Psychometric Inferences From a Meta-Analysis of Reliability and Internal Consistency Coefficients

Juan Botella, Manuel Suero, and Hilda Gambara
Autonomous University of Madrid

A meta-analysis of the reliability of the scores from a given test, also called reliability generalization, allows the quantitative synthesis of its properties from a set of studies. It is usually assumed that part of the variation in the reliability coefficients is due to some unknown and implicit mechanism that restricts and biases the selection of participants in the studies’ samples. Sometimes this variation has been reduced by adjusting the coefficients by a formula associated with range restrictions. We propose a framework in which that variation is included (instead of adjusted) in the models intended to explain the variability and in which parallel analyses of the studies’ means and variances are performed. Furthermore, the analysis of the residuals enables inferences to be made about the nature of the variability accounted for by moderator variables. The meta-analysis of the 3 studies’ statistics—reliability coefficient, mean, and variance—allows psychometric inferences about the test scores. A numerical example illustrates the proposed framework.

Keywords: reliability, reliability generalization, meta-analysis

Meta-analysis of the reliability of scores from a given test has grown rapidly in the last decade under the label reliability generalization (Thompson, 2003; Thompson & Vacha-Haase, 2000; Vacha-Haase, 1998). However, although it is already discussed in general books on psychometrics (e.g., Brennan, 2006), there is not yet a generally accepted set of procedures for doing it. The main issues under debate have to do with the choice of the effect size index to represent the characteristic under study (reliability coefficient, reliability index, standard error of measurement) and with the decisions related to weighting and normalizing transformations of the estimates. Furthermore, one of the obstacles to establishing a consistent method for reliability generalization has been that sampling variances often show large fluctuations; this has a well-known influence on reliability. In the present article we propose a framework for dealing with the problem of the variations in reliability that are due to fluctuations in sample variances.

Test reliability can be studied with multiple procedures and by assuming multiple perspectives and sources of error in measurement. Awareness of such variety allows researchers to reach more judicious conclusions (Dimitrov, 2002). Currently it is accepted that coefficients representing different facets of reliability, such as stability or test–retest consistency, internal consistency, or equivalence should not be mixed (Sawilowsky, 2000), given that they estimate different sources of error variance. It is also easy to make statements about the information that ideally should be included in any published application of a test, to maximize the benefits of any later meta-analytic integration (Dimitrov, 2002). Despite multiple suggestions (e.g., Thompson, 1994; Wilkinson & Task Force on Statistical Inference, 1999), many authors are not aware of the importance of reporting the empirical reliability (at least as internal consistency, for one spot designs) of the scores in their studies. Rather, many authors do what has been called a “reliability induction” from the value reported in the manual of the test (Vacha-Haase, Kogan, & Thompson, 2000). However, it is also true that for many years to come, the bulk of the studies that will be available for meta-analysis are those already published today, no matter how limited they are. In the present article we focus on the meta-analysis of test–retest reliability coefficients (\(r_{xx}\)) and of internal consistency (alpha coefficient; Cronbach, 1951), the two measures of reliability that are most often employed (Hogan, Benjamin, & Brezinsky, 2000).

The goals in the meta-analysis of reliability usually involve the same three general issues as any other form of meta-analysis: (a) the combined estimation issue, (b) the homogeneity issue, and (c) the explanation of the variability issue. These three issues must be addressed in the building of a good statistical model for the empirical estimates collected from the literature, when applied to score reliabilities obtained with a specific test.

We focus here on the third issue. After obtaining a significant result in a homogeneity test, we studied the adjustment to models based on moderator variables included in the meta-analysis. After obtaining a significant fit, we reached conclusions of the following type: “the reliability is associated with the categorical variable \(x\)” or “the reliability increases (or decreases) with the quantitative variable \(x\).”

However, we believe that the meta-analytical procedures allow us to go one step further in inferences related to scores’ reliability. Specifically, something can be said about the nature of the source
of the variation observed if we study the relationship between reliability coefficients and the samples’ means and variances, instead of analyzing only the variability of the reliability coefficients.

**Terminology**

Our theoretical framework is the classic test theory (CTT), which assumes that the observed score equals the true score plus some random error \((x = \overline{x} + \varepsilon)\), plus a well-known set of assumptions (Lord & Novick, 1968).

Suppose there exists a covariance matrix for a test composed of \(j\) items in which the total score equals the simple sum of the scores in the \(j\) items \((x = \Sigma I_i)\). Then, \(\Sigma \sigma_i^2\) represents the sum of diagonal elements (sum of the item variances) and \(\Sigma \sigma(x,x_m)\) the sum of all the elements external to the diagonal [sum of the \(j \times (j - 1)\) covariances]. We will also employ the expression \(\Sigma \sigma(x,T_{T_m})\) for the sum of the covariances between the true scores of the items. As can be demonstrated from CTT assumptions, the covariance between the observed scores of two items equals the covariance between their true scores, \(\sigma(x,x_m) = \sigma(T,T_m)\), for any pair of items \(i, m\).

The variance in the test equals the sum of the variances in the items, plus the covariances between them,

\[
\sigma^2_x = \Sigma \sigma_i^2 + \Sigma \sigma(x,x_m).
\]

Of course, the expected value in the test equals the true score, which equals the sum of the true scores in the items,

\[
E(x) = \sum_{i=1}^{j} I_i = \sum_{i=1}^{j} T_i.
\]

Also from CTT, the variance of each item equals the variance of the true scores plus that of the errors, and the covariance between the observed scores from any two items equals the covariance between their true scores, \(\sigma(x,x_m) = \sigma(T,T_m)\), so we have

\[
\sigma^2_i = \Sigma \sigma_i^2 + \Sigma \sigma_{12}^2 + \Sigma \sigma(T,T_m).
\]  
(1)

The reliability is defined (Feldt & Brennan, 1989; Lord & Novick, 1968) as

\[
\rho_{xx} = \sigma^2_x/\sigma^2_x,
\]

and can be expressed in the same terms as employed in Equation 1,

\[
\rho_{xx} = \Sigma \sigma_i^2 + \Sigma \sigma(T,T_m)/\Sigma \sigma_i^2 + \Sigma \sigma(T,T_m) + \sigma_{xx}^2.
\]  
(2)

In the same vein, the internal consistency as expressed in the Cronbach’s alpha coefficient is defined as,

\[
\alpha = \left(\frac{j}{j-1}\right) \left(\frac{\Sigma \sigma(x,x_m)/\sigma^2_x}{\Sigma \sigma_i^2 + \Sigma \sigma(T,T_m) + \sigma_{xx}^2}\right).
\]  
(3)

### Three Sources of Variation

Our proposed framework recognizes three sources of systematic variation of the reliability coefficients beyond the variability generated by mere random sampling; of course, random sampling yields variations both in the samples’ variances and their reliabilities. However, if all the variation in those statistics is produced by truly random sampling, it should be within the range expected for random noise, and the null hypothesis will survive a test of homogeneity.

On the other hand, if the variation is unexpectedly large, then it is worth trying to account for it, realizing there are at least three sources of variability: (a) biased sampling of true scores, (b) variations in the error variance, and (c) variations in the correlations between the items. Except when indicated, the assumptions of CTT will be generally assumed.

### Sampling Schemes of the True Scores

It is easy to generate situations that differ from an ideal scene in which the participants are selected from the population by a strictly random procedure. Generally speaking, the only case in which a test is applied to a sample specifically designed to represent a general population is the process of test building and norms elaboration that finally makes up the test manual. On the contrary, in any study in which the test is employed as a tool for assessment or research, the degree to which the sample represents the general population is simply ignored. The procedure for recruiting the participants is limited by the needs and goals of each specific study. As a consequence, we must expect a large heterogeneity of the composition and variability of the samples to which the test is applied, and in general, they can be very different from the general population (Dawis, 1987; Vacha-Haase et al., 2000). Any procedure for recruiting the participants for a study produces biases, truncations, or partial censures of different magnitude of the true scores, as related to the distribution of the general population. Those biases, or sampling schemes, have important effects in the reliability estimation for the scores from each study.

However, if there is no other source of perturbation, the variations in reliability can be predicted from the variances generated by the mechanism of bias (B), following the equation,

\[
\rho_{ab} = 1 - \frac{\sigma^2_x}{\sigma^2_x(1 - \rho_{ab})},
\]  
(4)

where \(\rho_{ab}\) and \(\sigma^2_x\) are, respectively, the reliability and the variance of the scores in the general population, whereas \(\rho_{ab}\) and \(\sigma^2_x\) are the same for the population of scores defined under B, the mechanism of bias involved. Figure 1 shows the function that relates the population variance under different schemes of biased sampling and the expected reliability, according to Equation 4, for a case in which the reliability is 0.90 for a population variance of 100. As we explain below, the line depicted in Figure 1 is where we expect to find the pairs of variance/reliability values obtained under varied sampling schemes. Zones I and II are those where we expect to find the pairs of values under different types of departures from the model of measurement. Thus, the conditions with an increased error variance will yield pairs in Zone I, as described in the next section.
For example, suppose that in performing a meta-analysis we use the following three categories to classify the set of studies in which a test for measuring trait anxiety has been applied: (a) normal participants, volunteer students; (b) medical outpatients; and (c) individuals receiving a psychological treatment. Suppose we find that the variances of the observed scores increase from the first to the second group and from the second to the third group. The explanation can be in the nature of the sample recruitment mechanism. The participants in the first group are more homogeneous, as almost all have relatively low scores. The medical patients include a broader range of scores, given that at least some should have higher levels of anxiety. Finally, the group under psychological treatment involves some even more extreme scores and many scores in the middle but still includes individuals with low scores, given that not all the participants have anxiety-related problems. Assuming that all other elements are equal (especially the standard errors of measurement), this increase in the sample variances from Categories 1 through 3 is accompanied by a parallel increment in the reliability coefficients. This positive relationship between the variances and the reliabilities comes from the fact that the variability in the variances is due exclusively to the presence of heterogeneous sampling schemes.

Sometimes the biases are easily identifiable, but in other cases they remain hidden in the sampling fluctuations. In general, the presence of different biases in the true scores of the recruited participants generates reliability and variance pairs that are closely related. The relationship between those statistics \( r_{xx} \) and \( \alpha \) is positive (the same happens if the reliability is assessed by internal consistency, as with Cronbach’s alpha).

As illustrated in the anxiety test example above, the selective biases can be associated with a categorical moderator variable, but they can also be associated with a quantitative variable, as in the example in the next paragraph. In those cases part of the variation (in the sample variances and reliability coefficients, simultaneously) can be explained by that moderator variable.

Consider an example with a quantitative variable. Suppose that the population is in fact composed of a mixture of several subpopulations (or in the simplest case, only two subpopulations—e.g., men and women), with different means and variances in the scores: \( \mu_M, \mu_F, \sigma^2_M \) and \( \sigma^2_F \). If the other elements in the psychometric model are the same for the two subpopulations (error variance, homogeneity of the items), the mean and the variance of the sample scores are a function of the percentage of contribution to the sample of those subpopulations. For a sample composed only of men, the expected values for the sample’s mean and variance are \( \mu_M \) and \( \sigma^2_M \) (\( \mu_F \) and \( \sigma^2_F \) for women-only samples). However, when the percentages of men and women in the sample vary, the expected means and variances are different. Specifically, if we test a model with the percentage of women (or men) as the moderator variable, then the sample means must show a linear relationship with it, whereas the sample variances must show a quadratic relationship (see the Appendix). In short, it is possible to identify the source of perturbation if the mentioned relationships are observed between the percentage of women and the averages and variances of the samples’ scores. Furthermore, the fluctuations in the sample variances associated with the different percentage compositions of men and women must be accompanied by concomitant fluctuations in the reliability coefficients.

This source of variation of the reliability coefficients is sometimes considered a problem for reliability generalization. On the contrary, we believe that it should not be eliminated or adjusted (see Rodriguez & Maeda, 2006) but rather incorporated into the model while an attempt is made to identify the operations that yield differential sampling schemes. The reason is that the adjustments can eliminate interesting variability that we are trying to explain; we will return to this point in the discussion.

**Variance of the Measurement Errors**

Equations 1, 2, and 3 show what will happen if a moderator variable has any influence on the error variance.\(^1\) In those cases the moderator variable can be considered to be a proxy measure of the error variance. The moderator can be some personal characteristic of the participants but also other types of variables, such as contextual or situational variables. An example of the first type is if the moderator variable is the type of diagnosis in the sample. Some pathologies involve memory impairments that reduce the reliability of memories on which self-report is based or increase distractibility that can reduce the attention paid and/or the care taken while answering the questionnaire. An example of a contextual variable can be whether the test is administered in an individual or group setting, because in collective administrations people may answer less accurately. An example of a situational variable can be the testing location, as the physical characteristics of the environment can contribute to higher or lower performance.

The examples above illustrate cases for which it is likely that the moderator variable has an influence on the measurement errors, thereby increasing the error variance, whereas the other elements in Equations 1, 2, and 3 remain the same.

Equation 1 follows from the CTT assumption that the error variance is independent of the true score, so that if a moderator variable is associated with increments in the error variance, then it must also be associated with increments in the variance of the observed scores. As to the reliability, any moderator variable

---

\(^1\) Here the door is open for relaxing the CTT’s assumption of constant error variance for any true value but holding the remaining assumptions.
related to increases in the variance of errors will necessarily be related to reductions in the coefficient of reliability (Equation 2) and the internal consistency (Equation 3). Notice that the direction of this relationship is the opposite of that described in the previous section. Here there is a negative relationship between the variances and the reliabilities. This is because here the source of the heterogeneity in the variances is the variability of the error variances. Nevertheless, as errors are still independent of the true scores, the averages of the observed scores must remain unchanged.

A consequence of the above is that the studies in which the variance of error is different will not pertain to the general function relating the reliability to the variance. Suppose that the studies done under some specific circumstances imply an increase in the error variance. Those studies would show reliability coefficients that are in general smaller than expected from their sample variances. In other words, they should show a negative residual, as the increment in the error variance will increase the variance of the observed scores and will reduce the reliability. For a given variance, \( \rho \), will be lower in conditions of increased error variance. When the data are adjusted to a model that does not distinguish between the two conditions, the residuals of the studies under conditions of increased error variance will tend to be negative. Zone I of Figure 1 shows where the points representing those increased-error-of-measurement studies will lie.

In short, increases in the error variance of measurement imply increases in the variance of the observed scores and decreases in the coefficient of reliability. If a moderator variable, \( x \), is systematically associated in one direction with the observed variance and the opposite direction with reliability (either \( r_{xx} \) or \( \alpha \)), whereas the mean scores do not change, then a credible hypothesis is that some mechanism associated with \( x \) increases the error of measurement when the test is administered.

Correlations Between the Items

The correlations between the items, as a source of variations in reliability, are related to the possibility that a variable, \( x \), is associated with a systematic modification of the correlations between the true scores in the items; the moderator variable would be a proxy measure of the degree to which the items correspond to the definition of tau-equivalence (Lord & Novick, 1968).

A single set of items can be more homogeneous in one subpopulation when the items better represent the construct for the first subpopulation than for the second one; that is, when the items reflect better the nucleus of the construct for the first one than for the second one.

The general population can be disaggregated into subpopulations for which the specific items forming the test reflect somewhat different constructs. It is possible that in a subpopulation the construct is simpler, and the items correlate more with the conceptual nucleus. On the contrary, in other subpopulations, for which the construct is more complex, the items are farther from the conditions required for two items being parallel (the correlations among them are lower). In this case, the construct is not identical in both subpopulations, as it is differentially sampled by a single test. For example, in adaptations and translations of tests, problems sometimes arise with the similarity of the constructs in different languages, cultures, or even cohorts (Hambleton, Merenda, & Spielberger, 2005). Differential homogeneity of the items is a well-known difficulty in cross-cultural assessments. This problem can arise when there is an unequal correspondence between the domain of observations from several subpopulations (languages, cultures) and the universe of generalization (van de Vijver & Poortinga, 2005). This can show up, for example, in constructs such as quality of life or even anxiety. Beyond the difficulties of a verbatim translation, the meaning conveyed by each item can differ between different communities. Although the same patterns of correlations between the items should be observed for the adaptation of the test to be acceptable, we should not expect them to be identical. The variations in the correlations can explain some differences in the reliability coefficients.

From CTT the degree to which two items are tau-equivalent is reflected by the degree to which the true scores correlate (but see Boyle, 1991). As a consequence, in the terms employed here, the key is in the covariances between the true scores of the items, \( \sigma(T_iT_j) \). Going back to Equations 1, 2, and 3, we can see that if a moderator variable conveys increments in the correlations between the true scores, then it also implies that (a) increments should occur in the observed variance; and (b) increments should also occur in the reliability (either \( r_{xx} \) or \( \alpha \)), given that the variance of the true scores (\( \sigma_T^2 \)) is also increased, whereas that of the errors (\( \sigma_e^2 \)) remains constant.

That is, for a given observed variance, \( \rho \), will be larger in conditions with higher correlations among the items. Let us suppose that an adjusted model does not distinguish between two categories of studies, one with higher correlations between the items than those from the other category. The consequence will be that the studies with increased correlations between the items will tend to show positive residuals. Zone II of Figure 1 shows where the points will lie that represent those increased-correlation studies (or, in relative terms, reduced-error-of-measurement studies).

In short, increments in the correlations between the true scores of the items imply increments in the variance of the observed scores and increments in the coefficient of reliability. As in the first case described above (varied sampling schemes, whereas the measurement error variance and the correlations between the items remain constant), a positive relationship between the variances and the reliabilities should be observed. The difference is that here the samples means should remain homogeneous. If a moderator variable, \( x \), has associated systematic and direct changes in the observed variance and reliability (either \( r_{xx} \) or \( \alpha \)), whereas the mean does not change, then a credible hypothesis is that some mechanism associated with \( x \) conveys a shift toward the definition of tau-equivalent items.

A Framework for Analysis

A consequence of the preceding arguments is that a more detailed analysis of the psychometric properties of the scores from a test is possible if the meta-analysis of a set of reliability estimates is accompanied by similar analyses of the variances and means of the samples’ scores. Sometimes significant levels of heterogeneity are found in a set of estimates. We have identified three different scenarios that, jointly or separately, can account for that heterogeneity. We have also pointed out procedures that can allow distinctions between them.

It is important to highlight at this point the complex relationships between the variances of the observed scores in the samples and the corresponding reliability coefficients. As reflected in Equations 1, 2,
and 3, and on the basis of the arguments in the previous section, that relationship can be direct or inverse. That is, when the variability in the variances is due to changes in the measurement errors, the relationship is inverse (large variances will be associated with smaller reliability coefficients). However, when the variability of the variances is due to differential sampling schemes or to changes in the correlations between the items, the relationship is direct (larger variances are associated with larger reliability coefficients).

In the next section we describe how to complement the meta-analysis of the reliability or internal consistency coefficients with meta-analysis of the samples’ means and variances. Then we describe an example in which this framework is employed to diagnose the source(s) of perturbation in the set of reliability coefficients synthesized.

**Meta-Analysis of Means and Variances**

To apply the above framework, it is necessary to complement the analysis of the reliability coefficients with parallel analysis of the means and variances. The meta-analysis of the means and variances of the scores allows more judicious interpretations of the source of the fluctuations in the reliability coefficients. Specifically, the presence of significant heterogeneity in those two sample statistics must be assessed, and the relationships between them and also with the reliability must be studied.

Before explaining how to do that, a few words about weighting are in order. Traditionally, the studies included in a meta-analysis are weighted in some way (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper & Hedges, 1994; Lipsey & Wilson, 2001). There is some controversy as to whether the reliability coefficients should be weighted, and several simulations have tried to answer the question of which is the most appropriate weighting scheme (e.g., Feldt & Charter, 2006; Mason, Allam, & Brannick, 2007). In the present article we have always weighted by the inverse of the variance of the estimate, as it minimizes the variance of the joint estimate (Cochran, 1954) and is the most common weighting method in meta-analysis (Rodriguez & Maeda, 2006). Nevertheless, the procedures described in this article are not tied to any weighting scheme or even to weighting itself.

The meta-analysis of the means can be done directly by employing the sample averages as the effect size; the variance of the mean is

\[ \sigma^2(\bar{x}) = \sigma_x^2/N, \]

where \( \sigma_x^2 \) is the population variance of the scores. Substituting \( \sigma_x^2 \) with the sample variance yields a reasonable approximation, so that employing the weights \( w_i = 1/\sigma_x^2 \) we can apply the more usual procedures in meta-analysis (Hedges & Olkin, 1985).

The case for the variance is more complex. Suppose that the scores in a test, \( x \), are distributed \( N(\mu, \sigma^2) \). Then the sample variances obtained from \( N \) observations, \( S^2 = \sum(x_i - \bar{x})^2/N - 1 \), have as expected value and variance,

\[ E(S^2) = \sigma^2 \quad \text{and} \quad \sigma^2(S^2) = (\sigma^2 \cdot \frac{2(N-1)}{N})^2 = 2 \cdot \sigma^4/(N-1), \]

and although its distribution is asymmetrical (chi-square), it approximates normality with large samples (\( N \geq 100 \)). Again, substituting \( \sigma_x^2 \) by the sample variance we have a reasonable approximation, so that by employing the weights,

\[ w_i = 1/(\sigma_x^2(N-1)/2 \cdot (S_i^2)^2), \]

we can also apply the more usual procedures in meta-analysis. As a way to reinforce the validity of the conclusions about the homogeneity of the variances, a parallel analysis can be done that applies Bartlett’s test to the observed variances (Bartlett, 1937).

**Normalizing the Measures and Adjusting the Model**

Before adjusting the model, some normalizing transformations are often recommended. Neither \( r_{xx} \) nor \( \alpha \) are normally distributed. Instead of analyzing the reliability coefficients, \( r_{xx} \), directly, the well-known transformation to Fisher’s \( z \) is applied, \( z_r = 1 \cdot \log(1 + r)/(1 - r) \). Many authors recommend Fisher’s transformation although some potential biases can also arise (Schulze, 2004). Something similar happens with the alpha coefficients, for which the following normalizing transformation can be applied (Hakstian & Whalen, 1976; Rodriguez & Maeda, 2006):

\[ T_i = (1 - \alpha_i)^{-1}, \]

where \( T_i \) stands for the transformation of the \( i \)th estimate and \( \alpha_i \) is the corresponding internal consistency coefficient.

It is commonly accepted that in doing a meta-analysis the statistics employed must be weighted in some way, with the inverse of the variance being the most popular choice. In this way the weight of each study is larger as its sample size increases. In our case, the variances of the transformations above are, respectively,

\[ \text{var}(Zr_i) = 1/(N_i - 3) \]

and

\[ \text{var}(T_i) = \frac{18 \cdot j \cdot (N_i - 1) \cdot (1 - \alpha_i)^{2/3}}{(j - 1) \cdot (9 \cdot N_i - 11)} \]

where \( N_i \) is the number of participants in the \( i \)th study, \( j \) is the number of items in the test, and the other symbols are defined in the same way as above.

Related to the normality assumption for the meta-analysis of means and variances, there is no problem with the means, given that the normality is guaranteed with samples of moderate size (\( N \geq 30 \)). On the contrary, we have already mentioned that the sample variances have an asymmetrical distribution that only approximates normality with relatively large samples (\( N \geq 100 \)).

The steps, more or less ordered, for a meta-analysis of reliability or internal consistency, from our proposed framework, include the following: (a) Study the homogeneity of the averages and variances.

---

\(^3\) Sometimes the analysis is restricted to the subset of samples reaching that size.

\(^2\) This approximation is strongly conditioned by the normality assumption.

\(^4\) Not all authors agree that the transformation is necessary, but neither is there any cue that it is prejudicial.
ances of the samples test scores; (b) if there is a significant heterogeneity, look for any moderator variable that can explain it; (c) adjust a regression weighted model that includes the sample variance as a predictor of the reliability; and (d) study the residuals and look for moderator variables that allow a significant increase in the explained variance with a simpler model that only includes the sample variances. In the next section we illustrate this scheme with a practical example.

Example With the State–Trait Anxiety Inventory (STAI)

The trait version of the well-known STAI of Spielberger, Gorsuch, and Lushene (1970) is a self-report of 20 items. The individuals assess the frequency with which they usually experience emotions, have thoughts, or perform behaviors that reflect high levels of anxiety.

We have collected data on three values—the alpha coefficients plus the means and variances—from 29 administrations of the test. In this example we conduct several analyses to try and understand what variables are related to these three values. They are in four different languages (Spanish, English, French, and Dutch), a feature that suggests a potential source of variation in the internal consistency of the applications. The main statistics are shown in Table 1.

The models adjusted throughout this article are random effects models. Compared to the fixed effects models, the random effects models do not assume a single population effect size. Rather, the random effects models assume that although the studies included in a meta-analysis have enough in common that it makes sense to synthesize the information, there is no reason to assume that the true effect size is identical in all studies (Hedges & Vevea, 1998). The generalizability of the conclusions of a meta-analysis is higher with random effects models than with fixed effects models.

The alpha coefficients were transformed to $T$ by Equation 5. On the basis of the 29 $T$ transformations included in our meta-analysis, the combined weighted estimate is 0.898 ($\pm 0.010$) with a random effects model (Hedges & Vevea, 1998).

The homogeneity of the estimates has been tested by the $Q$ statistic (Hedges & Olkin, 1985), applied to the $T$ values, and weighted again by the inverse of the variance,

$$Q = \sum w_i \cdot (T_i - \bar{T})^2,$$

where $w_i$ and $T_i$ have been defined above, and $\bar{T}$ is the combined estimate (0.898 in the example). The $Q$ statistic is distributed approximately as a chi-square with $k - 1$ degrees of freedom, where $k$ is the number of independent estimates (typically, the number of studies).

The $T$ values show a statistically significant heterogeneity, $Q(28) = 200.98$, $p < .0001$. When analyzing the variability of the means and variances, again with the $Q$ statistic, we also obtain significant heterogeneities [$M_s$, $Q(28) = 210.77$, $p < .0001$; variances, $Q(28) = 279.71$, $p < .0001$]. The combined and weighted mean and variance by random effects models are 40.7 and 99.88, respectively. As we have previously explained, we have studied the relationships among the three statistics. The Pearson and Spearman correlations are shown in Figure 2. They suggest that, as expected, the alpha coefficients show a positive relationship with the variances. On the other side, the variances also show a positive relationship with the means. As the homogeneity test for the variances suggests that this set of studies is heterogeneous in their sampling schemes, an attempt to identify its source is in order. When meta-analyzing the variances we find the following results. A regression model indicates that the sample means explain a substantial part of the variability in the variances, $Q(1) = 59.35$, $p < .001$, $R^2 = .629$, whereas the variable mean age does not (see Table 2). The same happens with a categorical model that employs a moderator variable the type of sample, but not with the variable language (see Table 2). Perhaps the moderator variables that yield significant results are explanatory only because they are proxy measures of the differential sampling schemes (e.g., the type of sample—normal students, medical outpatients, or psychiatric patients), or perhaps because they do indeed have an influence on other sources of perturbation discussed in previous sections (e.g., the variance of the measurement errors). We go back to this point in the next section.

Our interpretation of the results is that the way the different samples are recruited yields different sample variances, due to differences in the sampling schemes applied for selection. For example, the average scores are lower in the samples of students, normal controls, or medical patients (39.2) and larger in the samples of psychiatric patients in general or those with an anxiety disorder (52.8). Naturally, these biases show up also in the variances; specifically, the estimated combined variance of the studies for the first group is 94.7, whereas for the second group it is 150.5. As a consequence, we conclude that there is some variability in the alpha coefficients directly due to the variability in the composition and homogeneity of the samples, produced by the different sampling schemes. The sample variances, as reflecting different sampling schemes, should be incorporated into the model. Part of that variation is explained by the type of sample, but probably there are other unknown sources not yet identified; for example, the language, age, and/or gender, can also involve small differences in the sampling schemes of the true scores. As all those sources have a combined influence on the variance, they are all packaged into the model when the sample variance is incorporated into the model as a predictive variable for reliability.

Variations in the Sampling Schemes as a Source of Variation

Table 2. Perhaps the moderator variables that yield significant results are explanatory only because they are proxy measures of the differential sampling schemes (e.g., the type of sample—normal students, medical outpatients, or psychiatric patients), or perhaps because they do indeed have an influence on other sources of perturbation discussed in previous sections (e.g., the variance of the measurement errors). We go back to this point in the next section.

Exploring Potential Moderator Variables

We have studied the adjustments in several models, assessing the fit for the three statistics: the alpha coefficients ($T$ transformed), the means, and the variances. The studied moderator variables have been

\[\text{5 Given the problems already noticed with the normality assumption, we have repeated the test for the homogeneity of the variances taking only the 21 samples with } N \geq 100; \text{ the conclusion does not change, } Q(20) = 100.39, p < .0001.\]
the mean age of the sample, the sex (percentage of women in the sample), the language, and the type of sample. The main results, again by adjusting random effects models, are shown in Table 2. The table includes the \(Q\) statistics (Hedges & Olkin, 1985) for two categorical (analysis of variance) and two quantitative (regression) models, for the three statistics analyzed as dependent variables.

The variable language explains a significant part of the variability in the alpha coefficients; in examining the categories, a difference is observed between the administrations in Spanish and the rest. The combined reliability estimations from the studies with administrations in the different languages are as follows: Spanish, 0.873; English, 0.906; French, 0.909; Dutch, 0.916 (the administrations in Spanish include some with the translations developed in Spain and in Puerto Rico).

The type of sample also explains a significant part of the variability in the coefficients. However, because there are variations in the variances and means, the more parsimonious conclusion is that its influence is explained through variations in the sampling schemes.

Although apparently neither sex nor age have any association with the statistics, we have observed a noteworthy regularity for age. Specifically, a careful inspection of a scatterplot of age and alpha coefficients (see Figure 3) calls our attention to a unique behavior of the samples with larger mean age that deserves a deeper analysis. There is a trend for alpha coefficients to decrease, but only for high average ages. In order to pursue this point, we have created a new dichotomous variable that distinguishes the samples of older people (mean age \(> 65\)) from the others (mean age \(< 65\)). In the following section we discuss the role of that variable in explaining the variability of the coefficients.

### Table 1

**Main Statistics of the 29 Studies Included in the Example**

<table>
<thead>
<tr>
<th>Study</th>
<th>(\alpha)</th>
<th>(N)</th>
<th>(M)</th>
<th>Variance</th>
<th>Language(^a)</th>
<th>Age(^b)</th>
<th>% Women</th>
<th>Sample type(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.900</td>
<td>783</td>
<td>40.65</td>
<td>95.45</td>
<td>3</td>
<td>1</td>
<td>54.50</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.910</td>
<td>122</td>
<td>35.00</td>
<td>105.06</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.860</td>
<td>52</td>
<td>37.00</td>
<td>66.42</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.890</td>
<td>100</td>
<td>38.00</td>
<td>89.11</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.860</td>
<td>36</td>
<td>35.50</td>
<td>68.72</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.900</td>
<td>391</td>
<td>41.72</td>
<td>105.06</td>
<td>3</td>
<td>1</td>
<td>65.47</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>.930</td>
<td>384</td>
<td>42.31</td>
<td>112.36</td>
<td>3</td>
<td>1</td>
<td>100.00</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>.910</td>
<td>221</td>
<td>41.48</td>
<td>104.24</td>
<td>3</td>
<td>1</td>
<td>0.00</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>.861</td>
<td>57</td>
<td>40.50</td>
<td>103.63</td>
<td>1</td>
<td>2</td>
<td>75.40</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>.915</td>
<td>54</td>
<td>39.20</td>
<td>158.51</td>
<td>1</td>
<td>2</td>
<td>51.90</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>.820</td>
<td>120</td>
<td>41.84</td>
<td>93.90</td>
<td>1</td>
<td>2</td>
<td>68.30</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>.898</td>
<td>888</td>
<td>41.59</td>
<td>90.44</td>
<td>2</td>
<td>1</td>
<td>50.00</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>.919</td>
<td>888</td>
<td>40.82</td>
<td>102.74</td>
<td>2</td>
<td>1</td>
<td>50.00</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>.876</td>
<td>417</td>
<td>42.61</td>
<td>97.70</td>
<td>1</td>
<td>1</td>
<td>78.00</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>.883</td>
<td>391</td>
<td>41.73</td>
<td>99.53</td>
<td>1</td>
<td>1</td>
<td>81.00</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>.920</td>
<td>279</td>
<td>35.35</td>
<td>108.37</td>
<td>3</td>
<td>1</td>
<td>68.90</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>.890</td>
<td>143</td>
<td>41.80</td>
<td>76.74</td>
<td>3</td>
<td>1</td>
<td>82.52</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>.900</td>
<td>432</td>
<td>37.45</td>
<td>76.91</td>
<td>4</td>
<td>1</td>
<td>81.00</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>.860</td>
<td>270</td>
<td>38.76</td>
<td>89.68</td>
<td>1</td>
<td>1</td>
<td>100.00</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>.860</td>
<td>137</td>
<td>36.32</td>
<td>78.85</td>
<td>1</td>
<td>1</td>
<td>0.00</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>.930</td>
<td>103</td>
<td>40.29</td>
<td>153.76</td>
<td>3</td>
<td>1</td>
<td>100.00</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>.940</td>
<td>101</td>
<td>43.93</td>
<td>170.30</td>
<td>3</td>
<td>1</td>
<td>0.00</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>.880</td>
<td>50</td>
<td>48.01</td>
<td>141.61</td>
<td>3</td>
<td>2</td>
<td>72.00</td>
<td>4</td>
</tr>
<tr>
<td>24</td>
<td>.790</td>
<td>94</td>
<td>26.80</td>
<td>32.50</td>
<td>3</td>
<td>2</td>
<td>69.10</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>.900</td>
<td>217</td>
<td>53.59</td>
<td>167.52</td>
<td>3</td>
<td>2</td>
<td>71.00</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>.891</td>
<td>93</td>
<td>40.23</td>
<td>89.66</td>
<td>1</td>
<td>1</td>
<td>84.00</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>.879</td>
<td>62</td>
<td>43.46</td>
<td>101.94</td>
<td>1</td>
<td>1</td>
<td>90.30</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>.930</td>
<td>567</td>
<td>56.10</td>
<td>141.61</td>
<td>3</td>
<td>1</td>
<td>63.00</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>.930</td>
<td>366</td>
<td>37.11</td>
<td>89.46</td>
<td>4</td>
<td>1</td>
<td>71.86</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\) Language: 1 = Spanish; 2 = French; 3 = English; 4 = Dutch.  
\(^b\) Age: 1 = mean age \(\leq 65\) years; 2 = mean age \(> 65\).  
\(^c\) Sample types: 1 = students or normal people; 2 = medical patients; 3 = varied psychological disorders; 4 = anxiety disorder.

### Some Specific Models

After performing several exploratory analyses of the available moderator variables to understand the fluctuations of the coefficient alpha, we have defined some new regression models. On one side, we assume that part of the variance has to do with the sampling schemes, so that for studies with larger sample variances, larger alpha coefficients are expected. However, we have not found any moderator variable that explains this variability in the variances. Model 1 includes only the sample variance as moderator, assuming that it is based on varied sampling schemes of unknown origin, with the exception of the part explained by the type of sample. Models 2 and 3 add one dichotomous moderator variable. In Model 2 that variable is the age group (0, adults; 1, older), whereas for Model 3 the variable incorporated is a dichot-
omy associated with the language (1, Spanish; 0, other). We have also adjusted a model (Model 4) with the three predictors (i.e., sample variance, age, and language). The results are in Table 3.

We have studied the incremental adjustment when the moderator variables are incorporated one by one, employing the statistic for nested models (Judd & McClelland, 1989; Maxwell & Delaney, 1990). We have found that Models 2 and 3 involve significant increments in the variance accounted for, compared with Model 1 [Model 2 over 1: \( F(1, 25) = 22.015, p = .001 \); Model 3 over 1: \( F(1, 25) = 11.290, p = .01 \)]. In the same vein, Model 4 conveys a significant increment as related to Models 2 and 3 [Model 4 over 2: \( F(1, 24) = 12.391, p = .01 \); Model 4 over 3: \( F(1, 24) = 23.152, p = .001 \)].

The Final Model

Model 4 is our final proposal, with which we explain 77.9% of the observed variance in the internal consistency coefficients studied. It is expressed in terms of the \( T \) transformation. In this model, the value of \( T \) is smaller (larger alpha) as the sample variance increases, but \( T \) is also larger (smaller alpha) when the language is Spanish (\( L = 1 \), versus \( L = 0 \) for other languages) and when it is applied to samples of older people (\( A = 1 \) for mean age \( > 65 \) years vs. \( A = 0 \) for other mean ages). Specifically, the adjusted model is

\[
T' = 0.546 - 0.001 \times \text{var} + 0.039 \times L + 0.063 \times A.
\]

Table 2

<table>
<thead>
<tr>
<th>Model</th>
<th>( \alpha ) Coefficient</th>
<th>( M )</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language (( df = 3 ))</td>
<td>12.801*; ( p = .005 )</td>
<td>0.780*; ( p = .850 )</td>
<td>1.181; ( p = .758 )</td>
</tr>
<tr>
<td>Type (( df = 3 ))</td>
<td>9.714; ( p = .021 )</td>
<td>37.236; ( p &lt; .001 )</td>
<td>18.415; ( p = .004 )</td>
</tr>
<tr>
<td>Quantitative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex (( df = 1 ))</td>
<td>0.562; ( p = .454; R^2 = .017 )</td>
<td>0.0066; ( p = .909; R^2 = .001 )</td>
<td>0.172; ( p = .678; R^2 = .007 )</td>
</tr>
<tr>
<td>Age (( df = 1 ))</td>
<td>0.879; ( p = .348; R^2 = .024 )</td>
<td>0.435; ( p = .510; R^2 = .015 )</td>
<td>1.621; ( p = .203; R^2 = .051 )</td>
</tr>
</tbody>
</table>

* Apparently due to smaller alpha in Spanish (within-class heterogeneity in English).  ** Showing heterogeneity only within the English category.  * Only with the 25 studies reporting this variable.
What does this mean with an example inversely transformed to alpha values? The model predicts that in a sample for which the individuals have an adult mean age lower than 65, with the English, French, or Dutch version, and with a sample variance of 100, the predicted alpha coefficient is 0.911; if it is in Spanish, the prediction is 0.886; finally, if the sample includes only older Spanish speakers, the prediction is 0.835. Related to the effects of the sampling scheme, we can also specify its magnitude: For each increment (or decrement) of 10 points in the sample variance due to variations in the sampling scheme, the coefficient is increased (decreased) by 0.006. The model reflects nicely the relationship between the sample variances and the alpha coefficients, modulated by the predictive variables included (see Figure 4).

What this model expresses is that the variations in alpha can be explained by three sources of perturbation. First, the variation in alpha is associated with the different sampling schemes implicit in the different procedures for recruiting the participants for each study. We have identified the variable Type of Sample (i.e., normal participants, medical patients, varied psychologically disordered and anxious participants) as the one capturing that part of the variation. Second, there is smaller internal consistency when the Spanish version is employed. Third, there is smaller internal consistency when it is applied to older participants (>65 years). The fact that these two last sources of perturbation are necessary to include is reflected in the relative shortcomings of a model including only the sampling scheme as the source of variation. The alpha values of the samples in Spanish and/or with older participants are too low for the sample variances they show (conversely, because their sample variances are too large for their alpha values). This is graphically shown (see Figure 4) in that the points associated with those samples have negative residuals (Zone I of Figure 4). If these factors were producing an increment in the homogeneity of the items, then they would show alpha values larger than expected for their variances, so that their points in the scatter plot would appear with positive residuals (zone II of Figure 4).

In short, we must conclude that these two factors, the Spanish version and the participation of older people, yield scores with lower alpha because they increase the error variance. The first is probably due to slight variations in the meaning of the words employed in the translation or to cultural differences. The second is probably due to the fact that older people are more vulnerable to some extraneous variables when taking a test; thus, because older people in general have poorer memory (STAI is a self-report), they misunderstand more often what they read or make more errors in marking the responses.

### General Discussion

The meta-analysis of the reliability of test scores allows psychometric inferences about the scores and their sources of variance. This

<table>
<thead>
<tr>
<th>Model</th>
<th>Moderator variables</th>
<th>$Q_R$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Sample variances</td>
<td>17.892</td>
<td>.370</td>
</tr>
<tr>
<td>M2</td>
<td>Sample variances, age group</td>
<td>45.110</td>
<td>.665</td>
</tr>
<tr>
<td>M3</td>
<td>Sample variances, language</td>
<td>39.945</td>
<td>.566</td>
</tr>
<tr>
<td>M4</td>
<td>Sample variances, age group, language</td>
<td>80.846</td>
<td>.779</td>
</tr>
</tbody>
</table>

*Note. The second column shows the moderator variables included in each model (see text). The two last columns show the main statistics.

* $Q$ statistic for the meta-analytical regression (Hedges & Olkin, 1985).

* $p < .0001$. 

---

**Figure 3.** Scatter plot of the alpha coefficients with the average age of the sample.

**Figure 4.** Scatter plot of the alpha coefficients with the sample variances. The studies with samples of older people appear as empty circles (○); those with a Spanish version of the test, as triangles (△); those with both older people and the Spanish version, as plus signs (+); those with no older people and in a language other than Spanish, as filled circles (●).
can be done by a parallel meta-analysis of the observed variability in the reliability coefficients and in the samples’ means and variances, by applying the more usual meta-analytic procedures (Hedges & Olkin, 1985). The coefficients must be previously transformed for normality, e.g., by Fisher’s $z$ for the test–retest reliability (although see also Schulze, 2004, for possible biases with this transformation) or by Hakstian and Whalen’s $T$ (1976; Rodriguez & Maeda, 2006) for the internal consistency (Cronbach’s alpha). These two types of coefficients should not be mixed, as they reflect different sources of measurement error (Dimitrov, 2002). We have identified three sources of variation that, when associated with a moderator variable, produce specific effects in that triad of statistics.

The first source of variation is the sampling scheme. The different studies often involve varied procedures for recruiting the participants. Those procedures convey different sampling schemes from the general population of true scores. The consequence is that under those different sampling schemes the population variance estimates are different and changes are produced in their reliability coefficients. When the sampling schemes vary, the variation in the reliability coefficients is partly due to this source. A study of the homogeneity of the samples’ means and variances allows the identification of a scenario of varied sampling schemes. The identification of a scenario like that legitimizes the inclusion of the samples’ variances as a moderator variable. However, it is also interesting to identify any moderator variable capitalizing partly on these differences in sampling schemes. The practice of adjusting the alpha coefficients with Equation 4 before performing the analyses (e.g., Rodriguez & Maeda, 2006) can be misleading. In fact, if a moderator variable simultaneously explains part of the differences in the sampling schemes apart from any of the other sources of variation, then the adjustment will eliminate what we are trying to explain. A better way to deal with this type of situations is to adjust a multiple regression model with the samples’ variances as one of the predictors.

The second source is error variance. Whenever a moderator variable modulates the error variance, the empirical variances of the samples will show some variability associated with that moderator. Specifically, looking at Equations 1, 2, and 3, we must anticipate that the categories with larger error variances will show larger sample (empirical) variances and smaller reliability or internal consistency coefficients (error variances are in the denominator of Equations 2 and 3). However, the averages do not change, as the expected value equals the mean true score. The third source is the correlation between the items. Whenever a moderator variable modulates the correlation between the items, it is accompanied by larger alpha coefficients as the sample variances increase, whereas the averages do not change (remember that the relationship between the empirical variances and the reliability coefficients is inverse when the variability of the variances is due to differences in the error variance; however, it is direct when the variability is due to differential sampling schemes or variations in the correlations between the items).

Our framework assumes CTT, in which a number of exigent assumptions are made. The violation of those assumptions in a specific test could invalidate many predictions. However, one of the goals of meta-analysis is an attempt to identify sources of systematic variations in the reliability of the scores, a goal shared with the generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) rather than with CTT.

We have illustrated our framework with an example, although it cannot be considered a meta-analysis providing sound answers about the test involved, given that we have not tried to exhaust the published studies. Specifically, we have shown how in the 29 studies included there is a variation in the alpha coefficient of the Trait STAI scores that is associated with a direct variation in the variances. We have identified this source with some variability in the sampling schemes for the true scores, given that they included studies with university students and normal controls but also medical and psychiatric patients, including patients with an anxiety disorder. We have also identified two moderator variables with significant associations with the alpha coefficients and the variances, although in a negative way. Specifically, the use of the Spanish version and the application to samples of older people, with an average age over 65, introduces two sources of variation that we have interpreted as increments in the error variance (larger standard error of measurement). The model that includes these three sources of perturbation explains 77.9% of the variation in the alpha coefficients.

**Statements About “Typical Reliability”**

Another issue related with reliability generalization is the estimation of a general value, representing the internal consistency of the scores in a test. It is not clear whether there is any sense in estimating a general value, as we are acknowledging that reliability statements should be accompanied by the role of a set of moderator variables. However, if an investigator does wish to compute such a value, then it is considered advisable that some cautionary steps be taken.

Naturally, the value obtained by the weighted combination of the coefficients included in the studies of the meta-analysis can be reported. However, this value can be severely misleading, for example, if the set of studies includes many clinical trials or intervention programs with samples of individuals with clinical disorders. Then, both the estimation of the combined variance and the alpha coefficient will be misleading if they are interpreted as a typical value for the internal consistency of the scores with that test for the general population. Often the participants in the available studies oversample subpopulations with specific problems or with features not representative of the general population (e.g., university students, especially from psychology departments). Apparently, it would be more appropriate and informative to report the specific estimates for the subpopulations, showing significant deviations from the general value, besides (or in place of) that general value.

But the question remains whether there is any sense in adjusting the coefficient to a specific population value for the variance by applying Equation 5. Our view is that the value reported as more typical must be the value obtained with the combined estimation in the meta-analysis, once studies with characteristics identified as sources of perturbation are eliminated. The combined estimate must be adjusted for range differences to the best available estimate of the population variance; that is, an estimate obtained from a large sample that is composed in such a way that it is representative of the population. The large samples employed in the test-building processes and the elaboration of the norms for the manual often will be the best choice.

Let us go back to the example. After eliminating the samples where the Spanish version of the STAI was applied or older people were studied (15 studies), the combined internal consistency estimate is 0.909. The estimation of the population variance from
these studies is 97.20. Suppose that for a sample specifically
designed to be representative, the manual of the test reports a
variance of 100 in the United States. Applying Equation 4, we find
that the adjusted coefficient for that variance is .911. This is the
value for internal consistency that better represents the majority
of the versions in English or other languages and with adult ages (not
older people). This value should be modulated for other circum-
stances by the model reached above.

The procedures and the framework proposed can be useful in the
coming years. We hope that journals will be stricter about the infor-
mation that must be included in the research reports, besides imple-
menting a wider general policy of sharing the raw data. In such
circumstances the psychometric inferences could be based in the
analysis of the raw scores—although that is not a meta-analysis,
strictly speaking. However, by now we must be familiar with a
typically meta-analytical scenario in which the researcher has avail-
able the statistics that are reported in the articles that are published
today. Often those data, together with the descriptive information in
the text of the study, are all that the meta-analyst has available. In a
context like this, the challenge is to extract all the information that a
judicious and rigorous analysis can offer about the psychometric
characteristics of the scores obtained with a test. The framework
proposed here is an attempt to contribute to that task.

Take-Home Message and Conclusion

Very often psychologists include in research projects the measure-
ment of a psychological characteristic or construct. One is commonly
advised to take the psychometric quality of the instrument as part of
the basis for the choice of a specific test. However, it is also true that
most often the only index of reliability is that from the manual.
The results of the reliability generalization studies, for which the frame-
work described here have been proposed, allow a better basis for
choosing a measurement tool. The moderators included in the ad-
justed models are valid cues for that choice. A test with high reliability
for the general population can show significantly lower values for
specific subpopulations. In those cases an alternative test can be a
better choice; the meta-analysis of reliability and/or internal consis-
tency can provide that basis.

However, for most of the more popular tests (e.g., the STAI), the
analyses under the present framework would probably predict
small changes in the reliabilities associated to the moderator vari-
bles. That is also good news, as it means that the test is robust
under the potential effects of the moderator variables analyzed. In
fact, when the moderators have no explanatory role or the size of
their influence is small, it can be concluded that the variance of
error measurement is basically stable. It can be a source in arguing
the soundness of the choice of a specific test.

In short, the reliability generalization studies performed under the
present framework have two main advantages. First, they are more
informative, as the role of the sample variances is taken in account.
Second, the results are usually revealing, both when the moderators
predict relevant changes in the coefficients and when they do not. In
the first case the result is important in helping motivate the choice of a
different test when the reliability is significantly lower. In the second
case, results can help support the robustness of the test. Often the
arguments for using a specific test in a study rely on the estimation of
the reliability obtained in the process of building the test and on
specific psychometric studies. From our point of view, a reliability
generalization study that shows homogeneous results, once the vari-
ance due to the heterogeneity of the samples variances has been taken
into account, is an excellent and sound complementary argument for
employing the test.

References

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Pro-
ceedings of the Royal Society, Series A: Mathematical and Physical

Introduction to meta-analysis. Chichester, United Kingdom: Wiley.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency
or item redundancy in psychometric scales? Personality and Individual
Differences, 12, 291–294.

Praeger.

Cochran, W. G. (1954). The combination of estimates from different

Cooper, H., & Hedges, L. V. (Eds.). (1994). The handbook of research

Psychometrika, 16, 297–334.


ogy, 34, 481–489.

and potential problems with generalization across studies. Educa-
tional and Psychological Measurement, 62, 783–801.


educational and psychological tests for cross-cultural assessment.

Orlando, FL: Academic Press.

meta-analysis. Psychological Methods, 3, 486–504.

ods: A note on the frequency of use of various types. Educational
and Psychological Measurement, 60, 523–531.

Judd, C. M., & McClelland, G. H. (1989). Data analysis: A model com-


scores. Reading, MA: Addison-Wesley.

coefficients of stability estimates: Some recommendations based on Monte

Maxwell, S. E., & Delaney, H. D. (1990). Designing experiments and ana-

Psychological Methods, 11, 306–322.

Sawilowsky, S. S. (2000). Psychometrics versus datametrics: Comment on
Vacha-Haase’s “reliability generalization” method and some EPM editorial

bridge, MA: Hogrefe & Huber.
Appendix

Total Mean and Variance for a Mixture of Two Populations

Let there be two populations: A ~ N(μ_A, σ_A^2) and B ~ N(μ_B, σ_B^2).

We define

\[ \mu_B = \mu_A + c_1 \]
\[ \sigma_B^2 = c_2 \cdot \sigma_A^2, \]

where \( c_1 \geq 0 \) and \( c_2 > 0 \) (under these conditions, \( \mu_B \equiv \mu_A \)).

When selecting an element to form part of the sample, the probability of taking an element from population A is \( \pi_A \), and that of taking an element from population B is \( 1 - \pi_A \) (\( 0 \leq \pi_A \leq 1 \)).

The expected value of the sample mean in these conditions is

\[ \mu_T(\pi_A) = \mu_A + c_1 - c_1 \times \pi_A. \]

That is, the expected value for the sample mean is a linear function of \( \pi_A \) (intercept, \( \mu_A + c_1 \); slope, \(-c_1\)). The slope will be larger as larger is the difference between the subpopulations’ means (\( \mu_A - \mu_B \)).

It can also be demonstrated that the expected variance is

\[ \sigma_T^2 = -c_1^2 \cdot \pi_A^2 + (\sigma_A^2 - c_2 \cdot \sigma_A^2 + c_1^2) \cdot \pi_A + c_1 \cdot c_2 \cdot \sigma_A^2, \]

that is, a quadratic function of \( \pi_A \). Again, the difference between the means of the subpopulations will have a strong effect in the composed variance.

As an example, suppose two subpopulations with parameters \( \mu_A = 50, \mu_B = 80, \sigma_A^2 = 100, \) and \( \sigma_B^2 = 150. \)

Figure A1. Expected mean (left panel) and variance (right panel) for a sample composed of a proportion \( \pi \) from subpopulation A and \( (1 - \pi) \) from subpopulation B; in the example, \( \mu_A = 50, \mu_B = 80, \sigma_A^2 = 100, \) and \( \sigma_B^2 = 150. \)