .82 | .60 | 12 | .74 | .70 | 8 | 0 | 1 | 0 | 85 | 182 |
| cb97 | .75 | .60 | 13 | | | | 1 | | 1 | 86 | 378 |
| c92 | .75 | .62 | 8 | .55 | .86 | 5 | 1 | 0 | 0 | 70 | 170 |
| gd84 | .78 | | 9 | .75 | | 7 | 1 | 0 | 0 | 75 | 208 |
| jkd93 | .89 | .79 | 13 | .82 | 1.11 | 9 | 1 | 0 | 0 | 62 | 26 |
| mf88 | .63 | | 11 | | | | 1 | 0 | 0 | | 230 |
| sp94 | .74 | .59 | 9 | .66 | | 7 | 1 | 0 | 0 | 90 | 110 |
| a94 | .76 | .40 | 9 | .56 | .50 | 7 | 1 | 0 | 1 | 96 | 112 |
| mbdg96 | .75 | .61 | 8 | .68 | .89 | 6 | 1 | 0 | 1 | 93 | 298 |
| hlw98 | .74 | .65 | 6 | .70 | .82 | 7 | 1 | 1 | 0 | | 239 |
| msp95 | .87 | .66 | 9 | | | | 1 | 1 | 0 | 100 | 78 |
| ps93 | .75 | | 10 | .65 | | 6 | 1 | 1 | 0 | 83 | 240 |
| rlf | .74 | .61 | 8 | .65 | .81 | 5 | 1 | 1 | 0 | 80 | 194 |
| r92 | .69 | .55 | 9 | .73 | .80 | 7 | 1 | 1 | 0 | | 18 |
| sp96 | .80 | | 11 | .73 | | 4 | 1 | 1 | 0 | 79 | 310 |
| wp97 | .84 | .58 | 9 | .72 | .81 | 6 | 1 | 1 | 0 | 83 | 82 |
| wrh90 | .81 | | 12 | .77 | | 8 | 1 | 1 | 0 | 91 | 55 |
| sp93 | .76 | .68 | 8 | .70 | .80 | 7 | 1 | 1 | 1 | 68 | 192 |

*Note*. Experience 0=preservice, 1=inservice. Level 0=elementary school, 1=mixed. Area 0=regular education, 1=special education. Gender is proportion of majority gender. Size is sample size.

Table 9

*STEBI Data Recovered from Primary Studies during Replication*

| Study | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Alpha | SD | Items | Alpha | SD | Items | Exp | Level | Area | Gender | Size |
| sh95 | 0.84 | 0.44 | 13 | 0.73 | 0.35 | 10 | 0 | 0 | | 88 | 84 |
| esr95 | 0.85 | | 13 | 0.76 | | 10 | 0 | 0 | 0 | 89 | 71 |
| sb98 | 0.89 | | 13 | 0.80 | | 10 | 0 | 0 | 0 | | 619 |
| er90 | 0.90 | 0.59 | 13 | 0.76 | 0.46 | 12 | 0 | 0 | 0 | 87 | 212 |
| re90 | 0.92 | 0.61 | 13 | 0.77 | 0.43 | 12 | 1 | 0 | 0 | 88 | 331 |

*Note*. Experience 0=preservice, 1=inservice. Level 0=elementary school, 1=mixed. Area 0=regular education, 1=special education. Gender is proportion of majority gender. Size is sample size.

Table 10

*Summary Statistics for TES Reliability Coefficients, Unweighted and Untransformed*

| | *Original RG results* | | *Replication* | |
|---|---|---|---|---|
| *Instrument Subscale* | *M (SD)* | *n* | *M (SD)* | *n* |
| TES: PTE | .778 (.057) | 25 | .775 (.061) | 20 |
| TES: GTE | .696 (.072) | 21 | .692 (.070) | 17 |
| STEBI: PSTE | .885 (.050) | 13 | .880 (.034) | 5 |
| STEBI: STOE | .761 (.025) | 11 | .764 (.025) | 5 |
| TLC: I+ | .740 (.020) | 7 | .724 (.057) | 3 |
| TLC: I– | .700 (.130) | 7 | .776 (.014) | 3 |
| RSA+ | .760 (.030) | 5 | .760 | 1 |
| RSA– | .840 (.040) | 5 | .830 | 1 |

Table 11

*Summary Statistics for TES Reliability Coefficients, Adjusted for Variance and Number of Items, T-Transformed and Weighted*

| | *n* | *Mean* | *SE* | *Q* |
|---|---|---|---|---|
| PTE Original RG Study | 25 | .778 | .011 | |
| PTE Fixed Effects | 14 | .845 | .001 | 105* |
| PTE Random-Effects | 14 | .847 | .016 | 12 |
| GTE Original RG Study | 21 | .696 | .016 | |
| GTE Fixed-Effects | 11 | .782 | .008 | 63* |
| GTE Random-Effects | 11 | .775 | .020 | 9 |

** Significant heterogeneity, p<.001.*

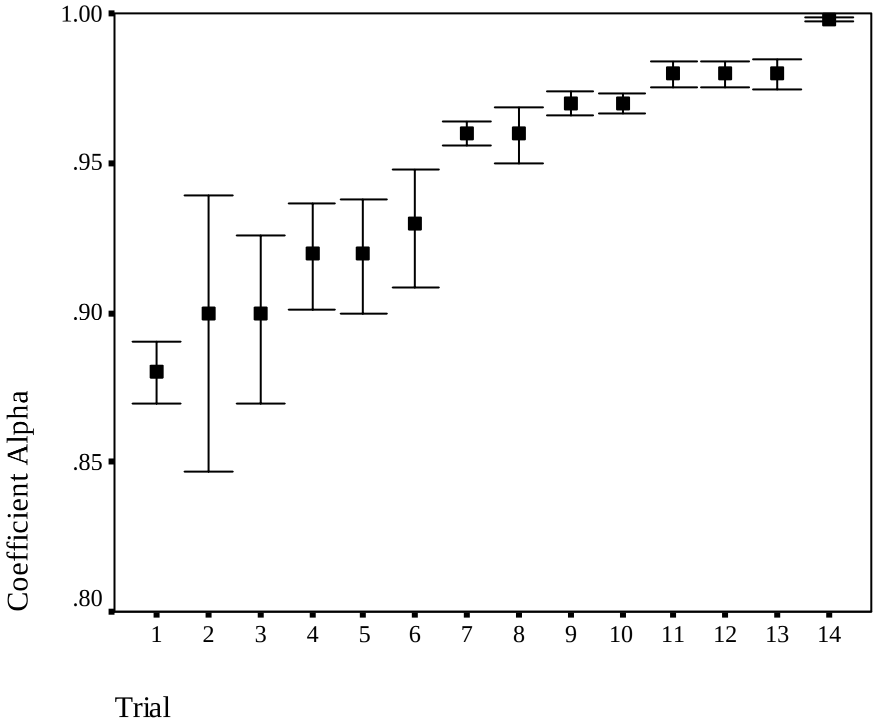
Figure Captions

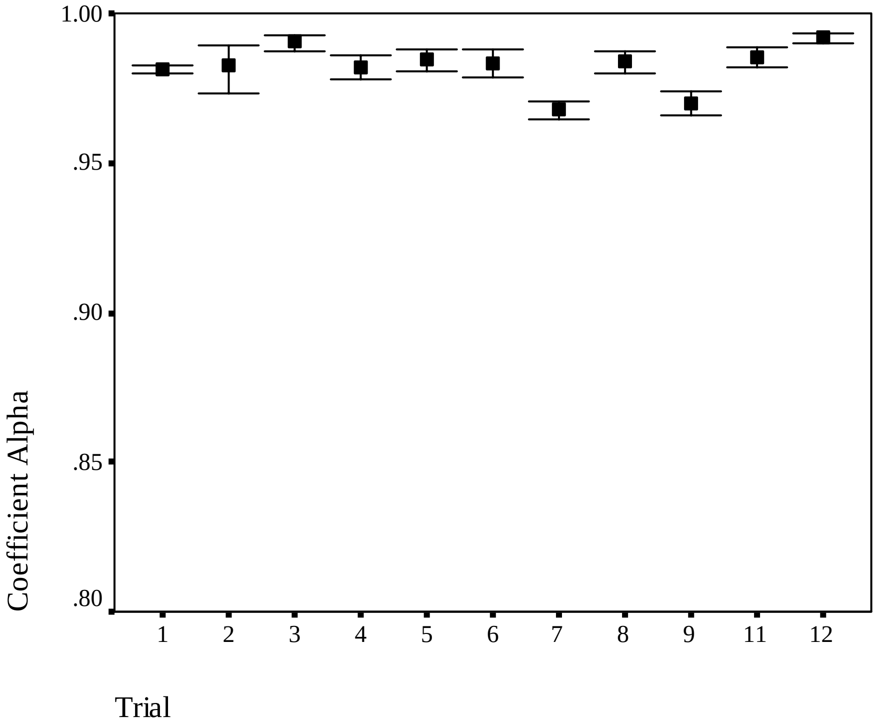
*Figure 1*. T-transformed MARS coefficient alpha fixed-effects 95% confidence intervals (*n*=14).

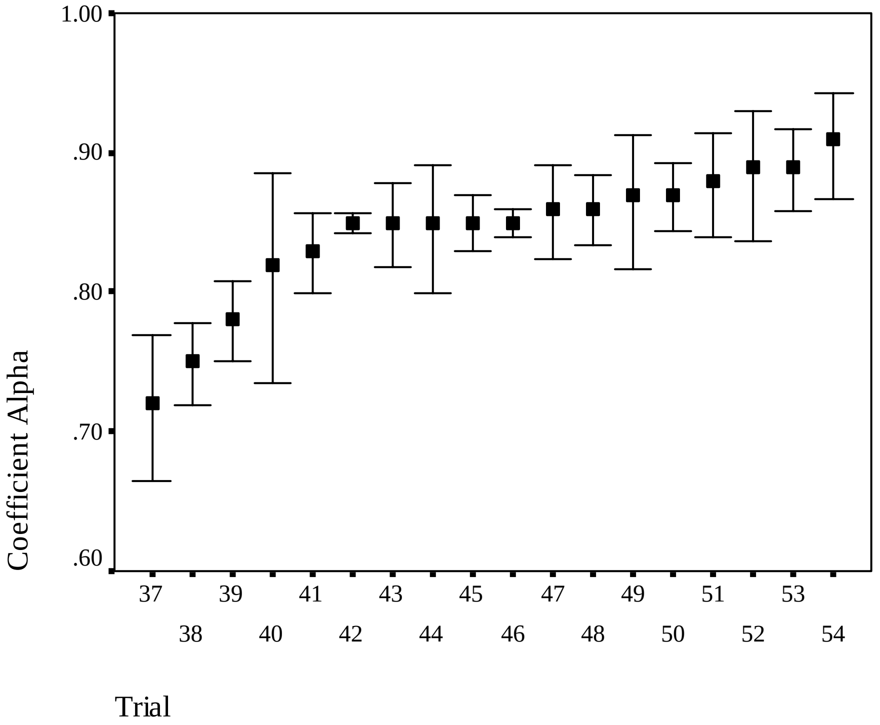
*Figure 2*. T-transformed MARS coefficient alpha fixed-effects 95% confidence intervals adjusted for number of items and group heterogeneity of 11 MARS coefficients.

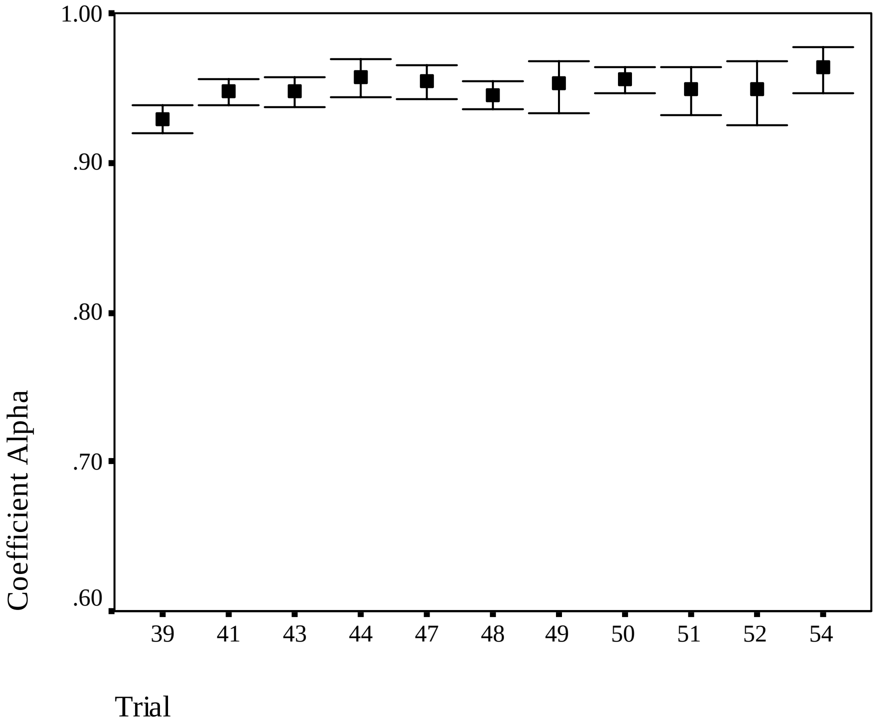
*Figure 3.* T-transformed NEO-FFI coefficient alpha fixed-effects 95% confidence intervals (*n*=18).

*Figure 4*. T-transformed NEO-FFI coefficient alpha fixed-effects 95% confidence intervals adjusted for number of items and group heterogeneity of 11 NEO-FFI coefficients.


1.00
.95
.90
.85
.80
Coefficient Alpha
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Trial


1.00
.95
.90
.85
.80
Coefficient Alpha
1
2
3
4
5
6
7
8
9
11
12
Trial


1.00
.90
.80
.70
.60
Coefficient Alpha
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
Trial


1.00
.90
.80
.70
.60
Coefficient Alpha
39
41
43
44
47
48
49
50
51
52
54
Trial