Statistical Issues of Reliability Generalization and an Application to Achievement Data

Yukiko Maeda and Michael C. Rodriguez

University of Minnesota

April5, 2002

Revised, April 13, 2002

Paper presented at the annual meeting of the American Educational Research Association,

New Orleans, LA.

Copies of this manuscript may be obtained from M. C. Rodriguez at 206 Burton Hall, University of Minnesota, 178 Pillsbury Drive, Minneapolis, MN 55455; or at the following website: http://measurement.class.umn.edu/aera/papers.html

Statistical Issues of Reliability Generalization and an Application to Achievement Data

Meta-analysis has become an important tool in the social sciences, providing a strong class of techniques to the integration of volumes of research. Moving beyond narrative reviews, meta-analysis employs statistical methods to the synthesis of quantitative findings of many researchers.

There are many approaches to meta-analysis. Five methods were described by Bangert-Drowns (1986). Each is closely related to Glass' (1976) original conceptualization with primary differences in unit of analysis and treatment of study variation, and each is briefly described here in terms of Bangert-Drowns' labels. "Glassian" meta-analysis begins with the collection of all relevant studies with liberal inclusion criteria. The empirical outcomes of these studies (allowing multiple outcomes per study) are transformed into a common effect size metric. The distribution of these outcomes is described and study level characteristics are employed to explain variation in outcomes. A related method, "study effect meta-analysis," restricts the unit of analysis to one effect per study with strict selection criteria for including studies in the synthesis.

A "combined probability method" combines p values and effect sizes (separately) for each study, but pays less attention to study characteristics. "Approximate data pooling with tests of homogeneity" is a method that accounts for not only variability among study outcomes but also the variance of each effect size with a corresponding test of homogeneity of effects. In addition, researchers in this method encourage optimal weighting of study effects to minimize the variance of the resulting summary statistic. Bangert-Drowns (1986) identified Hedges (1981, 1982) and Rosenthal and Rubin (1982a, 1982b) as primarily responsible for developments in these areas.

Finally, “approximate data pooling with sampling error correction” is closely related to the previous method, with the exclusion of homogeneity testing and the inclusion of corrections for artifactual mediators, such as sampling error, variation in reliability of measures, and range restriction. Hunter and Schmidt (1990) and Hunter, Schmidt, and Jackson (1982) are primarily responsible for this line of development.

In psychology, meta-analyses involving artifactual corrections are most notably found in the area of validity generalization. Schmidt and Hunter (1977) use multiple corrections for error and bias in research findings to complete a synthesis of criterion-related validity coefficients (typically Pearson correlations) as reported in research. These methods include corrections for sampling error, measurement error in the criterion or predictor, range restriction (group homogeneity), and other statistical artifacts that impact the magnitude of effect sizes (for a review, see Schmidt & Hunter, 1999). In validity generalization, the argument is essentially that the degree to which criterion-related validity evidence from one setting can be used in another setting is a function of the accumulation of findings (quantitative synthesis of prior research). If validity generalization evidence is limited or weak, local validity evidence will be necessary to justify the use of test scores for important decisions.

Recently, Vacha-Haase (1998) and others have transferred these methods to the generalization of reliability coefficients in a method called <u>reliability generalization</u> (RG). She suggested that the application of this technique to reliability coefficients was an appropriate way 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). Although the synthesis of reliability coefficients has been occurring for decades, the classification of these studies as RGs is recent. Since the introduction of the RG label, several RG studies have been reported in the literature. Most appear to employ the methods of Hunter and Schmidt (1990), as cited by the various RG authors, although few engage in the level of error and bias correction employed by Hunter and Schmidt. In addition, it is not clear how closely these methods are aligned with those of Hunter and Schmidt because authors of RG studies do not describe their synthesis methods explicitly.

In many of the RG studies, the classical test theory conceptualization of reliability is also presented. In their comprehensive chapter on reliability, Feldt and Brennan (1989) presented two ideas that have been overlooked by RG authors. Without completely reviewing conceptions of classical test theory here, recall that observed scores are composed of two components, the true score and the error score. The true score is classically considered the long run average (limit) of observed scores as the number of observations (or items) increases infinitely. The variance of those observed scores for a single individual is considered error variance. One limitation of this conception is the ambiguity regarding the conditions that can be permitted to vary across observations, which leads to various definitions of true score. Because observed scores vary about the true score, different definitions of the true score result in different definitions of error variance. “This, in turn, implies that any measuring instrument can have many reliabilities, even for a clearly identified population of examinees” (Feldt & Brennan, 1989, p. 107).

Feldt and Brennan (1989) also suggested that there are many situations where we may be interested in comparing reliability coefficients, including modification of testing procedures, scoring formulas, and training of scorers, observers, or raters; studies of item format and item-selection procedures; and comparison of multiple instruments, forms, or examinee (sample) populations. In doing such comparisons, they aptly reviewed the literature on the sampling distribution theory of reliability. Some researchers have considered reliability as a correlation (theoretically the squared correlation between observed and true scores or the correlation between two parallel forms). Sawilowsky (2000) in his critique of RG suggested a Fisher Z normalizing and variance stabilizing transformation of the reliability coefficient because of its distributional properties as a correlation. Some RG researchers have since employed the Fisher Z-transformation.

In their review of the sampling distribution theory of reliability, Feldt and Brennan (1989) demonstrated how Fisher's transformation is appropriate when parallel forms result in a product-moment correlation or when a single exam is evaluated with a split-halves correlation. Once a split-half coefficient is corrected (i.e., employing Spearman-Brown), Fisher's transformation is no longer appropriate (Lord, 1974). They also argued that the sampling distribution for coefficient alpha is not the same as that for product-moment coefficients.

It has been suggested that one of the appealing aspects of the RG technique is that the options of analysis are wide open (Henson & Thompson, 2001), with no standard or "single genre of analyses. For example, some authors will use box-and-whisker plots, others regression, some ANOVA"(Thompson & Vacha-Haase, 2000). Specific analytic method aside, we must not ignore what we know about statistics and the assumptions underlying each method. The quantitative synthesis of effect sizes has undergone a great deal of development and we now have at our disposal an integrated set of principles that can be applied under various specific analytical methods. This body of work is grounded in a long tradition of synthesis of statistics. To adequately address these issues and present a sound statistical and theoretical framework to the meta-analysis of reliability coefficients, we present a framework for meta-analysis grounded in statistical theory, review the sampling distribution theory for coefficient alpha (the most common reliability coefficient); and provide an application to statewide achievement test data.

A Principled Approach to Meta-Analysis

The meta-analytic methods described in The Handbook of Research Synthesis (Cooper & Hedges, Eds., 1994), hereinafter referred to as the Handbook, are based on the developmental work on the quantitative synthesis of the empirical results of multiple studies. The synthesis of study effects can be dated back to the early 1930s in studies conducted by Tippitt, Fisher, K. Pearson, E. S. Pearson, as described by Bangert-Drowns (1986) in his historical review. However, Glass (1976) is typically credited with the introduction of the term meta-analysis.

Meta-analysts who rely on homogeneity testing and optimal weighting of effects also employ a variety of tools based on the nature of the effects to be synthesized and the nature of the research questions. One clear departure from Hunter and Schmidt (1990) includes the use of statistical null-hypothesis testing. Although the debate regarding the utility of the null-hypothesis continues (Frick, 1996; Schmidt, 1996), the methods of homogeneity testing and optimal weighting are not exclusive of other methods. Most meta-analysts employ a wide variety of techniques to summarize study effects including a central role for confidence intervals.

The methods employed here include a basic set of principled steps through which the story of the synthesis can be told. These are briefly described in terms of the steps a Hedges (1981, 1982) and Rosenthal and Rubin (1982a, 1982b) meta-analyst may take. The following steps are discussed at length in the Handbook. This particular set of steps is most directly attributable to Becker (1999) who presented these steps at a recent AERA workshop.

(1) Gather studies. There are a wide variety of techniques employed at this stage just as there are a wide variety of techniques employed in the selection of relevant articles in a narrative review of the literature. We leave the intricacies of this step to the researcher, but caution is warranted because of the file drawer problem (e.g., unreported studies and rejected manuscripts)

for generalizability, among others. Researchers are encouraged to use liberal selection criteria and code characteristics they believe may impact the quality or nature of the study results.

(2) Code relevant study characteristics. Theory drives this step, as it does when gathering information on relevant variables in any study. A coding form is designed through which researchers can record relevant study characteristics to use in the meta-analysis, including the study effects or information needed to compute effect sizes. The choice of which effect size to synthesize also needs to be made by the researcher. Here, Hunter and Schmidt (1990) would argue for artifactual corrections to make study effects comparable and thus require that information be gathered to make such corrections. The approaches of many others would also allow for relevant corrections.

(3) Quantify study effect precision. The work of meta-analysis researchers has been strong in this area. Once significant issue in synthesizing effects regards the comparability of the effects (aside from any corrections which may be desired). If studies were actually identical replications of an original study, combining results would be a relatively straightforward exercise. In reality, studies differ in many ways. To deal with these differences, the synthesist has available a number of weighting schemes based on three assumptions:

> (a) Theory or evidence suggests that studies with some characteristics are more accurate or less biased with respect to the desired inference than studies with other characteristics, (b) the nature and direction of that bias can be estimated prior to combining results, and (c) appropriate weights to compensate for the bias can be constructed and justified. (Shadish & Haddock, 1994, p. 263)

Since each effect essentially results from a different study (different in terms of sample size, group variability, and other quality indicators including randomization, treatment fidelity, etc.), each effect is estimated with a different level of precision. The weights with the strongest statistical properties are a function of the variance of the effect. This weight minimizes the variance of the combined effects and accounts directly for the level of precision of the estimated effect; generally the variance of the effect is a function of its sampling distribution (e.g., commonly based on sample size). The researcher may desire other weights or combinations of weights, including an index of study quality. The choice of additional weighting factors are best made by the researcher; however, weighting is clearly necessary to preserve the integrity of each effect and account statistically for its precision in the combination of effects.

(4) Combine study effects. This step and the next are thought of as being simultaneous, but necessarily occur in stages. Researchers can combine study effects by computing a weighted mean and then quantify the uncertainty of this mean by computing a standard error of the weighted mean. The variability of study effects is also of primary interest.

(5) Test homogeneity of study effects. The first research question asked by most meta-analysts is: Do study effects appear similar across studies? We want to know if there is variation in the effects as reported in various research studies, potentially after correcting for study-based statistical artifacts. What follows is a clear departure from Hunter and Schmidt (1990). The test of homogeneity is based on the null hypothesis of equality of effects, where θ (theta) is any unknown parameter (e.g., a correlation, a standardized mean difference, an odds ratio, etc.), such that H_0: $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$ for K studies.

This test is grounded in statistical theory regarding the distribution of K independent, asymptotically normally distributed estimates $\hat{\theta}_k$ of parameter θ_0, each with a large-sample variance $\sigma^2_{\hat{\theta}_k}$. The derivation of this test was developed by Marascuilo (1965), who showed that the quantity $M = \sum_{k=1}^{K} \frac{\left(\hat{\theta}_k - \hat{\theta}_0\right)^2}{\mathrm{var}\left(\hat{\theta}_k\right)}$ (a kind of effect size variance ratio) is approximately distributed Chi-square with df = K –1. The estimate $\hat{\theta}_0$ of θ_0 is the weighted average of each K estimate,

$\hat{\theta}_0 = \frac{\sum_{k=1}^{K} \hat{W}_k \hat{\theta}_k}{\sum_{k=1}^{K} \hat{W}_k}$, where $\hat{W}_k = \frac{1}{\mathrm{var}\left(\hat{\theta}_k\right)}$. Here we see that the weight used is a function of the variance of the study-based parameter estimate. The "approximate data pooling with tests of homogeneity" camp has adopted these procedures to complete the null hypothesis test of homogeneity. The homogeneity test has also been evaluated (Harwell, 1997), and although it generally behaved as expected, there were conditions where the homogeneity test should be used with caution, including cases of nonnormal score distributions, unequal sample sizes, and particularly when the ratio of within-study sample size to the number of studies is less than 1.0.

The following steps depend, partly, on the results of the homogeneity test. Step 6 is appropriate if the study effects appear homogeneous across studies (i.e., there is no study effect variation). The subsequent steps are appropriate if there is significant heterogeneity in effects across studies.

(6) Describe the common effect. Under failure to reject the null hypothesis of effect homogeneity, we conclude that there is a common effect – study effects do not vary across studies. We can then estimate θ_0 by reporting the weighted mean as described in step 5 and quantify its uncertainty in terms of the standard error of the mean. Since the weights used to compute the weighted mean are a function of the variance of each estimate, $\hat{W}_k = \frac{1}{\mathrm{var}\left(\hat{\theta}_k\right)}$, the

variance of the mean will be the inverse of the sum of the weights, $v_{\bullet} = \frac{1}{\sum_{k=1}^{K} \hat{W}_k}$, so that the standard error of the mean will be its square root, $\sqrt{v_{\bullet}}$.

With the standard error, a confidence interval can be computed with the usual formula, $\hat{\theta}_0 \pm t_{crit} \sqrt{v_{\bullet}}$, with the critical t-value associated with the desired width of the confidence interval.

(7) Model the heterogeneity of effects. Under rejection of the null hypothesis of effect homogeneity, we conclude that there is no common effect – there is significant variation in effects across studies. At this point, it becomes necessary to decide how to best describe the model under which this variation may be explained – a fixed-effects, random-effects, or mixed model. This is a step that many meta-analysts fail to select explicitly and so default to a fixed-effects model. The decision rules here are not without some judgment on the part of the researcher; however, we argue that the decision should be made explicit and on theoretical and empirical bases.

Briefly, the decision regarding the model to employ depends on the universe to which generalizations are made. This addresses issues related to the choice between fixed or random effects inferences and subsequent statistical analyses. Hedges (1994) described the differences:

> In the fixed effects, or conditional, model, the universe to which generalizations are made consists of ensembles of studies identical to those in the study sample except for the particular people (or the primary sampling units) that appear in the studies. Thus, the studies in the universe differ from those in the study sample only as a result of the sampling of people into the groups of the studies. The only source of sampling error or uncertainty is therefore the variation resulting from the sampling of people into studies (p. 30).
>
> In the random effects, or unconditional, model, the study sample is presumed to be literally a sample from a hypothetical collection (or population) of studies. The universe to which generalizations are made consists of a population of studies

> from which the study sample is drawn. Studies in this universe differ from those in the study sample along two dimensions. First, the studies differ from one another in study characteristics and in effect size parameter. …Second, in addition to differences in study characteristics and effect size parameters, the studies in the study sample also differ from those in the universe as a consequence of sampling of people into the groups of the study (p. 31).

The argument to support the use of random effects analysis suggests that the studies we observe in the literature are, more or less, accidental or due to chance as much as anything. The question, as clarified by Hedges (1994), is not “What is true about these studies?” but “What is true about studies like these that could have been done?” Such generalizations can be handled through statistical means by incorporating the additional uncertainty due to the inference to studies not identical to those in the sample. Upon identification of the appropriate universe of generalization, the meta-analyst can proceed. However, there are additional considerations.

Hedges (1994) suggested that all of the available methods commonly used in statistical analysis of study results assume two conditions. The studied effect (or appropriate transformation) is normally distributed in large-samples with a mean that approximates the parameter of interest. In addition, the standard error of the estimated effect is a continuous function of its study sample size, the magnitude of the effect, and potentially other factors that can be estimated from the resulting study data. Once these considerations have been addressed, results from studies can be combined.

The analysis of a random effects model requires the estimation of an additional source of uncertainty that is due to the sampling of studies from the universe of generalization in addition to the sampling error within studies. Uncertainty, when considering sample studies to be representative of a larger universe, comes from the fact that study contexts, treatments, and administration procedures differ in many ways that potentially impact results. The inclusion of the additional random variance component in the weighting scheme and confidence intervals captures the degree of additional uncertainty. These specific methods are described more fully below.

(8) Conduct random effects synthesis. If the model selected is a random effects model, we can answer the question: How much variation is found in the population effects? We estimate the random variance component, compute the average effect ($\hat{\theta}_0$) using the random-effects weighted mean and quantify uncertainty using the random-effects augmented standard error of the mean. Confidence intervals and, if desired, a test of the null-hypothesis H_0: $\theta_0 = 0$ can be computed.

(9) Conduct fixed-effects or mixed model synthesis. If the model selected is a fixed-effects model or mixed model, we can answer the question: What kinds of study characteristics explain variation in study effects? Depending on the nature of the study characteristics, any number of analytical tools may be employed, including a wide range of categorical and continuous data analytic methods. The questions answered are of the general form: Do study characteristics explain between-study differences? Meta-analysts who employ the Q-test of homogeneity would then use this test again to assess potential reductions in heterogeneity.

If significant variation remains, the meta-analyst may then want to employ a mixed model by quantifying the additional (random-effects) uncertainty and including it in the computation of results, as in step 8 above.

This constitutes a rough outline of the procedures in the “approximate data pooling with tests of homogeneity” camp of meta-analysis. What follows is hopefully a user-friendly adaptation of these procedures to the meta-analysis of reliability coefficients, with attention to appropriate conditions and cautions as they arise.

Sampling Distribution Theory

Authors of RG studies conducted since 1998 include, to some extent, a discussion of two facts regarding reliability: (1) reliability is a property of test scores and not the test itself and (2) reliability is thus sample specific and its magnitude is impacted by sample characteristics. These facts have been reviewed and cited with significant authority. However, what has been absent is an appreciation of the statistical sensitivity of the analytical methods employed given the nature of the variables included, particularly the sampling distribution of various reliability coefficients.

The assumptions of general linear models are well known. Structural assumptions that impact the validity of our interpretations of results include independence of observations, error free measurement and independence of explanatory variables, normally distributed variables, linear interrelationships, and correct model specification. Stochastic assumptions that impact the integrity of statistical tests include the independent and identical distribution of errors or model residuals, that is, residuals are independent and normally distributed with a mean of zero and constant variance. To satisfy these assumptions, we typically design studies grounded in theory with sound measurement controls, rely on known sampling distribution theories of our statistics, and conduct residual analyses.

Implicit in this presentation of the sampling distribution theory of coefficient alpha is that researchers conducting RG studies have ignored the distributional properties of coefficient alpha (also suggested in a critique of RG methodology by Sawilowsky, 2000). The sampling distribution of coefficient alpha can be expected to become more skewed as the population parameter approaches unity. In the case of RG studies, which utilize results of measurement instruments that have been employed in the research literature at a moderate to high level of frequency, the reliabilities are likely to be fairly high. In a synthesis of coefficient alphas reported in 24 journals, two conference proceedings, and a sample of unpublished manuscripts from 1960 to 1992, the 4,286 “harvested” coefficients ranged from 0.06 to 0.99 with an unweighted mean of 0.77 and median of 0.79, with a negative skew (Peterson, 1994). Certainly at this level, we cannot expect the distribution of coefficient alpha to be normal.

Kristof (1963) and Feldt (1965) independently derived the sampling distribution for the sample coefficient alpha, based on a transformation of coefficient alpha they showed to be distributed as F. They used the results of this derivation to test hypotheses about coefficient alpha and compute confidence intervals. Feldt (1969) then developed a test for two independent coefficients and Hakstian and Whalen (1976) extended this test to K independent coefficients, where K is any number of coefficients. Feldt (1980) introduced a test of two dependent coefficients (where reliabilities were based on the same sample) and Woodruff and Feldt (1986) extended this test to K dependent coefficients. In all of these derivations, simulation studies were conducted to investigate the integrity of the procedures. Feldt, Woodruff, and Salih (1987) provided a review of these results in an integrated discussion of statistical inference for coefficient alpha.

Sampling distributions for special forms of coefficient alpha and other reliability coefficients have also been derived. Methods are available for testing the equality of independent coefficient alpha adjusted for test length (Alsawalmeh & Feldt, 1999) and equality of intraclass reliability coefficients (Alsawalmeh & Feldt, 1992, 1994; Feldt, 1990; Kraemer, 1981).

The predominant reliability coefficient in the literature is coefficient alpha (or the equivalent KR-20 for dichotomously scored items), an estimate of internal consistency, because it is easily obtained by the researcher through a single administration of a single form of a measurement instrument. In the context of meta-analysis, this requires a method for combining K

independent alpha coefficients. For this we relied on the work of Hakstian and Whalen (1976) and their K-sample significance test.

Feldt (1965) and Kristof (1963) who developed the sampling theory for the sample coefficient alpha (r_α) as an estimate of the true parameter (ρ_α) for scores from n subjects on J items, demonstrated that the ratio $\frac{1-r_\alpha}{1-\rho_\alpha}$ is distributed as F with df = (n – 1)(J – 1). By relying on the work of Paulson (1942), Hakstian and Whalen (1976) used the normalizing transformation of F to obtain the nonlinear monotonic normalizing transformation of the sample coefficient alpha, $(1-r_\alpha)^{1/3}$. They also noted that like the Fisher Z transformation of a sample product-moment correlation, this transformation is biased; however, the bias was less than that possessed by the Fisher Z, which is usually ignored.

The overall significance test is then based on the work of Marascuilo (1966) presented above in step 5, given K independent asymptotically normally distributed estimates, $\hat{\theta}_k$, of the unknown parameter θ_0. Here we will employ some of the notation of the Handbook to facilitate utilization of these techniques by the meta-analyst. In the translation, the unknown parameter θ_0 is the population coefficient alpha, ρ_α, while each estimate $\hat{\theta}_k$ is noted as $r_{\alpha k}$. In the generalized notation employed in the Handbook, each of the K study effects are T_k, with a weighted mean $\bar{T}_\bullet$. In our case, the study effect is the transformed sample coefficient alpha: $T_k = (1-r_{\alpha k})^{1/3}$. The weighted mean study effect is $\bar{T}_\bullet = \frac{\sum w_k T_k}{\sum w_k}$, where the weights are the reciprocal of the variance of each study effect, $w_k = \frac{1}{v_k}$. The variance of study effect k is $v_k = \frac{18J_k(n_k-1)(1-r_{\alpha k})^{2/3}}{(J_k-1)(9n_k-11)^2}$ (as derived by Hakstian & Whalen, 1976) and the variance of the mean is $v_{\bullet} = \frac{1}{\sum w_k}$ with a standard error of the mean $\sqrt{v_{\bullet}}$.

With these values, the meta-analyst has all that is needed to proceed with the synthesis and complete the test of homogeneity of study effects, where $Q = \sum \frac{\left(T_k - \overline{T}_{\bullet}\right)^2}{v_k}$ is distributed χ^2 with df = K – 1. This test is not without assumptions, primarily having samples large enough to satisfy the asymptotic χ^2 distribution. Hakstian and Whalen (1976) employed monte carlo methods to evaluate the impact of sample size and found that the test maintained good control of Type I error rates with as few as 20 subjects per instrument under conditions of nonnormality and heterogeneity of variance.

In the spirit of completeness, an important consideration that was addressed earlier (step 7) is the choice of analytic model. The statistics computed here are all based on a weight that is strictly a function of study effect variance under a fixed-effects model, the conditional variance. In a random effects model, the population coefficient alpha of interest, $\rho_{\alpha k}$, is not fixed but random with its own distribution. Note here the population parameter has an additional subscript k. Total variability of an observed study estimate includes both conditional variation, v_k, of the estimate around each population $\rho_{\alpha k}$ and random variation, σ^2_{ρ}, of $\rho_{\alpha k}$ around the mean population parameter.

Estimation of the random variance component (the between-studies variance) is no easy matter. Various estimation procedures result in different estimates with important consequences (see Raudenbush, 1994). In this context, the weights, w_k, that minimize the variance of the estimated mean population parameter $\overline{T}_{\bullet}$ are inversely proportional to the sum of the conditional and random-effects variance, where $w_k = \frac{1}{v_k + \sigma_\rho^2} = \frac{1}{v_k^*}$. Uncertainty, when considering sample studies to be representative of a larger universe, comes from the fact that study contexts, treatments, and administration procedures differ in many ways that potentially impact results.

Threats to valid inferences from such a synthesis remain. Among those described by Raudenbush (1994) are uncertainty about the random effects variance component, the tenability of the assumption that the random effects are normally distributed with constant variance, model misspecification, and including multiple effects from single studies resulting in dependent data. None of the reliability generalization studies published to date have addressed these issues.

An Application to State Test Data

Sample

The data utilized for this study were obtained from the Minnesota Basic Standard tests, which are administered as a part of the state's accountability system. The Minnesota Basic Standard test consists of three subtests: the mathematics and reading tests, which are first given in February to eighth grade students and writing composition test, which is first given in February to tenth grade students. Two sets of data resulting from eighth grade mathematics and reading tests, which were administered in the spring of 2000, were used for this study.

The mathematics test, consisting of 68 four-option multiple-choice items, covers material that students in general initially learn before sixth grade and frequently appear in adult life. These include simple problems of arithmetic, geometry and algebra. The utilized mathematics test data set originally included the test scores from 827 schools, representing a total of 71,033 students. The test scores of students who answered less than 95% of the total number of questions were excluded from the data set and the schools with fewer than 20 students who took the test were excluded because of the instability of sample statistics that could result from this small school size. The final mathematics data set consisted of the remaining 512 schools, representing the test scores of 68,721 students.

The reading test, consisting of 40 four-option multiple-choice items, measures students' understanding of factual information. The utilized reading test data set initially included the test scores from 824 schools, representing a total of 70,180 students. Again, applying the criteria used for the mathematics data, the final reading data set included 515 schools, representing 67,919 students.

Procedure

For both the mathematics and reading test score data sets, coefficient alpha was computed for each school and then transformed into the corresponding T (employing our effect size notation) using the normalizing transformation of coefficient alpha introduced by Hakstian and Whalen (1976).

The test score mean and variance (observed variance) were computed for each school. The ratio of the error variance to the true variance, $\frac{\sigma_E^2}{\sigma_T^2}$, was also estimated for each school using the method described below to complete the analyses. Within classical test theory, we assume that the error variance is constant, so that coefficient alpha varies as a function of the observed variance. However, Feldt and Qualls (1999) demonstrated in their study that, since the commonly appearing formula of reliability can be redefined as $\rho_{xx'} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = \frac{1}{1 + \frac{\sigma_E^2}{\sigma_T^2}}$,

coefficient alpha varies as functions of the ratio $\frac{\sigma_E^2}{\sigma_T^2}$.

To obtain this ratio, first, error variance (σ_E^2) was estimated for each school from the observed variance (σ_X^2) and coefficient alpha, applying the classical test theory formula $S_e^2 = S_x^2(1 - r_{xx'})$. Second, true variance (σ_T^2) was estimated using the relationship among σ_E^2, σ_X^2 and σ_T^2, also defined by classical test theory. Using these estimated values, the ratio of error variance to true variance was computed for each school.

Finally, following the outlined steps (from step 4 to 9) described above, the 512 mathematics tests' Ts (transformed coefficient alphas) and the 515 Ts for the reading test were synthesized for the RG analyses.

In addition, to examine the effect of the variation of error variance among schools on the variation of alpha coefficients, the computed original alphas were adjusted in such a way that score variance for each school was equalized, based on the value of the score variance at the 50th percentile, using the relationship between the original reliability (r_{xx}) and variance (S_x^2) and the adjusted reliability (r_{uu}) and variance (S_u^2, in our case, the 50th percentile variance of our sample) $r_{uu} = 1 - \frac{S_x^2(1 - r_{xx})}{S_u^2}$ (Gulliksen, 1987). 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). Hereafter, this value is referred to as the adjusted coefficient alpha, accounting for variation in group heterogeneity (score variance).

Result

The three main objectives of RG studies are to identify 1) the typical score reliability of a specific instrument or measurement (i.e. the mathematics test and the reading test for this study),

2) the amount of variation of the score reliability for this instrument or measurement, and 3) the variable(s) affecting score reliability (Vacha-Hasse, 1998).

Typical Score Reliability

For the mathematics exam, the mean unweighted alpha was 0.907 with a median of 0.918. As can be seen in Figure 1, the distribution is highly skewed (skewness = -2.59). When reliability coefficients were adjusted to account for differences in variance, the mean coefficient was 0.918 with a median of 0.919. The distribution more closely resembled a normal distribution (skewness = -0.663). When reliability coefficients were transformed into the T metric, the mean unweighted reliability was 0.911, also with less skewness (1.42) than the original coefficient alpha distribution. The mean weighted T transformed reliability was 0.924.

In addition, standard errors were computed for each estimate. As expected, the standard errors for the original mean alpha (0.00176) and mean alpha based on unweighted T (0.00144) were much larger than those for mean adjusted alpha (0.000660) and mean alpha based on weighted T (0.000421).

This pattern of results was similar for the reading exam. The mean unweighted alpha was 0.840 with a median of 0.870. This distribution was also highly skewed (skewness = -2.71, Figure 2). Reliability coefficients adjusted for variance had a mean of 0.872 with a median of 0.874. The distribution was also closer to normality (skewness = -1.088). When reliability coefficients were transformed into the T metric, the mean unweighted alpha was 0.852, also with less skewness (1.68). The mean weighted T transformed alpha was 0.875. Again, standard errors of these means followed a similar pattern as well (Table 1).

To provide a summary of these results, mean estimates of coefficient alpha under each technique and their 95% confidence intervals are illustrated in Figures 3 and 4.

A comment on weighting is appropriate at this point. In several RG studies, study effects were weighted by sample size (in some cases optimal weighting is based on an effect size variance estimate that is largely a function of sample size). We examined sample size weighted coefficient alphas for the mathematics test data. The first attempt was to obtain a weighted mean coefficient alpha, weighted for sample size. This yielded a total n of 67821 (the total state sample size). The mean coefficient alpha under this model was 0.9186 with a standard error of 0.0000968 and a 95% confidence interval of (0.9185, 0.9188). Of course this standard error is based on the wrong n since there are only 512 data points, rather than 67821 (where SE is computed under the condition of simple random sampling, dividing the standard deviation by the square root of n). To adjust weights so the sum is equal to the original sample size (512), the weights were normalized by dividing each n by their sum (67821) and multiplying by 512. This resulted in a weighted mean of 0.9186 with a standard error of 0.00112 and a 95% confidence interval of (0.9165, 0.9208). Compared to previous methods for estimating the mean coefficient alpha, the n-weighted mean estimate yielded one of the widest confidence interval, a less precise estimate than the T-weighted mean alpha. In addition, the n-weighted distribution of coefficient alpha had the most severe skew, at –3.27.

We similarly computed the n-weighted mean alpha for the reading test, which was 0.8617, with a standard error of 0.00256 (computed employing the normalized sample size weights that sum to 515, the number of alpha coefficients). This resulted in a 95% confidence interval for coefficient alpha on the reading test of (0.8567, 0.8667). Similarly, the skew for the n-weighted coefficient alpha distribution was increased to –3.81.

The statewide reliability coefficient computed from all subjects simultaneously was 0.932 for the mathematics test and 0.890 for the reading test.

Evaluating Variation in Score Reliability

The results of the K-sample significance tests indicated statistically significant variation in the computed alpha coefficients among the 512 schools for the mathematics scores and the 515 schools for the reading scores: χ^2 (511) = 1882, $p<0.01$ and χ^2 (514) = 2795, $p<0.01$, respectively (Figure 5 and 6). These are essentially the results regarding the heterogeneity of effect hypothesis.

Several variables related to scores and coefficient alpha were correlated using Spearman rank order correlations (r_s) to identify potential variables associated with variability in coefficient alpha. These variables included the original coefficient alpha, T-transformed coefficient alpha, mean test score, observed test score variance, and an estimated ratio of error score variance to true score variance. Each of the 512 mathematics coefficients and 515 reading coefficients contributed to these correlations, as reported in Table 2. Other variables were also assessed with no significant relationship, including school size and proportion of female students.

As expected because of the monotonicity of the T-transformation, original alpha and T-transformed alphas were perfectly correlated. We found moderate correlations between alpha and score means (-0.438 for mathematics and –0.673 for reading), suggesting that schools yielding higher mean scores also yielded lower score reliabilities. This may in part be due to the strong relationship between score means and score variances (-0.746 for mathematics and –0.876 for reading). This suggests that schools yielding higher mean scores also yield lower score variances, which may in part be due to a ceiling effect (the tests are basic standards tests). The correlations between alpha and score variance were very strong (0.900 for mathematics and 0.926 for reading), and likely responsible for some of the correlation between alpha and mean scores. Finally, alpha is perfectly correlated with the ratio of error-score variance to true-score variance, as suggested by classical test theory.

In addition to examining variability in coefficient alpha, we noted the variability in each of the variance components, including observed-, error-, and true-score variance estimates (Table 3). Compared to the mean value of each variance component, there existed significant variation in variance components across schools, including error variance that is assumed to be constant in classical test theory models of measurement error and reliability (more on this later).

Explaining Variability in Score Reliability

We assessed the ability of a linear model to explain variation in coefficient alpha. First, to develop a model capable of explaining variation in T-transformed coefficient alpha, a weighted linear regression analysis was carried out. Only results for the mathematics test data are reported here.

The weighted linear regression analysis for the mathematics test data indicated that the ratio of the error variance to true variance accounted for 92.9% of the variation in the weighted T among the 512 schools, with no significant variation remaining in T. Employing the heterogeneity test notation (Q), Q_{total} (1882) = $Q_{\text{regression}}$ (1749) + Q_{error} (134), where Q_{error} was substantially less than $Q_{\text{critical}(df=510,\ \alpha=0.05)} = 459$. In contrast, score variance accounted for only 69.8% of variation in the weighted T, after which significant unexplained variation in T remained. These results varied slightly from the Spearman correlations reported in Table 2 because of the linearity assumption in the regression. The nonlinear nature of the relationship between alpha and the error- to true-score variance ratio can be seen in Figure 7.

Finally, we examined the relationship of various score components with the adjusted coefficient alpha (adjusted for heterogeneity in observed score variance). As reported earlier and observed in Figure 1, the distribution of adjusted alpha approaches normality. Recall that this adjustment was made under the classical test theory assumption of constant error variance. Also recall that we observed significant variation in error variance across schools. This variation calls into question the appropriateness of making such an adjustment to coefficient alpha. The relationship between adjusted alpha and error variance is perfectly linear, as can be seen in Figure 8 ($\underline{r}$ = 1.00). However, under the condition where error variance is not constant, the adjustment based on heterogeneity of observed score variance is no longer appropriate.

Given that the test scores are based on dichotomously scored items, an alternate model of measurement error may be appropriate, such as the binomial-error model (Lord, 1965). This model assumes that the observed score is the number of items correct for a test that consists of locally independent items with equal difficulty. Where items are not of equal difficulty, a compound binomial-error model can be used (Lord, 1965). Binomial-error models suggest that we should expect different standard errors of measurement at different levels of true scores. Although we did not compute binomial standard errors, we can observe a similar situation when examining the relationship between mean score and the standard error of measurement (Figure 9). What we are likely seeing in this figure is a portion of a curve that increases as mean scores move from zero to about 30 (not observed in these data) and decreases as mean scores improve from about 30 to a high of 62. Such a curve is expected under the binomial-error model. There are additional conditions under which the classical standard error of measurement may not be appropriately applied to each individual (under assumptions of constant error variance), particularly when there are ceiling or floor effects (Allen & Yen, 1979).

Summary

Much of our investigation here yielded results that were directly predicted from classical test theory. Although a great deal more was presented earlier in terms of analytical options for the meta-analyst (e.g., random effects modeling choices), there was little need for complicated analyses.

Coefficient alpha varied significantly across schools in both the mathematics and reading tests. In our attempts to characterize the mean coefficient alpha and its variability, we addressed issues related to the distributional properties of the original coefficients and subsequent transformed coefficients, accounting for both predicted covariation (adjusted for heterogeneity of scores) and sampling distribution properties (normalizing T-transformation). As expected, adjusted alpha yielded a smaller standard error than did unadjusted alpha and unweighted T; however, the smallest standard error was obtained from the weighted T synthesis. We also noted that a typical weighting scheme, sample-size weighting, resulted in even more severe skewing of the coefficient alpha distributions for both instruments employed in this investigation.

We empirically confirmed subsequent predictions of classical test theory in that the correlation between alpha and the ratio of error- to true-score variance was at unity and nonlinear. Similarly, the relationship between adjusted alpha and error variance was also at unity and linear.

Somewhat surprising, although also predicted by the binomial error model, was the significant variation in error variance across schools. This result was also found be Feldt and Qualls (1999) who subsequently recommended that school-level standard errors be reported in large-scale testing programs. We suggested that the nonconstant error variance creates a condition making score heterogeneity adjustments to coefficient alpha inappropriate.

Because of these findings we recommend that the meta-analyst interested in synthesizing coefficient alphas harvested from test scores based on dichotomously score items to employ the weighted T-transformation for coefficient alpha. It relies directly on the sampling distribution of coefficient alpha and avoids complications due to the inapplicability of the score heterogeneity adjustments and the severe skewing of sample-size weighting of coefficient alpha. However, because of the strong relationship between reliability and group heterogeneity, score variance must be employed in syntheses of coefficient alpha as a covariate, to eliminate that source of variance before examining other study characteristics that may or may not explain remaining variation.

We also strongly recommend that a test of heterogeneity of coefficients be employed prior to evaluating explanatory variables. Liberal or conservative interpretation of the results of such tests can be employed based on the researcher's own preferences. However, investigating potential moderators of variability in coefficient alpha seems out of place if variation is limited to start with and particularly so if most if not all variation can be accounted for by score variability. We have also shown how strongly score variation is related to the magnitude of coefficient alpha and this must not be overlooked in future syntheses. The fact that several RG studies have been published ignoring the impact of score variability leads us to question the meaningfulness of such studies.

This investigation has been conducted without questioning the meaning of synthesizing reliability coefficients across studies. The context in this study is slightly different than those found in typical RG studies in that the exact same test was administered to similarly aged individuals in the same state. In RG studies, it is typical to find instruments altered or modified and administered to a wide variety of individuals under various conditions. We can certainly conclude that reliability is not measurement error. Measurement error, or error variance, is one component that contributes to the magnitude of score reliability. Most directly, the ratio of error-score to true-score variance is the factor affecting the magnitude of reliability. Both of these variances are components of the observed score variance. Without accounting for these mathematical relationships, the meaning of a synthesized reliability is in question.

Questions about the nature of the relationships remain, including the implications of the binomial error model for dichotomously scored tests and the appropriateness of the constant error variance assumption for polytomously scored tests. Further analyses will also help us understand the true ability of each method (original coefficient, adjusted coefficient, sample-size weighted, and unweighted and weighted T-transformed alpha) to estimate the mean coefficient alpha and its standard error. Simulations may provide additional evidence as to the appropriateness of these methods.

References

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing.

Alsawalmeh, Y. M., & Feldt, L. S. (1992). Test of the hypothesis that the intraclass reliability coefficient is the same for two measurement procedures. Applied Psychological Measurement, 16(2), 195-205.

Alsawalmeh, Y. M., & Feldt, L. S. (1994). Testing the equality of two related intraclass reliability coefficients. Applied Psychological Measurement, 18(2), 183-190.

Alsawalmeh, Y. M., & Feldt, L. S. (1999). Testing the equality of independent alpha coefficients adjusted for test length. Educational and Psychological Measurement, 59(3), 373-383.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. Psychological Bulletin, 99(3), 388-399.

Cooper, H., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 3-14). 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. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 34(3), 363-373.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the sample for two tests administered to the same sample. Psychometrika, 45(1), 99-105.

Feldt, L. S. (1990). The sampling theory for the intraclass reliability coefficient. Applied Measurement in Education, 3(4), 361-367.

Feldt, L. S., & Qualls, A. L. (1999). Variability in reliability coefficients and the standard error of measurement from school district to district. Applied Measurement in Education, 12(4), 367-381.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. Applied Psychological Measurement, 11(1), 93-103.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Gulliksen, H. (1987). Theory of mental tests. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hakstian, A. R. & Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. Psychometrika, 41(2), 219-231.

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., & Thompson, B. (April, 2001). Characterizing measurement error in test scores across studies: A tutorial on conducting "reliability generalization" analyses. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

Kraemer, H. C. (1981). Extension of Feldt's approach to testing homogeneity of coefficients of reliability. Psychometrika, 46(1), 41-45.

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. (1965). A strong true-score theory, with applications. Psychometrika, 20, 239-270.

Lord, F. M. (1974). Variance stabilizing transformation of the stepped-up reliability coefficient. Journal of Educational Measurement, 11(1), 55-57.

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

Paulson, E. (1942). An approximate normalization of the analysis of variance distribution. Annals of Mathematical Statistics, 13, 233-235.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. Journal of Consumer Research, 21, 381-391.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 301-321). New York: Russell Sage Foundation.

Sawilowsky, S. S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. Educational and Psychological Measurement, 60(2), 157-173.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 261-281). New York: Russell Sage Foundation.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115-129.

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

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. Intelligence, 27(3), 183-198.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. Educational and Psychological Measurement, 60(2), 174-195.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58(1), 6-20.

Woodruff, D. J., & Feldt, L. S. (1986). Tests for equality of several alpha coefficients with their sample estimates are dependent. Psychometrika, 51(3), 393-413.

Table 1

Summary of Descriptive Statistics Employing Various Estimates of Coefficient Alpha

		95% Confidence Interval			
	Median	Lower Limit	Mean	Upper Limit	SE
Math (n=512)					
Alpha (unweighted)	0.9182	0.9036	0.9071	0.9105	0.001763
Alpha adjusted	0.9185	0.9163	0.9176	0.9189	0.0006603
Alpha (n-weighted)	0.9241	0.9165	0.9186	0.9208	0.001116
T (unweighted)	0.4342	0.4413	0.4460	0.4508	0.002410
T converted to alpha	0.9181	0.9084	0.9113	0.9141	0.001442
T (weighted)	0.4210	0.4224	0.4239	0.4255	0.0007798
T converted to alpha	0.9254	0.9230	0.9238	0.9246	0.0004205
Reading (n=515)					
Alpha (unweighted)	0.8696	0.8318	0.8401	0.8483	0.004194
Alpha adjusted	0.8736	0.8695	0.8721	0.8746	0.001303
Alpha (n-weighted)	0.8741	0.8567	0.8617	0.8667	0.002562
T (unweighted)	0.5071	0.5213	0.5286	0.5359	0.003711
T converted to alpha	0.8696	0.8461	0.8523	0.8583	0.003118
T (weighted)	0.4902	0.4981	0.4999	0.5017	0.0009252
T converted to alpha	0.8822	0.8737	0.8751	0.8764	0.0006934

Table 2

Intercorrelations Between Coefficient Alpha, Transformed Alpha, and School-Level Score Characteristics for Mathematics and Reading Tests Data

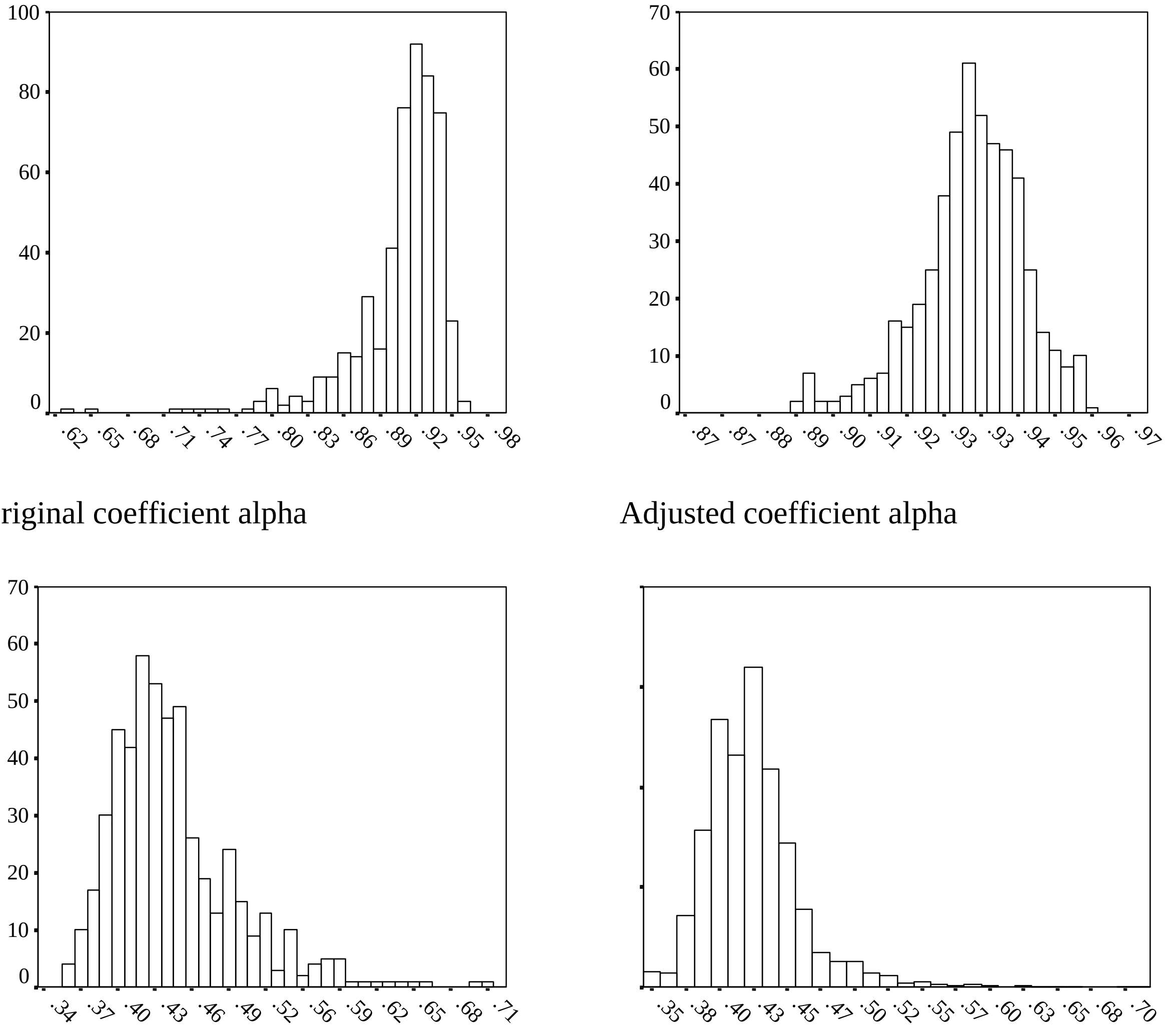
	1	2	3	4	5
1. Coefficient Alpha	--	-1.000	-0.438	0.900	-1.000
2. T (Transformed alpha)	-1.000	--	0.438	-0.900	1.000
3. Mean Score	-0.673	0.673	--	-0.746	0.438
4. Score Variance	0.926	-0.926	-0.876	--	-0.900
5. Ratio: $\frac{\sigma_E^2}{\sigma_T^2}$	-1.000	1.000	0.673	-0.926	--

Note. The upper triangle includes correlation for the mathematics test (n=512) and the lower triangle includes correlations for the reading test (n=515). All correlations are Spearman rank order correlations.

Table 3

Descriptive Statistics for Variance Components for the Mathematics and Reading Tests

	Observed Variance	Error Variance	True Variance
Mathematics			
Mean	105.076	8.517	96.559
Std. Deviation	39.907	1.545	38.875
Reading			
Mean	33.655	4.259	29.396
Std. Deviation	16.436	0.985	15.676



Original coefficient alpha

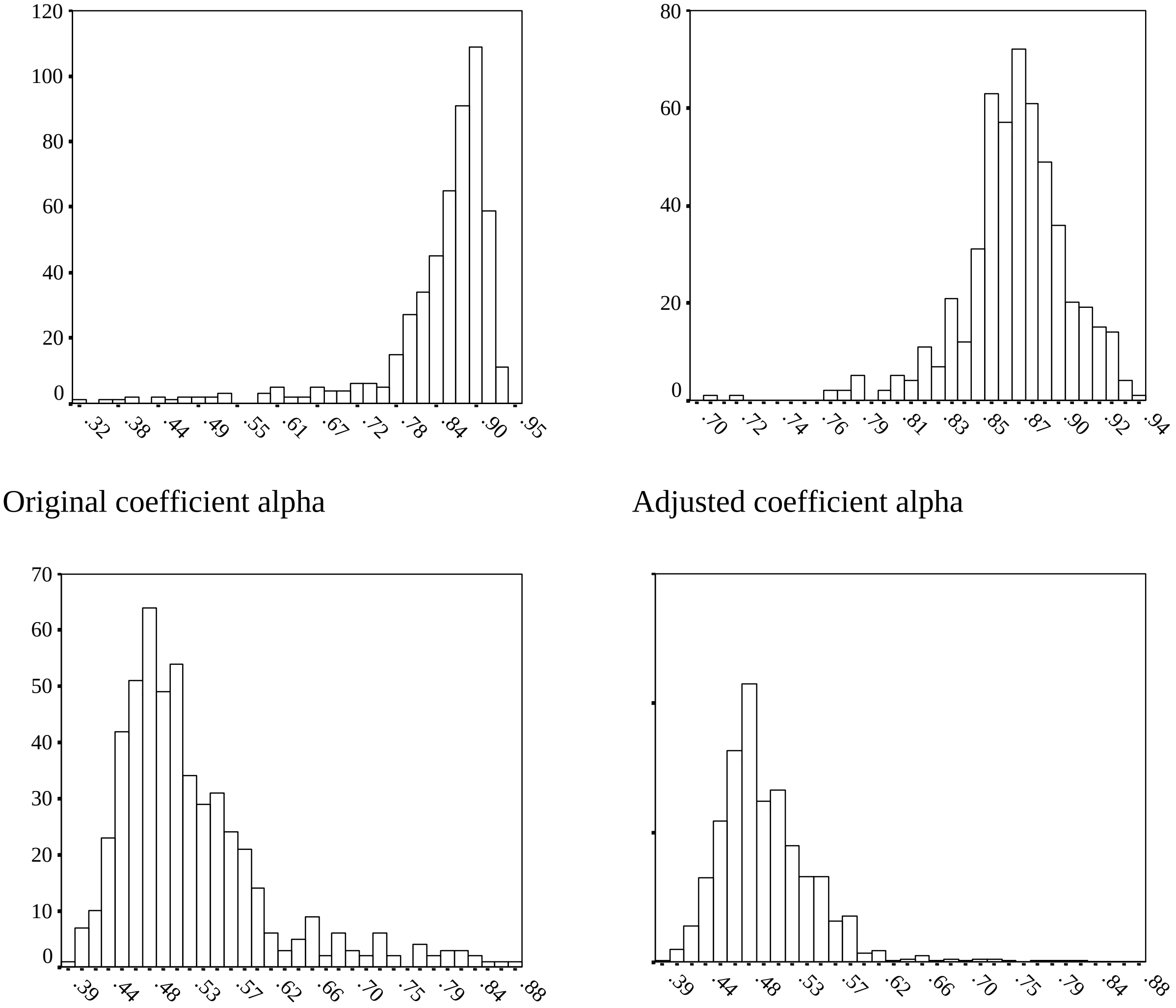
Adjusted coefficient alpha

T-transformed coefficient alpha

Weighted T-transformed coefficient alpha

Figure 1.

Distributions of coefficient alpha in original metric and under three transformations for mathematics test scores of 512 schools.



Original coefficient alpha

Adjusted coefficient alpha

T-transformed coefficient alpha

Weighted T-transformed coefficient alpha

Figure 2.

Distributions of coefficient alpha in original metric and under three transformations for reading test scores of 512 schools.

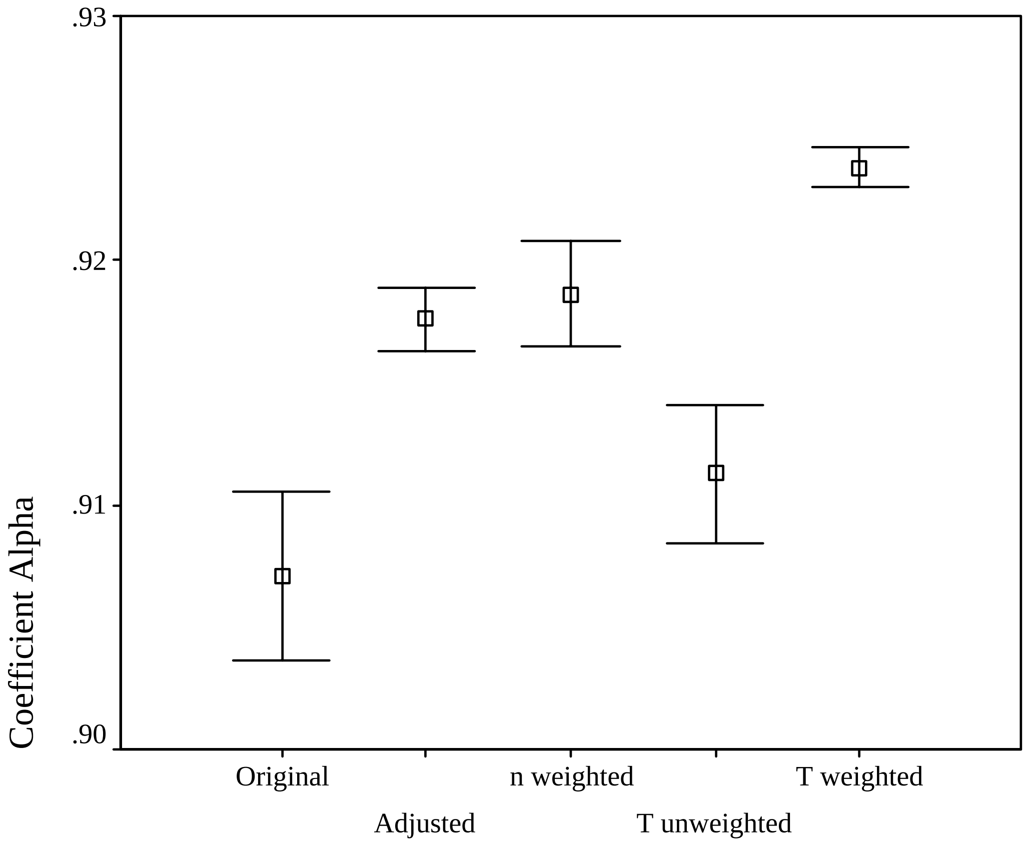


Figure 3.

Means and 95% confidence intervals for estimates of mean coefficient alpha computed under five methods for the mathematics test.

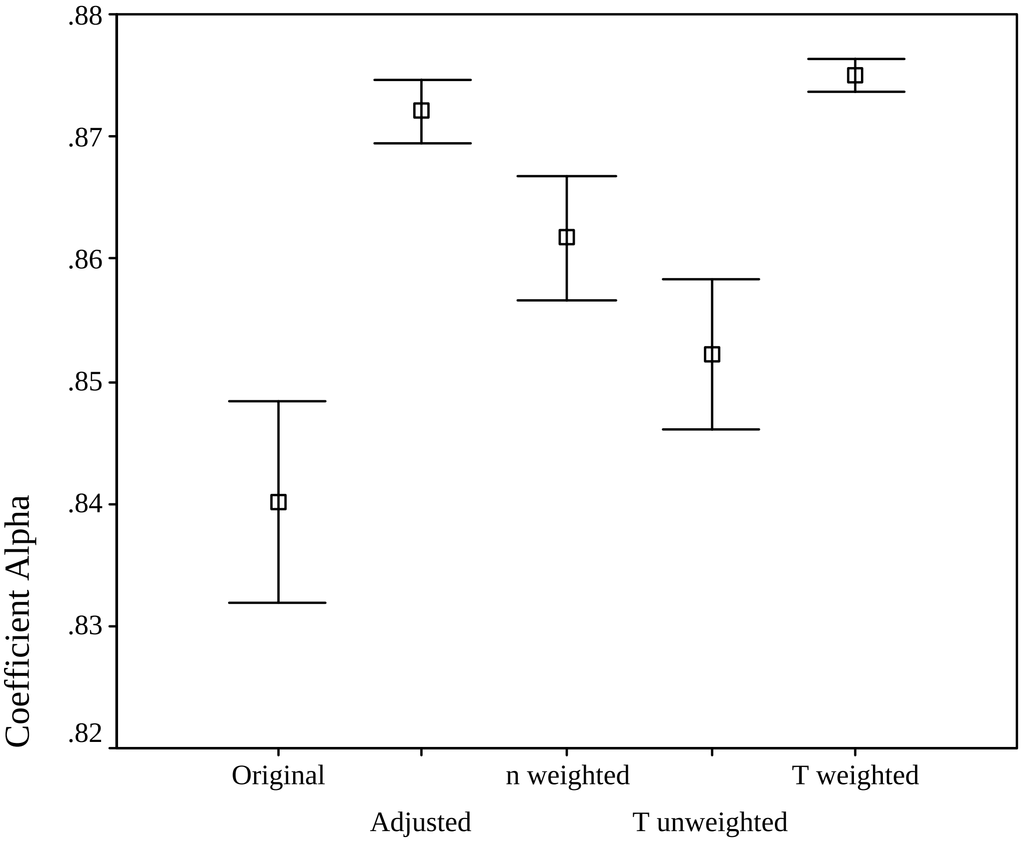


Figure 4.

Means and 95% confidence intervals for estimates of mean coefficient alpha computed under five methods for the reading test.

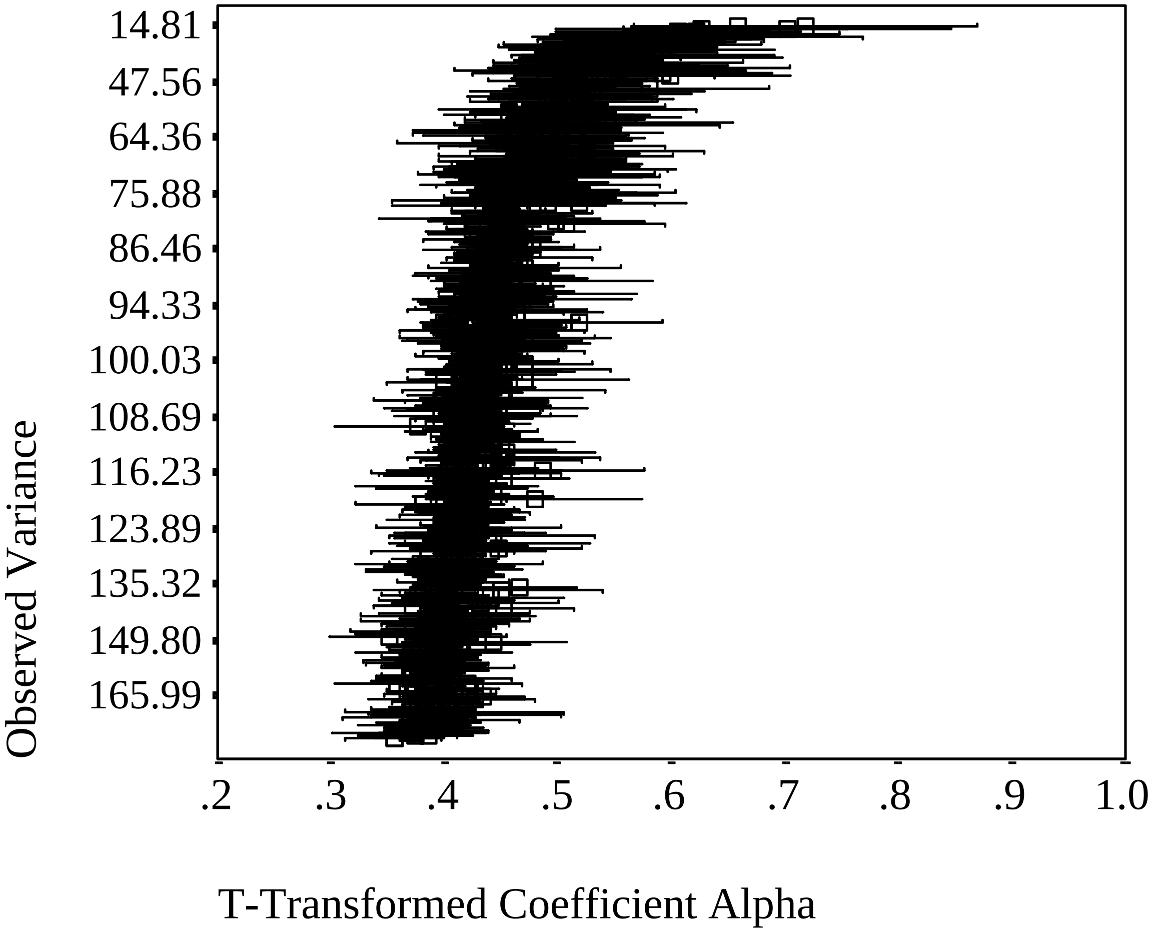


Figure 5.

Mathematics test T-transformed coefficient alphas and their 95% confidence interval based on the variance estimate for each T, ordered by observed variance.

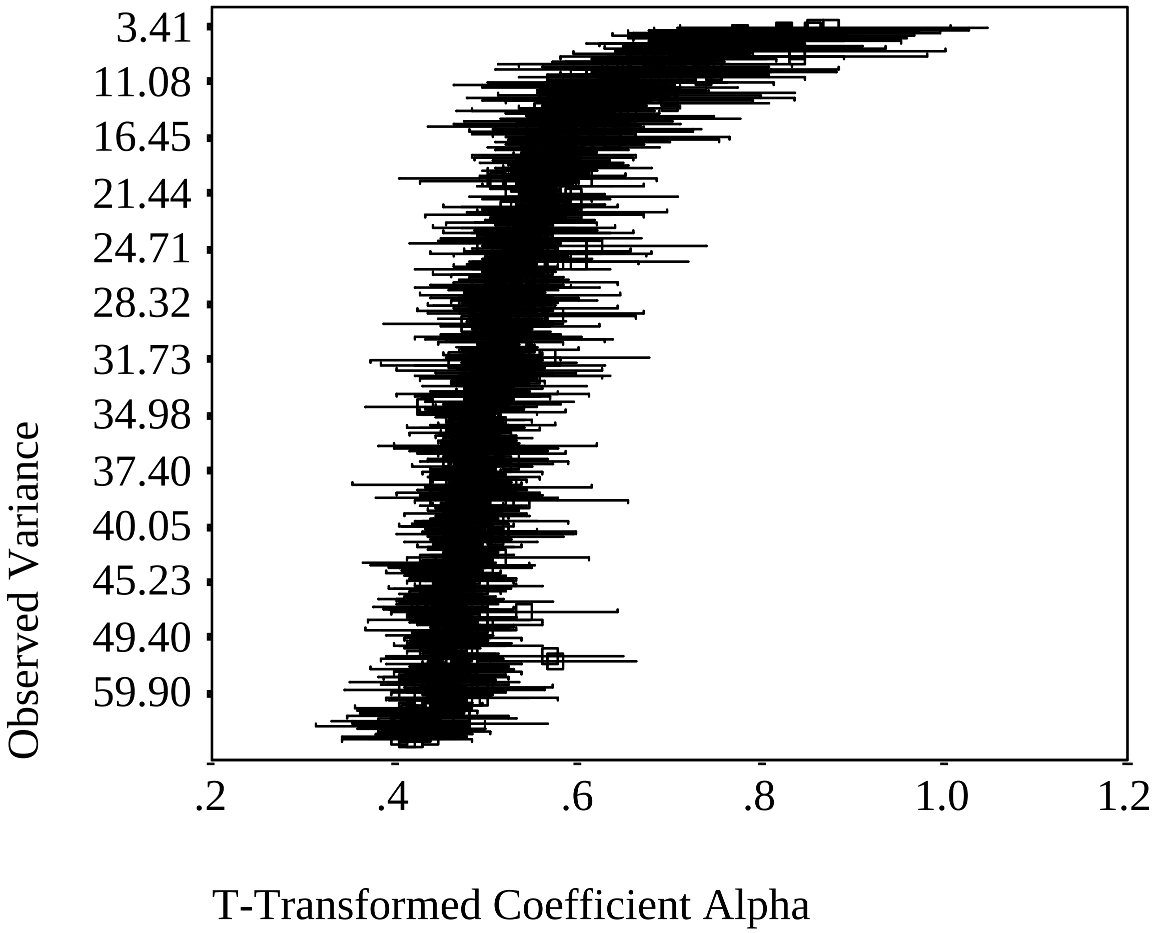


Figure 6.

Reading test T-transformed coefficient alphas and their 95% confidence interval based on the variance estimate for each T, ordered by observed variance.

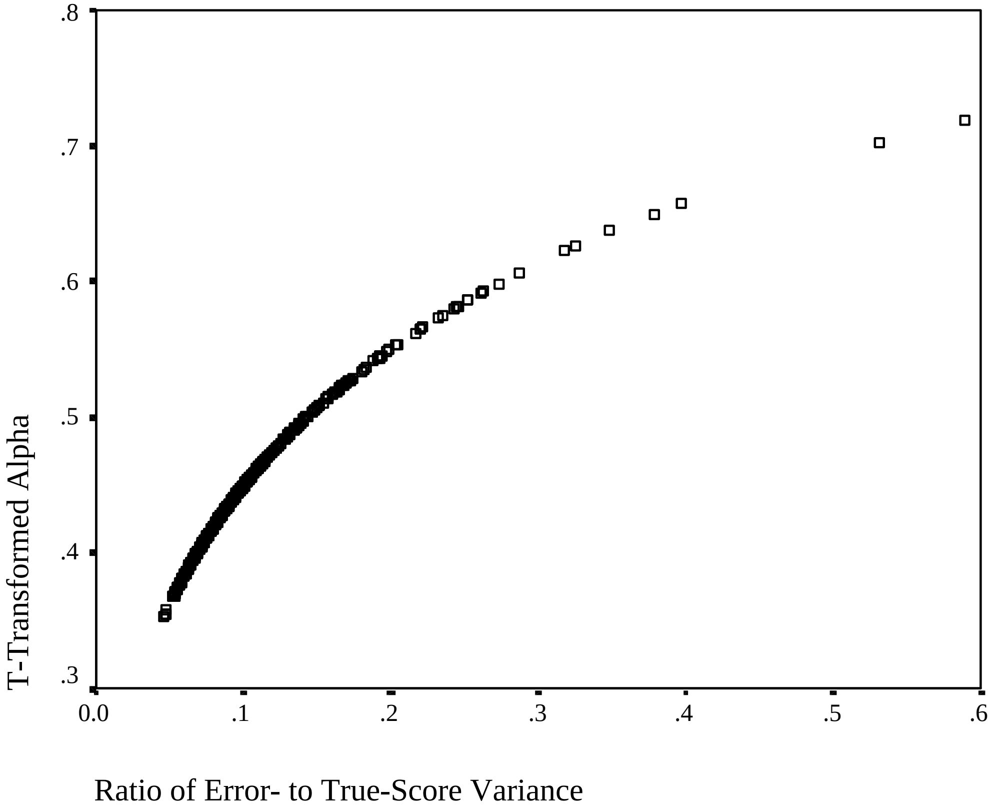
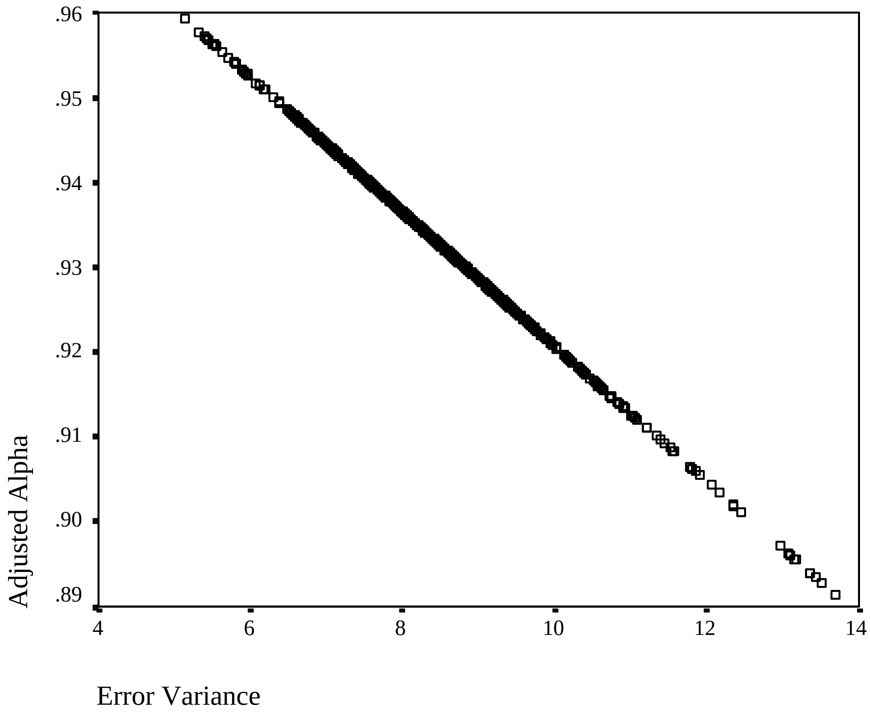


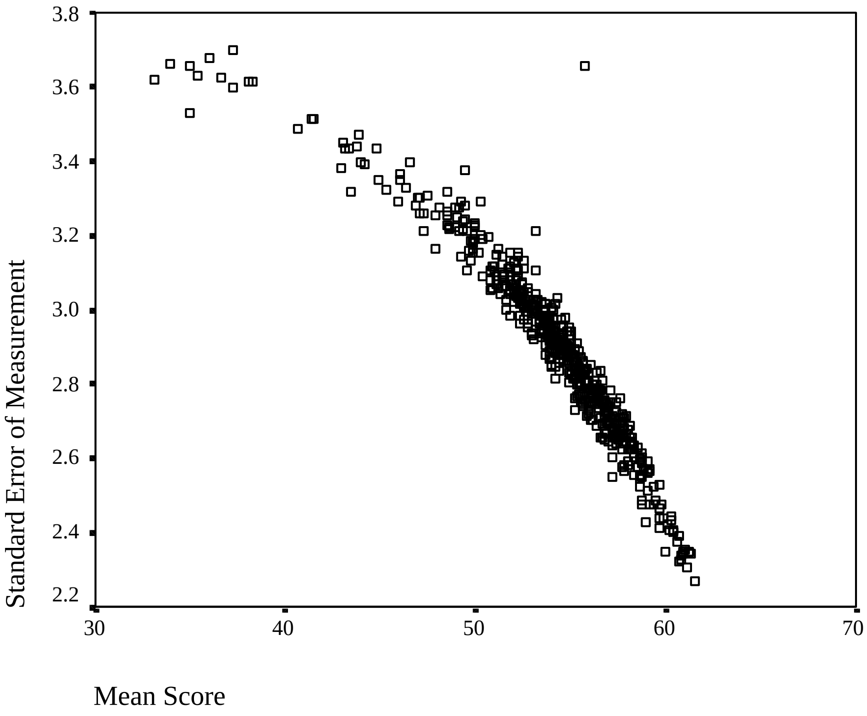
Figure 7.

Relationship between T-transformed alpha and the ratio of error-score to true-score variance for mathematics test scores of 512 schools.



<u>Figure 8.</u>

The relationship between adjusted coefficient alpha and error variance for mathematics test scores of 512 schools.



<u>Figure 9.</u>

Relationship between mean score and standard error of measurement for mathematics test scores of 512 schools.