Journal of Indian Economy and Business
Vol. 1 No. 1 (April 2024); pp 44-77

Experimental researchers do not know it, but item reliabilities are crucial for prediction

David Trafimow

New Mexico State University

Abstract

The importance of test reliabilities for predicting criterion variables has been well-established by psychometricians and is familiar to experimental researchers too. Thus, standard operating procedure for experimental researchers includes assessing and reporting Cronbach’s alpha. However, experimental researchers generally ignore item reliabilities. And yet, item reliabilities can be argued much more important than indices of single-administration whole test reliabilities, such as Cronbach’s alpha. The present goal is to make that argument and detail the complications that arise upon considering item reliabilities. Item reliabilities interact complexly with true item-criterion correlation coefficients, true interitem correlation coefficients, the number of items, and whether the researcher engages in amalgamating or unamalgamating test items. Standard operating procedure should include the assessment and reporting of item reliabilities.

Keywords: item reliability; amalgamating; unamalgamating; item-criterion correlation coefficients; interitem correlation coefficients

Wordcount: 7304

Imagine that participants are presented with a test item twice under idealized conditions, without any effect of the first test-taking occasion on the second one (no fatigue, practice effects, and so on); the correlation between scores on the two test-taking occasions can be considered to index the reliability of the item (Lazarsfeld, 1959).[1] Equivalently (see Gulliksen, 1987 for a well-cited review and mathematical derivation), it is possible to invoke the classical test theory notion of indefinite test taking occasions, with each participant's expectation across these test taking occasions as her or his true score. Under the assumption that each person's observed score on the item on a single test-taking occasion equals that person's true score plus error, item reliability can be considered the variance in true scores for the item, across participants, divided by the variance in observed scores for the item (true score variance plus error variance): $reliability = \frac{true\ score\ variance}{observed\ score\ variance} = \frac{true\ score\ variance}{true\ score\ variance + error\ variance}$. Experimental researchers practically never care about item reliability. Should they?

Based on the seminal work by Spearman (1904), it is possible to argue that item reliabilities are unimportant if the reliability of the whole test is acceptable. Spearman's (1904) attenuation equation is presented below as Equation 1:

$$r_{ac} = r_{Tac}\sqrt{r_{aa'}r_{cc'}}. \tag{1}$$

Equation 1 includes the following components:

- r_{ac} denotes the observed correlation between a test a and a criterion c,
- r_{Tac} denotes the correlation between true scores on the test and the criterion, the correlation that would be obtained sans random measurement error,

[1] Even when conditions are not ideal, item reliabilities usually are estimated by computing the correlation across two test-taking occasions. Practically all statistics programs, and even many spreadsheets, are capable of performing the calculation. For example, in Excel, the CORREL command will work. In jamovi, it is possible to click on 'correlation matrix' under 'regression.'

- $r_{aa'}$ denotes the reliability of the test, and
- $r_{cc'}$ denotes the reliability of the criterion.

Equation 1 clarifies that if the reliabilities of the test and criterion are close to 1.00, then the test-criterion correlation coefficient a researcher is likely to observe will be close to the true correlation coefficient. Poor reliabilities would cause the observed correlation coefficient to attenuate substantially relative to the true correlation coefficient; hence, the label 'attenuation' equation.

Although impressive interitem correlation coefficients benefit the reliability of the whole test, traditional formulas, such as Cronbach's alpha (1951), show that even poor interitem correlation coefficients can be compensated merely by including many items (see Crocker & Algina, 1986; Gulliksen, 1987; Lord & Novick, 1968 for well-cited reviews). For instance, suppose a researcher has 40 items, and the average interitem correlation coefficient equals 0.20. In that case, according to the Cronbach's alpha formula, alpha is 0.91. And other reliability formulas would give very positive overall reliability assessments too. In general, if the interitem correlation coefficients are large, few items are needed to obtain an impressive value for Cronbach's alpha; if the interitem correlation coefficients are small, many items are needed to obtain an impressive value for Cronbach's alpha. Thus, provided there are sufficient items, item reliability is unimportant; it is the reliability of the whole test that matters. Item reliability only matters insofar as it affects overall test reliability. If the overall reliability of a test is sufficient for the researcher's goal, item reliabilities can be ignored. Consistent with this thinking, experimental researchers practically never report item reliability coefficients; they typically report Cronbach's alpha for whole tests.

Hence, we arrive at a conclusion that item reliabilities are unimportant if the overall reliability of the test is impressive. And this can be accomplished simply by including sufficient items. The present goal is to argue to the contrary, with much nuance.

Guilford and Fruchter (1973), and implications

A limitation of Equation 1 is that it does not include items, thereby rendering difficult determining the effect of items on a test's ability to predict a criterion (Trafimow, Hyman, & Kostyk, 2023). Guilford and Fruchter (1973) explicitly considered items with Equation 2 below (also see Gulliksen, 1987),

$$r_{cs} = \frac{\sum r_{ci}\sigma_i}{\sqrt{\sum \sigma_i^2 + 2\sum r_{ij}\sigma_i\sigma_j}}. \qquad (2)$$

Using their notation, Equation 1 has the following components:

- r_{cs} denotes the correlation between the single test, including all items, with the criterion,
- r_{ci} denotes the correlation between any one item X_i and the criterion,
- σ_i denotes the item's standard deviation, and
- r_{ij} denotes the correlation between X_i and any other item X_j, with j greater than i.[2]

Equation 2 implies that adding items can aid prediction but can harm prediction too. If the added items correlate reasonably well with the criterion, or at least approximately as well as the other items, then including them will increase the test's ability to predict the criterion. However, if the added items correlate sufficiently poorly with the criterion, relative to the other items, then including them will decrease the test's ability to predict that criterion (Trafimow et al., 2023). For a quick example, suppose we set all interitem correlation coefficients at 0.40 and

[2] Equation 2 assumes equal weights for the items. Guilford and Fruchter (1973) provided a more complex equation for unequal weighting too, but the simpler equation is sufficient for present purposes.

all item standard deviations and variances at 1.00. Two items each correlate with the criterion at the 0.50 level, and a third item also correlates with the criterion at the 0.50 level. In that case, the overall prediction is 0.60 with the original two items and 0.65 with the third item included. In contrast, if the third item correlates with the criterion at the 0.10 level, then including it results in an overall prediction equal to 0.47, a substantial decrease from 0.60 using only the original two items.

It is not surprising that adding ‘good’ items betters criterion prediction whereas adding ‘bad’ items worsens criterion prediction. However, Equation 2 implies surprising news too. Consider that to publish in top experimental journals, researchers must report impressive reliabilities. As researchers usually favor Cronbach’s alpha, these need to exceed a threshold of .70 or 0.80, depending on the journal editor or reviewers. Furthermore, because Cronbach’s alpha depends on (1) interitem correlation coefficients and (2) the number of items, if we hold the number of items constant, the larger the interitem correlation coefficients, the more impressive the value for Cronbach’s alpha. Although a researcher can overcome poor interitem correlation coefficients by having many items, it is often inconvenient to use long tests. Hence, many researchers experience pressure to have interitem correlation coefficients be as large as possible to maximize Cronbach’s alpha and the probability of publication.

Another perceived advantage to having large interitem correlation coefficients is the common belief that large interitem correlation coefficients maximize the ability to predict a criterion. After all, large interitem correlation coefficients maximize Cronbach’s alpha, the most typical reliability index, and Equation 1 indicates that better reliability increases the ability of a test to predict the criterion. However, a careful investigation of the denominator of Equation 2 belies all this. To see why, note that the terms containing the interitem correlation coefficients r_{ij}

are connected to each other and other terms by plus signs. Hence, the larger the interitem correlation coefficients, the larger the denominator of Equation 2, and the worse the overall prediction of the criterion r_{cs}. For example, imagine a two-item test where the standard deviations of the items are set at 1.00, item-criterion correlation coefficients both equal 0.6, and the interitem correlation coefficient equals 0.90 or 0.10. Although researchers typically would rather have the larger (r_{12} = 0.90) than smaller (r_{12} = 0.10) interitem correlation coefficient, to demonstrate the reliability of the test, it is the smaller value that results in superior prediction of the criterion (r_{cs} = 0.81), and the larger value that results in inferior prediction of the criterion (r_{cs} = 0.62). The typical insistence on large interitem correlation coefficients incurs a large cost on researchers with respect to criterion prediction, though experimental researchers are unaware of it.

This counterintuitive effect, that large interitem coefficients that are good for single-administration reliability, such as Cronbach's alpha, are deleterious for criterion prediction, suggests that perhaps there is something wrong with single-administration reliability. Several researchers have argued that single-administration reliability does not properly index classical reliability (Revelle & Condon, 2019; Subkoviak, 1976; Trafimow et al., 2023), though this is not widely understood (Dunn et al., 2014; Lee and Hooley, 2005). Indeed, the thrust of classical test theory emphasizes that true scores are expectations of indefinite independent measurements, which seems inconsistent with single administration reliability such as Cronbach's alpha.[3] The present argument adds that although, by Equation 1, reliability is supposed to aid criterion prediction, single administration reliability decreases criterion prediction, keeping all else

[3] Lazarsfeld (1959) provided a widely cited interpretation that involves repeated testing with mind-washing between tests to ensure independence. A person's true score is the expectation across these indefinite tests.

constant. Hereafter, ‘reliability’ denotes classical reliability that is distinguishable from Cronbach’s alpha.

Although the surprising deleterious effect of large interitem correlation coefficients is both interesting and crucial, the effect does not yet address the issue of item reliabilities. To move in the direction of item reliabilities, it is necessary to modify Equation 2.

Modifying Equation 2

There are two categories of correlation coefficients in Equation 2. These are item-criterion correlation coefficients r_{ci} and interitem correlation coefficients r_{ij}. It is possible to express each of these in terms of true correlation coefficients and item reliabilities, simply by invoking Equation 1 and applying it to both categories of correlation coefficients. Thus, we have the following modified components:

- r_{ci} becomes $r_{Tci}\sqrt{r_{ii'}r_{cc'}}$ and
- r_{ij} becomes $r_{Tij}\sqrt{r_{ii'}r_{jj'}}$.

In both cases, we have the true correlation coefficient multiplied by the square root of the product of the reliability coefficients. Based on the modified components, Equation 2 becomes Equation 3:

$$r_{cs} = \frac{\sum r_{Tci}\sqrt{r_{ii'}r_{cc'}}\sigma_i}{\sqrt{\sum \sigma_i^2 + 2\sum r_{Tij}\sqrt{r_{ii'}r_{jj'}}\sigma_i\sigma_j}} \tag{3}$$

Equation 3 has an important advantage over Equation 2, for present purposes, which is that it includes item reliability coefficients. However, an important disadvantage is that the item standard deviations σ_i and σ_j and variances σ_i^2 are influenced by both variation in true scores and random variation. Likewise, both influence the reliability coefficients. Thus, there is no way to assess the effect of changes in one variable keeping the other variables constant; another modification is needed.

To move in this direction, consider two classical equations.[4] A classical definition of reliability is true score variance divided by the sum of true score variance and error variance: $r_{aa\prime} = \frac{\sigma^2_{Ta}}{\sigma^2_{Ta}+\sigma^2_{Ea}}$. Secondly, true score variance plus error variance compose variance: $\sigma^2_a = \sigma^2_{Ta} + \sigma^2_{Ea}$. Therefore, the components of Equation 3 can be modified as follows:

- $r_{ii\prime}$ becomes $\frac{\sigma^2_{Ti}}{\sigma^2_{Ti}+\sigma^2_{Ei}}$,
- $r_{jj\prime}$ becomes $\frac{\sigma^2_{Tj}}{\sigma^2_{Tj}+\sigma^2_{Ej}}$,
- $r_{cc\prime}$ becomes $\frac{\sigma^2_{Tc}}{\sigma^2_{Tc}+\sigma^2_{Ec}}$,
- σ^2_i becomes $\sigma^2_{Ti} + \sigma^2_{Ei}$,
- σ_i becomes $\sqrt{\sigma^2_{Ti} + \sigma^2_{Ei}}$, and
- σ_j becomes $\sqrt{\sigma^2_{Tj} + \sigma^2_{Ej}}$.

Instantiating the modified components into Equation 3 renders Equation 4:

$$r_{cs} = \frac{\sum r_{Tci}\sqrt{\left(\frac{\sigma^2_{Ti}}{\sigma^2_{Ti}+\sigma^2_{Ei}}\right)\left(\frac{\sigma^2_{Tc}}{\sigma^2_{Tc}+\sigma^2_{Ec}}\right)}\sqrt{\sigma^2_{Ti}+\sigma^2_{Ei}}}{\sqrt{\sum(\sigma^2_{Ti}+\sigma^2_{Ei})+2\sum r_{Tij}\sqrt{\left(\frac{\sigma^2_{Ti}}{\sigma^2_{Ti}+\sigma^2_{Ei}}\right)\left(\frac{\sigma^2_{Tj}}{\sigma^2_{Tj}+\sigma^2_{Ej}}\right)}\sqrt{\sigma^2_{Ti}+\sigma^2_{Ei}}\sqrt{\sigma^2_{Tj}+\sigma^2_{Ej}}}}. \qquad (4)$$

Although Equation 4 is inelegant, it has the advantage that crucial components can be varied, keeping the others constant. For example, item error standard deviations can be manipulated to

[4] Versions of these equations are provided in well-cited reviews (Crocker & Algina, 1986; Gulliksen, 1987; Lord & Novick, 1968), as well as in the original work by Spearman (1904).

influence item reliability coefficients, keeping true standard deviations constant. What lessons can we learn from Equation 4?

Consequences of Equation 4

For a preliminary consequence, consider again the case where there are 40 items and the average interitem correlation coefficient equals 0.2, so that the overall value for Cronbach's alpha equals 0.91. Nor is it necessary to use Cronbach's alpha. The traditional Spearman-Brown formula, using 0.20 for the item reliabilities, also renders a value of 0.91 (Brown, 1910; Spearman, 1910).[5] In addition, suppose that the item-criterion correlation coefficients equal 0.3, and the interitem correlation coefficients equal 0.10. Finally, suppose that the criterion is measured with perfect reliability. In that case, the ability of the 40-item test to predict the criterion equals 0.51. However, with perfectly reliable items, the value would equal 0.86. Converting to variance in the criterion explained by variance in the test, the observed variance explained is only 26%, relative to 73% that potentially could be explained, for a decrement due to unreliability equal to approximately 47%. These results are despite the impressive ostensible reliability of the whole test according to Cronbach's alpha (0.91). Therefore, item reliabilities are vital though this psychometric fact seems unknown to experimental researchers.

The four panels included in Figure 1 illustrate some consequences of Equation 4, in a more systematic way, where two test items predict a criterion.[6] In each panel, the correlation between the test and the criterion ranges along the vertical axis as a function of the error standard deviation along the horizontal axis (all true score standard deviations were set at 1.00 throughout

[5] Spearman-Brown requires parallel items, so the assumption would be that each item reliability equals 0.20, not just that the average equals 0.20. The value of 0.20 was used in making the calculations.

[6] To make the Figures, Equation 4 was converted to an Excel file using standard Excel commands. The Excel file is obtainable from the author by email request.

all explorations in this article). When the error standard deviation equals zero, then all items are measured without any random error, which represents the ideal case of perfect item reliability. As the error standard deviation increases along the horizontal axis, criterion prediction decreases. This decrease pertains to a single item and with the criterion reliability set at 1.00 (dotted curve), to both items and with the criterion reliability set at 1.00 (solid curve), or to both items and with the criterion reliability set at 0.70 (dashed curve).[7] Within each panel, it is easy to see that increasing the error standard deviation is generally deleterious for predicting the criterion, however, increasing the error standard deviation is less deleterious if it only happens to one item than to both items. And prediction worsens still more if the criterion is less than perfectly reliable.

---Insert Figure 1 about here---

The more interesting consequences of Equation 4 occur across panels. The panels differ in two respects: the true item-criterion correlations were set at 0.30 or 0.50 (left panels versus right panels) and the true interitem correlation coefficient was set at 0.10 or 0.90 (top panels versus bottom panels). It is interesting to consider the single best point in each panel, where the error standard deviation is set at zero. In the top panels, where the true interitem correlation coefficient was set at 0.10, the ability of both items to predict the criterion is substantially better than the ability of a single item to predict the criterion. However, in the bottom panels, where the interitem correlation is 0.90, even when the error standard deviation is set at zero, the ability of both items to predict the criterion is only slightly increased over the ability of a single item to

[7] To render the criterion reliability coefficient at 0.70, with the true standard deviation set at 1.00, the error standard deviation equals 0.654.

predict the criterion. Thus, we have another demonstration of the point that larger interitem correlation coefficients are harmful, not beneficial, for prediction.

An additional consequence can be seen by comparing each rightmost panel with its corresponding leftmost panel. When the true item-criterion correlation coefficients are set at 0.50, increasing random error with respect to one item, both items, or both items and the criterion, makes a larger difference than when the true item-criterion correlation coefficients are set at 0.30. In addition, the extent of the downward propagation of the curves is more pronounced when the true item-criterion correlation coefficients are set at 0.50 than when they are set at 0.30. In general, random error leads to a greater decrease in criterion prediction when true item-criterion correlation coefficients are larger than when they are smaller.

Too, it is possible to compare topmost panels with corresponding bottommost panels. Such comparisons show (a) that the differences between the curves are more pronounced when the true interitem correlation coefficient is small (0.10) than large (0.90) and (b) that the extent of the decrease in the curves as the error standard deviation increases is greater when the true interitem correlation coefficient is small than large. Although a small true interitem correlation coefficient is better than a large one for predicting a criterion, random measurement error can substantially decrease the gain in criterion prediction that researchers would otherwise enjoy with a small true interitem correlation coefficient.

Thus far, there has been no consideration of the number of items. Let us consider that now, setting all true item-criterion correlation coefficients equal at 0.5 and true interitem correlations coefficients equal at 0.1 or 0.9, as before; but have either a two-item or three-item test. Figure 2 illustrates what happens when the error standard deviations of the items range from

0 to 2.4, as in Figure 1. Like Figure 1, the various curves in all panels show that as more items have more random error, prediction of the criterion decreases.

---Insert Figure 2 about here---

A comparison of the first and second panels is of greater interest. Here, we see that criterion prediction is generally superior when there are three items than when there are two items. This result is due to the addition of a third good item. Had the third item been a bad item, adding it would have decreased, rather than increased, criterion prediction. More important, worst-case scenarios illustrated in the figure are less pronounced when there are three items than when there are two items; the curves propagate to better values for three-item than two-item tests. These conclusions also apply when comparing the third (two items) and fourth panels (three items), but where the true interitem correlation coefficients are at the 0.90 level as opposed to the 0.10 level.

Let us now consider the second and fourth panels, where there are three-item tests and where the true interitem correlation coefficients are set at 0.10 or 0.90. When there is little random error, there is an impressive advantage for setting the true interitem correlation coefficients at 0.10. as opposed to 0.90. However, when there is much random error, this advantage attenuates dramatically, especially when the criterion reliability coefficient is set at 0.70 as opposed to 1.00.

Unamalgamating Test Items

Everything stated thus far has been under the umbrella of amalgamating test items into a whole test, consistent with standard practice and standard recommendations. However, there is another option. Unamalgamating is possible where the researcher enters each item separately into a multiple regression equation. When there are two items, the standard equation for obtaining the multiple correlation coefficient $R_{y.12}$ from the interitem and item-criterion correlation coefficients is as follows, using Pedhazur's notation where *y* denotes the criterion (e.g., Pedhazur, 1997):

$$R_{y.12} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}}. \qquad (5)$$

To keep the notation consistent with Guilford and Fruchter (1973), we can represent the criterion with *c* rather than *y*. Substituting *c* for *y* in Equation 5 renders Equation 6:

$$R_{c.12} = \sqrt{\frac{r_{c1}^2 + r_{c2}^2 - 2r_{c1}r_{c2}r_{12}}{1 - r_{12}^2}}. \qquad (6)$$

Now let's consider reliability. We have the following components.

- r_{c1} expands to $\rho_{T_cT_1}\sqrt{r_{cc'}r_{11'}}$,
- r_{c2} expands to $r_{T_cT_2}\sqrt{r_{cc'}r_{22'}}$,
- r_{12} expands to $r_{T_1T_2}\sqrt{r_{11'}\rho_{22'}}$.

In turn, we also have the following.

- r_{c1}^2 expands to $r_{T_cT_1}{}^2 r_{cc'}\rho_{11'}$,
- r_{c2}^2 expands to $r_{T_cT_2}{}^2 r_{cc'}r_{22'}$,
- r_{12}^2 expands to $r_{T_1T_2}{}^2 r_{11'}r_{22'}$.

Instantiating all bullet-listed components into Equation 6 renders Equation 7.

$$R_{c.12} = \sqrt{\frac{\rho_{T_cT_1}{}^2\rho_{cc'}\rho_{11'} + \rho_{T_cT_2}{}^2\rho_{cc'}\rho_{22'} - 2\rho_{T_cT_1}\sqrt{\rho_{cc'}\rho_{11'}}\rho_{T_cT_2}\sqrt{\rho_{cc'}\rho_{22'}}\rho_{T_1T_2}\sqrt{\rho_{11'}\rho_{22'}}}{1 - \rho_{T_1T_2}{}^2\rho_{11'}\rho_{22'}}} \qquad (7)$$

Finally, to match the processes involved in deriving Equation 4, it is now necessary to expand all reliabilities to be expressed in terms of true and error variances.

- ρ_{11}, expands to $\frac{\sigma_{T1}^2}{\sigma_{T1}^2+\sigma_{E1}^2}$,
- ρ_{22}, expands to $\frac{\sigma_{T2}^2}{\sigma_{T2}^2+\sigma_{E2}^2}$,
- ρ_{cc}, expands to $\frac{\sigma_{Tc}^2}{\sigma_{Tc}^2+\sigma_{Ec}^2}$.

Instantiating the expanded reliabilities into Equation 7 renders Equation 8.

$R_{c.12} =$

$$\sqrt{\frac{{\rho_{T_cT_1}}^2\left(\frac{\sigma_{Tc}^2}{\sigma_{Tc}^2+\sigma_{Ec}^2}\right)\left(\frac{\sigma_{T1}^2}{\sigma_{T1}^2+\sigma_{E1}^2}\right)+{\rho_{T_cT_2}}^2\left(\frac{\sigma_{Tc}^2}{\sigma_{Tc}^2+\sigma_{Ec}^2}\right)\left(\frac{\sigma_{T2}^2}{\sigma_{T2}^2+\sigma_{E2}^2}\right)-2\rho_{T_cT_1}\sqrt{\left(\frac{\sigma_{Tc}^2}{\sigma_{Tc}^2+\sigma_{Ec}^2}\right)\left(\frac{\sigma_{T1}^2}{\sigma_{T1}^2+\sigma_{E1}^2}\right)}\rho_{T_cT_2}\sqrt{\left(\frac{\sigma_{Tc}^2}{\sigma_{Tc}^2+\sigma_{Ec}^2}\right)\left(\frac{\sigma_{T2}^2}{\sigma_{T2}^2+\sigma_{E2}^2}\right)}\rho_{T_1T_2}\sqrt{\left(\frac{\sigma_{T1}^2}{\sigma_{T1}^2+\sigma_{E1}^2}\right)\left(\frac{\sigma_{T2}^2}{\sigma_{T2}^2+\sigma_{E2}^2}\right)}}{1-{\rho_{T_1T_2}}^2\frac{\sigma_{T1}^2}{\sigma_{T1}^2+\sigma_{E1}^2}\frac{\sigma_{T2}^2}{\sigma_{T2}^2+\sigma_{E2}^2}}} \tag{8}$$

Equation 8 provides the mathematical basis for the conclusions explained below.

To see the potential value of unamalgamating, imagine a two-item test where all items and the criterion are measured without any measurement error, where Item 1 correlates with the criterion at 0.50, Item 2 correlates with the criterion at 0.10, and where Item 1 and Item 2 correlate at 0.90. We have already learned that a large interitem correlation coefficient decreases criterion prediction when amalgamating. In the present example, amalgamated criterion prediction is only 0.30, whereas it would be 0.40 if the true interitem correlation coefficient were dropped from 0.90 to 0.10. However, unamalgamating changes matters dramatically and a large true interitem correlation coefficient, such as 0.90, no longer decreases prediction. On the contrary, a large interitem correlation coefficient becomes very good for prediction when unamalgamating: the value is now 0.95! This is vastly improved over the 0.30 value for unamalgamated prediction. More generally, Trafimow et al. (2023) recently showed that unamalgamated prediction is always as good as, or better than, amalgamated prediction. Thus, Trafimow et al. advocated that future researchers should embrace unamalgamating, as opposed to the current practice of amalgamating, to best predict criterion variables.

However, although there is no attempt here to dispute Trafimow et al., the present focus suggests that their conclusion may be strongly qualified depending on item reliabilities. We have already seen that item unreliability attenuates criterion prediction under amalgamation, but potential harms under separate entry require exploration.

Let us commence by continuing the present example and considering Figure 3. As usual, each panel in Figure 3 relates criterion prediction with the error standard deviation, but with the solid curve representing amalgamating and the dotted curve representing unamalgamating. In each panel, unamalgamating is better than or equal to amalgamating. However, the degree of superiority depends heavily on whether the true interitem correlation coefficient is set at 0.90

(uppermost panels) or 0.10 (bottommost panels) and whether it is Item 1 (leftmost panels) or Item 2 (rightmost panels) that have varying levels of measurement error.

---Insert Figure 3 about here---

Consider the two uppermost panels. In the first panel, where the true item-criterion correlation coefficient is 0.50 for Item 1 but only 0.10 for Item 2, adding random error to Item 1 decreases criterion prediction when unamalgamating or amalgamating, but the effect is more pronounced when unamalgamating. In the second panel, where adding randomness applies to Item 2, as opposed to Item 1, unreliability still influences criterion prediction but to a lesser extent when unamalgamating; the dotted curve decreases more in the first panel than in the second panel. A possible explanation for these effects is that under perfect reliability, Item 2 profoundly influences criterion prediction when unamalgamating due to suppressing error in Item 1. Thus, adding random measurement error to Item 1 (a) decreases its ability to predict the criterion and (b) decreases the interitem correlation coefficient thereby reducing the ability of Item 2 to suppress error in Item 1. The combination of these effects contributes to the dramatic decrease in criterion prediction as the error standard deviation increases. In contrast, when it is Item 2 that is subject to various degrees of random error, although increasing the error standard deviation decreases the ability of Item 2 to suppress error variance in Item 1, it does not influence the ability of Item 1 to predict the criterion notwithstanding the error suppression effect of Item 2. Consequently, the extent of the decrease in criterion prediction is less pronounced in the second panel than in the first panel. The error suppression issue is less relevant under amalgamation; thus, the two solid curves are alike in both panels.

Moving to the bottommost panels, where the true interitem correlation coefficient is set at 0.10, there are two immediately obvious effects. As the opportunity for error suppression all but disappears, the dotted curves start at 0.51 as opposed to 0.95, under perfect reliability. Secondly, the solid curves are raised in the bottommost panels relative to the topmost panels. In summary, reducing the true interitem correlation coefficient is harmful for criterion prediction when unamalgamating but beneficial when amalgamating.

In addition, when the Item 1 error standard deviation increases in the third panel, it strongly influences criterion prediction, to the point where the advantage for unamalgamating over amalgamating eventually almost completely disappears. In contrast, when it is the Item 2 error standard deviation that increases in the fourth panel, the ability of Item 1 to predict the criterion is not affected, except for a miniscule error suppression effect that decreases so little that it is difficult to discern; thus, criterion prediction remains barely above the 0.50 level. The fourth panel is the only one in Figure 3 where more random error accentuates, rather than attenuates, the superiority of unamalgamating over amalgamating. Therefore, random measurement error can attenuate or accentuate the advantage of unamalgamating over amalgamating.

All panels in Figure 4 are like corresponding panels in Figure 3, with the single exception that the true item-criterion correlation for Item 2 was raised from 0.1 in Figure 3 to 0.4 in Figure 4. Because Item 2 correlates more with the criterion in Figure 4 than in Figure 3, it can be considered “better” in Figure 4 than in Figure 3. Hence, intuitively, criterion prediction ought to improve in Figure 4 relative to Figure 3. However, contrary to commonsense when unamalgamating (dotted curves), the “worse” item in Figure 3 sometimes leads to better prediction than the better item in Figure 4. This paradoxical effect is particularly evident in the uppermost panels. In these panels, when the error standard deviation is near zero, criterion prediction with the ostensibly better Item 2 in Figure 4 is paradoxically decreased relative to the ostensibly worse Item 2 in Figure 3. The reason for the paradoxical effect may pertain to error suppression. When Item 2 better predicts the criterion, less of its variance can be used to suppress error in Item 1, thereby paradoxically resulting in decreased criterion prediction. However, as error standard deviations increase, error suppression decreases, and the difference between the dotted curves in the uppermost panels of the two figures decreases. On the other hand, when amalgamating (solid curves), commonsense prevails, as each solid curve in Figure 4 represents better prediction than its corresponding curve in Figure 3.

---Insert Figure 4 about here---

It is also interesting to compare Figure 4 against Figure 3 with respect to the bottommost panels, when the true interitem correlation coefficient is 0.10. In this case, there is little difference between the dotted curves in the two figures, but there is an important difference in the solid curves. When amalgamating, the smaller true interitem correlation coefficient provides for better prediction when the error standard deviation is near zero but increasing the error standard deviation decreases that beneficial effect. In summary, under conditions favoring error suppression, a large true interitem correlation coefficient is beneficial when unamalgamating but harmful when amalgamating, with both effects qualified by the error standard deviation of Item 1 in the leftmost panels, or the error standard deviation of Item 2 in the rightmost panels.

It is interesting, too, to compare the four panels within Figure 4. As opposed to Figure 3, when the true interitem correlation coefficient is 0.1 in the bottommost panels, rather than 0.9 in the uppermost panels, even unamalgamating results in better prediction with the smaller true interitem correlation coefficient than with the larger one. This is because, as alluded to earlier, when Item 2 is “improved” in Figure 4, error suppression is dramatically diminished, and so what used to be a beneficial effect of a large true interitem correlation coefficient becomes a harmful effect. In fact, with error suppression reduced, and with a small true interitem correlation coefficient in the third and fourth panels of Figure 4, the curves representing unamalgamating (dotted curves) and amalgamating (solid curves) are quite similar. In the fourth panel, they are so similar that the difference is not visually discernable, and so it appears that there is only one curve. In fact, however, the dotted curve, though it cannot be seen, is very slightly above the solid curve.

Discussion

The examples and figures demonstrate that item reliabilities can dramatically influence criterion prediction. We have seen that even when the overall reliability of a test is impressive, unreliability at the level of items can nevertheless be problematic. A perhaps hidden issue is that single administration reliability indices, such as Cronbach's alpha, can be argued to poorly capture the essence of reliability (Dunn et al., 2014; Lee and Hooley, 2005; Revelle & Condon, 2019; Subkoviak, 1976). An advantage of single administration reliability indices is that they are easy and cheap, due to the lack of a necessity to measure people twice. However, this easiness is costly because the results provide a misleading reliability picture. Advantages of test-retest reliability are in (a) providing a less misleading overall reliability picture and (b) rendering possible the estimation of item reliabilities. Of course, even test-retest reliability is not perfect because the prior test-taking occasion can influence the subsequent one, but the advantages nevertheless outweigh the disadvantages. Furthermore, carryover effects can be mitigated in ways such as (a) embedding crucial items in a large set of unimportant ones to render memory more difficult, (b) increasing the delay between test-taking occasions, (c) providing distractor tasks, and (d) performing post-hoc analyses to determine whether there really are any substantial carryover effects. Regarding this last, although such carryover effects are possible and sometimes occur, they often do not occur (Trafimow & Rice, 2009).

In addition, there are many complex effects pertaining to true item-criterion correlation coefficients, true interitem correlation coefficients, and whether the researcher is amalgamating or unamalgamating test items. However, and at times crucially, these complex effects are, themselves, strongly qualified by item unreliability. Some complexities and qualifications are bullet-listed below.

- Adding items can benefit or harm criterion prediction, depending on true item-criterion correlation coefficients, true interitem correlation coefficients, and whether the researcher is amalgamating or unamalgamating.
- Large true interitem correlation coefficients harm criterion prediction when amalgamating, despite their desirability for obtaining impressive values on single administration reliability indices.
- Large true interitem correlation coefficients can benefit criterion prediction, if unamalgamating, when there is substantial error suppression.
- Large true interitem correlation coefficients can harm criterion prediction, even if unamalgamating, when true item-criterion correlation coefficients are sufficiently similar.
- Although unamalgamating is always equal to or superior to amalgamating for criterion prediction, the extent of the superiority can be immense, nonexistent, or anywhere in-between, depending on true item-criterion correlation coefficients and true interitem correlation coefficients.
- Perhaps most important, all foregoing bullet-listed effects are crucially qualified by item reliabilities. However, the extent of the qualification depends on complex configurations

of true item-criterion correlation coefficients, true interitem correlation coefficients, the number of items, and whether the researcher is amalgamating or unamalgamating.

- Item reliabilities can attenuate or accentuate the superiority of unamalgamating over amalgamating, depending on item-criterion correlation coefficients and interitem correlation coefficients.

The complexity of the bullet-pointed conclusions, especially when considered in totality, may seem daunting. How can a experimental researcher, who might not be an expert psychometrician, keep track of all the complexities? Fortunately, this may not be necessary, as there are simplicities buried in the complexities.

One such simplicity is that no matter the complexity of the configuration of true item-criterion correlation coefficients, true interitem correlation coefficients, and item reliabilities, unamalgamated prediction always equals or betters amalgamated prediction. Thus, a simple rule is that no matter the other complexities, researchers should favor unamalgamating. If unamalgamating fails to importantly increase criterion prediction over amalgamating, the researcher may or may not favor reporting findings obtained by unamalgamating. However, if unamalgamating importantly increases criterion prediction over amalgamating, this would constitute an important empirical reason for focusing on results obtained by unamalgamating.

Although the extent to which item unreliability harms criterion prediction varies greatly depending on configurations of true item-criterion correlation coefficients, true interitem correlation coefficients, number of items, and whether the researcher is amalgamating or unamalgamating, an underlying simplicity is that item reliability is generally beneficial for criterion prediction and item unreliability is generally harmful for criterion prediction. Therefore, it is worthwhile, when feasible, to include at least two test-taking occasions into the study design, to enable the estimation of item reliabilities. If criterion prediction is unimpressive, even if statistically significant, as is typical in the social sciences, item reliabilities may provide a strong clue as to why. If item reliabilities are near perfect, then perhaps the problem resides in small true item-criterion correlation coefficients. Then, too, true interitem correlation coefficients may be an issue, though this likely depends on whether the researcher is amalgamating or unamalgamating. However, if item reliabilities are poor, that may be the obvious first place to look to understand the reason for unimpressive criterion prediction. A strong suspicion is that many weak effect sizes in the social sciences are due to poor item reliabilities.

Is Unamalgamating Anti-Theory?

Quantitative demonstrations often suggest philosophical issues and the present work is no exception. One such issue pertains to theory. If a theory links a predictor construct to a criterion construct, then it seems *prima facie* sensible to amalgamate the items used to measure the construct. Moreover, it also seems sensible to create those items so that they correlate highly with each other; after all, they are all supposed to be measuring the same construct. However, we have seen that following this seemingly sensible strategy is deleterious for predicting the criterion. In turn, poor prediction, even if statistically significant, could be argued to undermine the worth of the theory. Thus, we have a dilemma, as the seeming theoretically sensible course of action and best strategy for criterion prediction oppose. However, there are potential circumventions. Unamalgamating is one, as it provides at least the possibility that large interitem correlations could then be beneficial rather than harmful for criterion prediction.

The obvious objection is that unamalgamating seems contrary to the idea that the test items are supposed to measure the same construct. However, it is possible to counter the objection. Consider, for instance, that extraversion, a popular Big 5 trait, includes ‘enthusiasm’ and ‘talkative’ items. A researcher could assume that both items measure extraversion, but a researcher could contrarily assume that the enthusiasm item measures enthusiasm and the talkative item measures talkativeness. There is no compelling reason to insist on an extraversion trait when one could instead assume enthusiasm and talkative traits. In that case, unamalgamating makes better theoretical sense than amalgamating, as well as being superior for criterion prediction, and the tension between theory and prediction disappears.

An objection to this line of reasoning hearkens back to the longstanding but not completely settled debate in the personality psychology area about whether personality traits cause behaviors or are merely convenient summaries of behaviors (see Buss & Craik, 1983, for a well-cited review). Either way can be considered problematic. Continuing with the extraversion example, to insist that enthusiasm and talkative items measure extraversion is a stretch. Yet, insisting that extraversion is merely a summary of behaviors seems a reversion to the bad old days of logical positivism and operationalism. However, there is a middle course to take. As alluded to earlier, it is possible to posit that enthusiasm and talkative items measure enthusiasm and talkativeness, respectively. In that case, there is no reversion to logical positivism and operationalism because there are clear traits here: enthusiasm and talkativeness. The stretch to extraversion is not necessary. Thus, the debate need not be about whether traits exist or whether they are merely summaries of behaviors but could be about which traits social scientists should assume to exist. It is one thing to insist that because enthusiasm and talkative items load on the same factor, they must measure extraversion, an unwarranted conclusion. It is quite another thing to hold that enthusiasm and talkative items measure enthusiasm and talkativeness, respectively, and that these traits happen to be sufficiently correlated, with each other and some other traits, that they load on the same factor.

The more nuanced philosophical thinking implies benefits. It justifies unamalgamating which is generally superior to amalgamating for criterion prediction. Secondly, the nuanced philosophical thinking opens the way to consider that the items might really be measuring different, though related, traits. Thirdly, a consideration that each item might be measuring a different trait renders reasonable the inclusion of traits that do not correlate well with the criterion, but that do correlate with another trait, thereby providing the possibility of spectacular error suppression. Recall that in Figure 3, error suppression caused criterion prediction to reach the gaudy level of 0.95, when the true item-criterion correlation coefficient was 0.50 for Item 1, and it was 0.10 for Item 2. Of course, as emphasized earlier, this depends, too, on item reliabilities.

Multicollinearity

A potential objection to unamalgamating is that if there are large interitem correlation coefficients, this constitutes a multicollinearity problem. It is even possible to argue that item unreliability is desirable because it decreases interitem correlation coefficients, thereby decreasing multicollinearity problems.

However, multicollinearity issues only apply if the researcher is interested in the regression weights. The present argument for unamalgamating does not pertain to regression weights, but rather to the multiple correlation coefficient. If a researcher is interested in the bivariate relations between individual items and a criterion, it is better to use zero-order correlation coefficients uncontaminated by relations with other items. Therefore, multicollinearity threats need not discommode researchers who are convinced that unamalgamating is desirable.

Conclusion

We commenced by considering and rejecting an argument about why item reliabilities are unimportant if the whole test is reliable. On the contrary, item reliabilities are crucial, and interact complexly with true item-criterion correlation coefficients, interitem correlation coefficients, the number of items, and whether one is amalgamating or unamalgamating.

It is interesting to consider the typically small effect sizes in psychology research (Schäfer & Schwarz, 2019). These are obviously problematic for application; although exceptions may exist, there typically is little reason to invest in applications associated with small effects. From a basic research standpoint, alternative explanations more plausibly explain small than large effect sizes. For example, although a correlation coefficient equal to 0.90 may be spurious, this is a difficult criticism to make; there are few outside variables that can plausibly explain such a large correlation coefficient. In contrast, to argue that a correlation coefficient equal to 0.10 is spurious is easily done; there are many outside variables that can explain such a small correlation coefficient. Thus, for both basic and applied research, small effect sizes can decrease the value of the research.

Are psychology researchers doomed to small effect sizes? One argument is that the psychological universe is so multi-causal that it is unreasonable to expect large effect sizes. Consequently, psychology researchers are doomed. However, an alternative possibility is that small effect sizes are due, in large part, to measurement problems. In that case, bettering measurement holds out the promise of likewise bettering effect sizes. In that spirit, the present work suggests two improvements. Researchers should stop amalgamating and instead embrace unamalgamating. Secondly, given the present demonstrations that item reliabilities are potentially crucial, researchers should obsess as much, or more, about item reliabilities as they currently do about reliabilities of whole tests. It is standard operating procedure to report Cronbach's alpha. However, item reliabilities are much more important than Cronbach's alpha for criterion prediction. Therefore, the assessment and reporting of item reliabilities should become standard operating procedure. Although stipulating researchers must assess and report item reliabilities would be a dramatic change in research and publication practice, the present quantification of the potential gains to be enjoyed more than justify the stipulation.

Declarations

Funding: Not applicable.
Conflicts of interest/Competing interests: Not applicable.
Ethics approval: Not applicable.
Consent to participate: Not applicable.
Consent for publication: Not applicable.
Availability of data and materials: Not applicable.
Code availability: Not applicable.

References

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322. SOME EXPERIMENTAL RESULTS IN THE CORRELATION OF MENTAL ABILITIES1 - BROWN - 1910 - British Journal of Psychology, 1904-1920 - Wiley Online Library

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review, 90*(2), 105-126. doi: 10.1037/0033-295X.90.2.105

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Belmont, CA: Wadsworth.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. doi: 10.1007/bf02310555

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399-412. doi: 10.1111/bjop.12046

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education*. New York, NY: McGraw-Hill.

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

Lee, N., & Hooley, G. (2005). The evolution of “classical mythology” within marketing measure development. *European Journal of Marketing*, *39*(3/4), 365-385. doi: 10.1108/03090560510581827

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: a study of a science*, Vol III (pp. 476-543). New York: McGraw-Hill.

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

Addison-Wesley.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395-1411. doi: 10.1037/pas0000754

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). United States: Wadsworth.

Schäfer, T. & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*(813), 1-13. doi: 10.3389/fpsyg.2019.00813

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72-101. http://www.jstor.org/stable/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295. CORRELATION CALCULATED FROM FAULTY DATA - SPEARMAN - 1910 - British Journal of Psychology, 1904-1920 - Wiley Online Library

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*(4), 265-276. doi: 10.1111/j.1745-3984.1976.tb00017.x

Trafimow, D., Hyman, M. R., & Kostyk, A. (2023). Enhancing predictive power by unamalgamating multi-item measures. *Psychological Methods*.

Trafimow, D., & Rice, S. (2009). Potential performance theory (PPT): Describing a methodology for analyzing task performance. *Behavior Research Methods*, *41*(2), 359-371. doi:10.3758/BRM.41.2.359

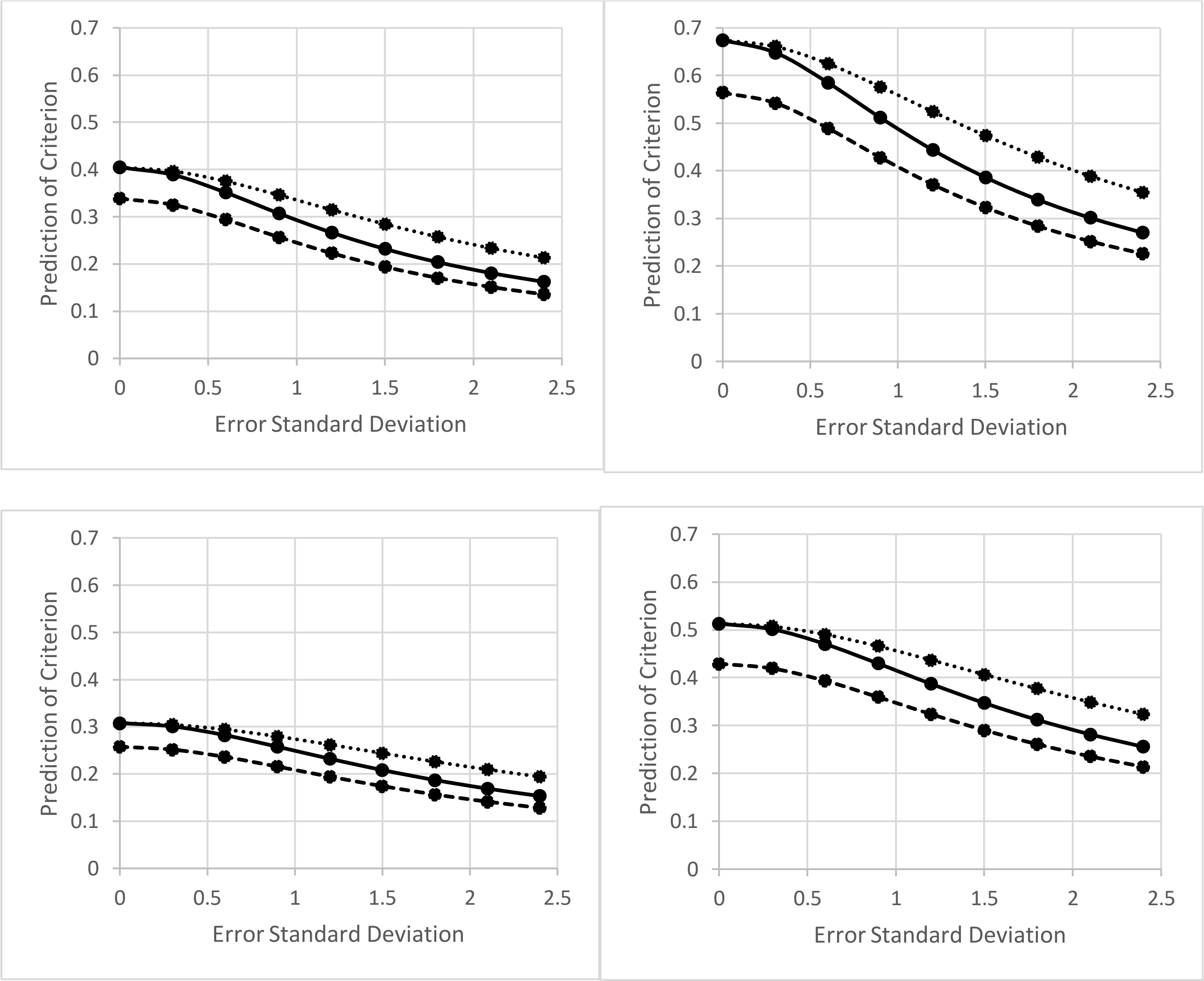

Figure 1. Prediction of criterion ranges along the vertical axis as a function of varying the error standard deviation for one item (dotted curve), two items (solid curve) or two items and setting the criterion reliability at 0.70 (dashed curve). The true item-criterion correlation coefficients were set at 0.30 (first and third panels) or 0.50 (second and fourth panels), and the true interitem correlation coefficient was set at 0.10 (first and second panel) or 0.90 (third and fourth panels).

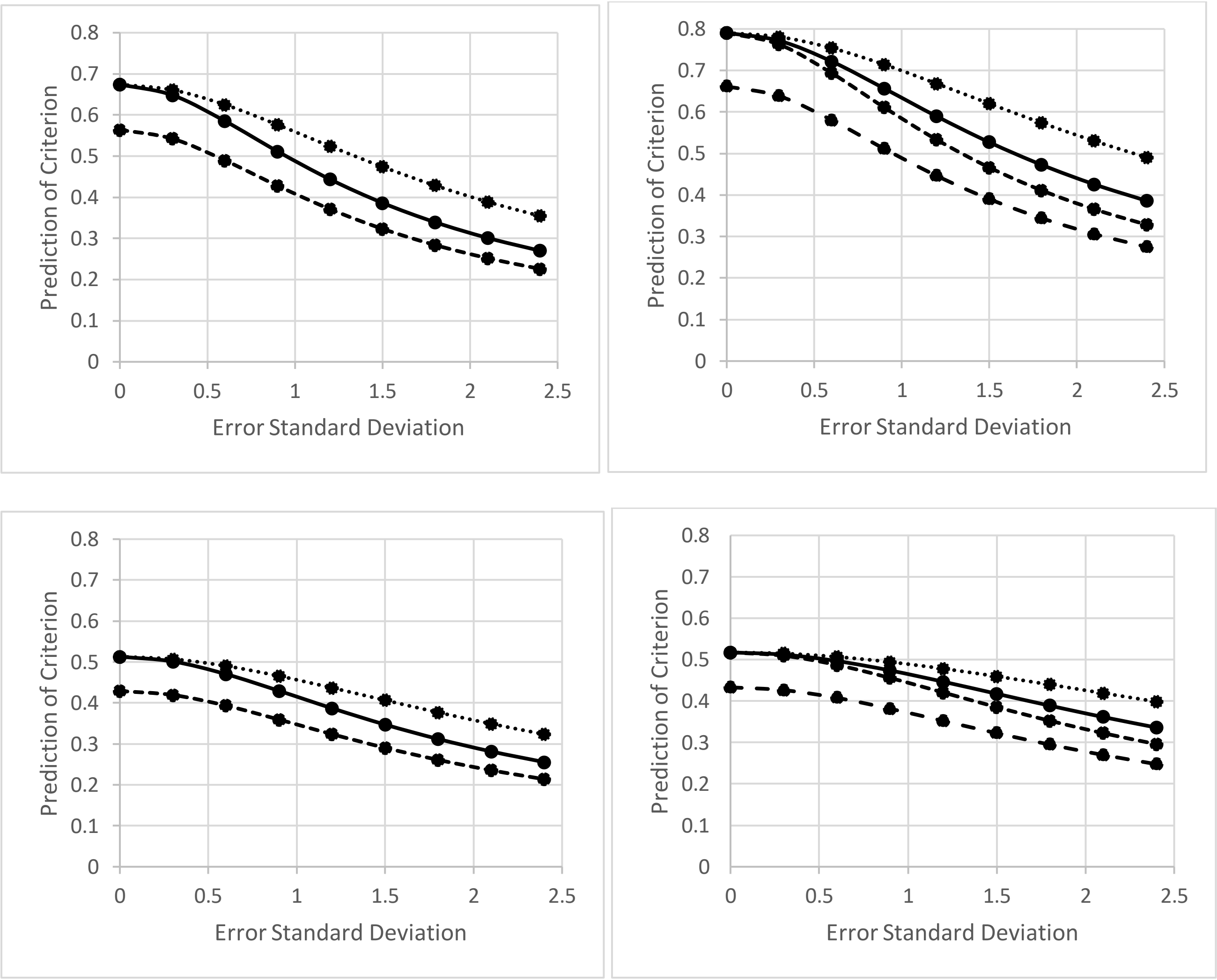

Figure 2. Prediction of criterion ranges along the vertical axis as a function of varying the error standard deviation. In the leftmost panels representing two-item tests, the error standard deviation varied for one item (dotted curve), two items (solid curve), or two items and setting the criterion reliability at 0.70 (dashed curve). In the rightmost panels representing three-item tests, the error standard deviation varied for one item (dotted curve), two items (solid curve), three items (dashed curve), or three items and setting the criterion reliability at 0.70 (long dashed curve). The true item-criterion correlation coefficients were set at 0.50, and the true interitem correlation coefficient was set at 0.10 (first and second panel) or 0.90 (third and fourth panels).

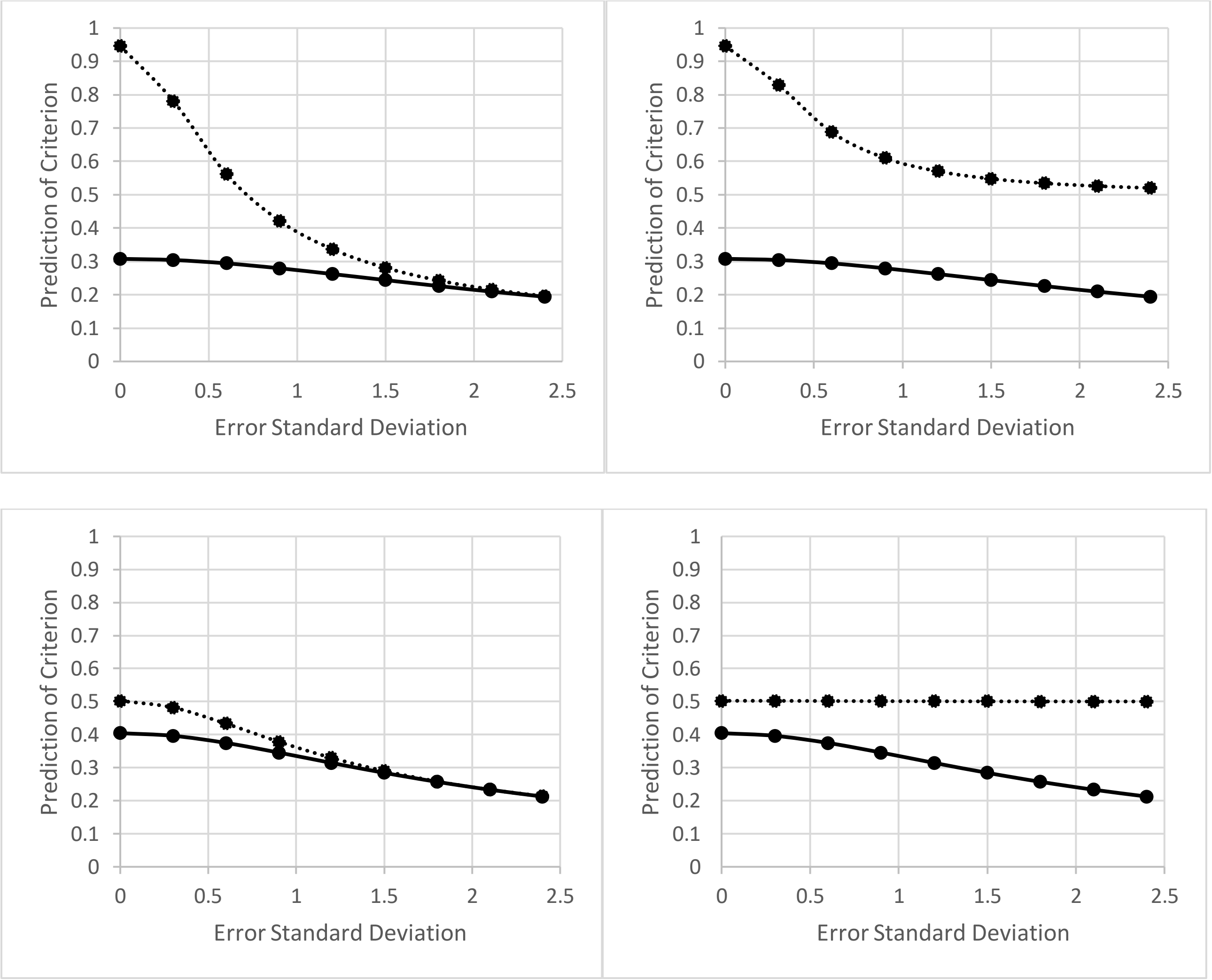

Figure 3. Prediction of criterion ranges along the vertical axis as a function of varying the error standard deviation, where Item 1 correlates with the criterion at 0.50 and Item 2 correlates with the criterion at 0.10. The dotted curve represents unamalgamating and the solid curve represents amalgamating. In the uppermost panels, the true interitem correlation coefficient is set at 0.90 and in the bottommost panels, the true interitem correlation coefficient is set at 0.10. In the leftmost panels, the Item 1 error standard deviation was allowed to vary, keeping the Item 2 error standard deviation at 0, whereas in the rightmost panels, the Item 2 error standard deviation was allowed to vary, keeping the Item 1 error standard deviation at 0.

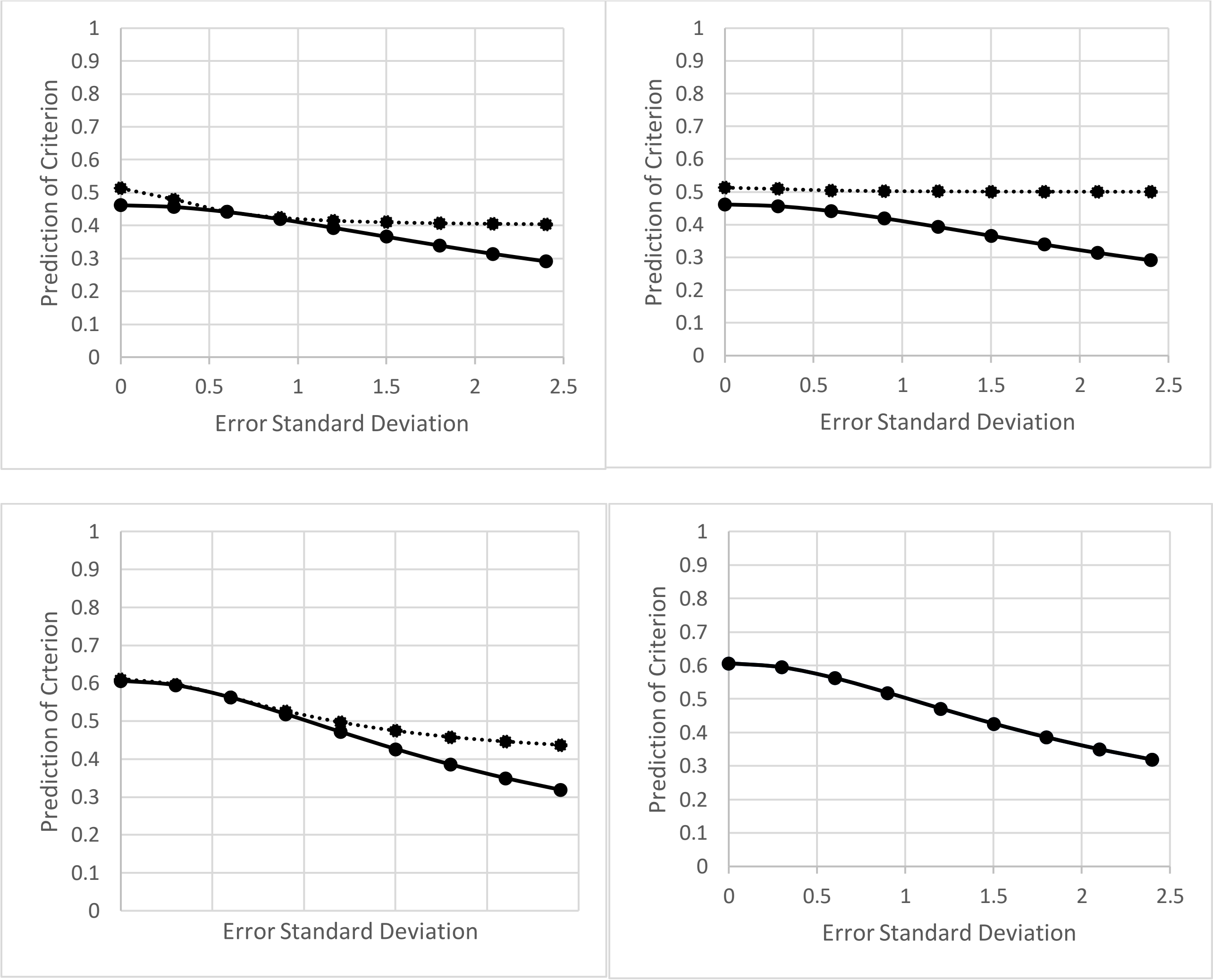

Figure 4. Prediction of criterion ranges along the vertical axis as a function of varying the error standard deviation, where Item 1 correlates with the criterion at 0.50 and Item 2 correlates with the criterion at 0.40. The dotted curve represents unamalgamating and the solid curve represents amalgamating. In the uppermost panels, the true interitem correlation coefficient is set at 0.90 and in the bottommost panels, the true interitem correlation coefficient is set at 0.10. In the leftmost panels, the Item 1 error standard deviation was allowed to vary, keeping the Item 2 error standard deviation at 0, whereas in the rightmost panels, the Item 2 error standard deviation was allowed to vary, keeping the Item 1 error standard deviation at 0.