International Journal of Aviation Research

Volume 11 | Issue 01 Peer-Reviewed Article

Drawing Conclusions about Reliability Without Measuring It

David Trafimow
New Mexico State University

The importance of having reliable measures in science research is obvious and cannot be overestimated. But in aviation research it often is impractical to collect test-retest reliability data because adding sessions can be expensive. What can be done? Based on classical measurement theory, the present article offers a procedure whereby researchers can draw limited conclusions about reliability even in the absence of reliability data.

Recommended Citation:
Trafimow, D (2019). Drawing Conclusions about Reliability Without Measuring It. *International Journal of Aviation Research, 11*(01), 1-10.

It is a truism that reliability sets an upper limit on validity. The truism is a direct consequence of a classical measurement theory equation that describes the relation between reliability and validity (Spearman, 1904). Because of the truism, it is not surprising that many journals in the social sciences require authors to collect reliability data and report the reliabilities of their measures. And yet, it sometimes is not feasible for researchers to collect reliability data. For example, consider an aviation researcher whose main dependent measure is performance on a flight simulator in various contexts of interest. Insisting that participants undergo a second session to facilitate a reliability computation may dramatically increase the cost of the experiment, rendering it unfeasible to perform. The present goal is to provide aviation researchers with a procedure to draw limited conclusions about reliability, even in the absence of reliability data.

Before continuing, it is important not to confuse the present use of the term "reliability," with Cronbach's alpha or its precursors, that also are commonly considered to index reliability. Cronbach's alpha is influenced by the number of items and inter-item correlations, thereby causing many to term it "internal consistency," rather than

reliability.[1] However, it also is well-known that items that load on different factors in a factor analysis nevertheless can be combined into a single set of items with high Cronbach's alphas. Thus, it is not necessarily clear that Cronbach's alpha even measures internal consistency, at least not in the sense that the items compose a single dimension. For present purposes, it is not necessary to delve into these complexities and issues pertaining to inter-item correlations will not be mentioned again. Instead, we can proceed directly to consider how classical measurement theory implies an ability to draw limited conclusions about reliability even without a direct reliability measure such as test-retest reliability. The first step in this direction involves a discussion of the classic attenuation and dis-attenuation equations, with implications. The implications are followed up subsequently with illustrative examples, extensions, and a discussion.

Attenuation, Dis-Attenuation, True and Observed Correlations

Let us commence with the famous attenuation formula from classical measurement theory listed below as Equation 1.[2] [3] In Equation 1, $\rho_{T_X T_Y}$ is the correlation between true scores, uncontaminated by random measurement error (the "true correlation" that the researcher really wants to know); ρ_{XY} is the correlation between observed scores (what the researcher obtains); and $\rho_{XX'}$ and $\rho_{YY'}$ are the reliabilities of measure *X* and measure *Y*, respectively.

$$\rho_{XY} = \rho_{T_X T_Y}\sqrt{\rho_{XX'}\rho_{YY'}} \quad (1)$$

Before continuing, it is worthwhile to pause for a moment to consider an implication of Equation 1 for the importance of reliability. If the reliabilities are perfect—that is, equal to 1.00—there is no random measurement error, and the true and observed correlations equal each other. This is the ideal case. In contrast, if the reliabilities equal zero, the correlation that the researcher will observe also will equal zero, even if the true correlation is perfect. The closer the reliabilities of the variables are to 1.00, the better the observed correlations will reflect the true correlations. Hence, it is difficult to overstate the importance of reliability (Spearman, 1904; Gulliksen, 1987; Lord & Novick, 1968; Trafimow, 2016; Trafimow, 2018).

Algebraic rearrangement of Equation 1 implies Equation 1*, which often is called the dis-attenuation formula.

$$\rho_{T_X T_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}} \quad (1^*)$$

Equation 1* shows how to "correct" for unreliability to estimate what the true correlation between the variables would be in the absence of random measurement error. Equation

[1] There are precursors, such as split-half reliability, that also depend on inter-item correlations. The present comments pertaining to Cronbach's alpha also apply to split-half reliability.
[2] Sometimes classical measurement theory is called classical test theory or classical true score theory.
[3] Equation 1 came originally from Spearman (1904); but was later expanded into the classical theory. Gulliksen (1987) and Lord and Novick (1968) provided well-cited reviews.

1* is widely used in the social sciences, and many statistical packages use it in structural equation modeling because models based on estimates of true correlations are much more accurate than models based on observed correlations.

For present purposes, it is convenient to consider the product of the reliabilities and to let *PROD* denote that product ($PROD = \rho_{XX'}\rho_{YY'}$). To provide context, imagine that the two measures have the same reliability. Well, then, if *PROD* is 0.49, it implies that the individual reliabilities are 0.70. If *PROD* is 0.64, it implies that the individual reliabilities are at the 0.80 level; and if *PROD* is 0.81, it implies that the individual reliabilities are at the 0.90 level. Of course, if the reliabilities of the two measures differ, and one of them is lower than, say, 0.70; the other would have to be greater than 0.70 to balance out the lower number. In line with typical pronouncements of 0.70 as being a lower limit of acceptable reliability, let us arbitrarily designate 0.49 as the threshold for an "acceptable" value for *PROD* ($0.70 \cdot 0.70 = 0.49$). Rewriting Equation 1* with the substitution of *PROD* for the reliabilities of the two measures renders Equation 2.

$$\rho_{T_X T_Y} = \frac{\rho_{XY}}{\sqrt{PROD}} \tag{2}$$

From here, we can perform algebra in easy steps to obtain an equation that renders *PROD* as a function of the correlation between true scores and the correlation between observed scores. First, let us multiply both sides of Equation 2 by $\sqrt{PROD}$ and divide both sides by $\rho_{T_X T_Y}$ to obtain Equation 3.

$$\sqrt{PROD} = \frac{\rho_{XY}}{\rho_{T_X T_Y}} \tag{3}$$

Squaring both sides of Equation 3 renders Equation 4.

$$PROD = \left(\frac{\rho_{XY}}{\rho_{T_X T_Y}}\right)^2 \tag{4}$$

Figure 1

Figure 1 illustrates values for $PROD$ along the vertical axis as a function of the observed correlation along the horizontal axis. Each curve represents different assumed values for the true correlation between the measures. The worst-case scenario for reliability is to assume that the true correlation maxes out at 1.00 (the solid curve in Figure 1), thereby resulting in the lowest possible values for $PROD$. Nevertheless, even this worst-case scenario does not have to be fatal for the researcher, for two reasons. First, and most convincingly, if the observed correlation is sufficiently large, even the worst-case scenario results in an acceptable result for *PROD*. Figure 1 shows this visually and instantiating values into Equation 4 provides a numerical demonstration. That is, if the observed correlation is 0.70, even assuming the true correlation is 1.00 gives the following: $PROD = \left(\frac{0.70}{1.00}\right)^2 = 0.49$. And, of course, if the true correlation is assumed to be anything less than 1.00, $PROD$ exceeds 0.49. For instance, suppose we assume that the true correlation is 0.90 instead of maximizing it at 1.00. In that case, $PROD = \left(\frac{0.70}{0.90}\right)^2 =$

0.60, a substantially more impressive value than the previous value of 0.49. In general, the smaller the value the researcher sets for the true correlation, the greater the value that will be obtained for *PROD*. The major point being made here is that even in the absence of reliability measures, researchers can draw conclusions about the lower limits of reliability. As we have just seen, an observed correlation of 0.70 implies that $PROD \geq 0.49$, an acceptable value. Of course, if the observed correlation is less than 0.70, the minimum possible value for *PROD* is less than 0.49. Table 1 provides necessary observed correlations so that *PROD* meets or exceeds the 0.49 standard, based on a variety of true correlations.

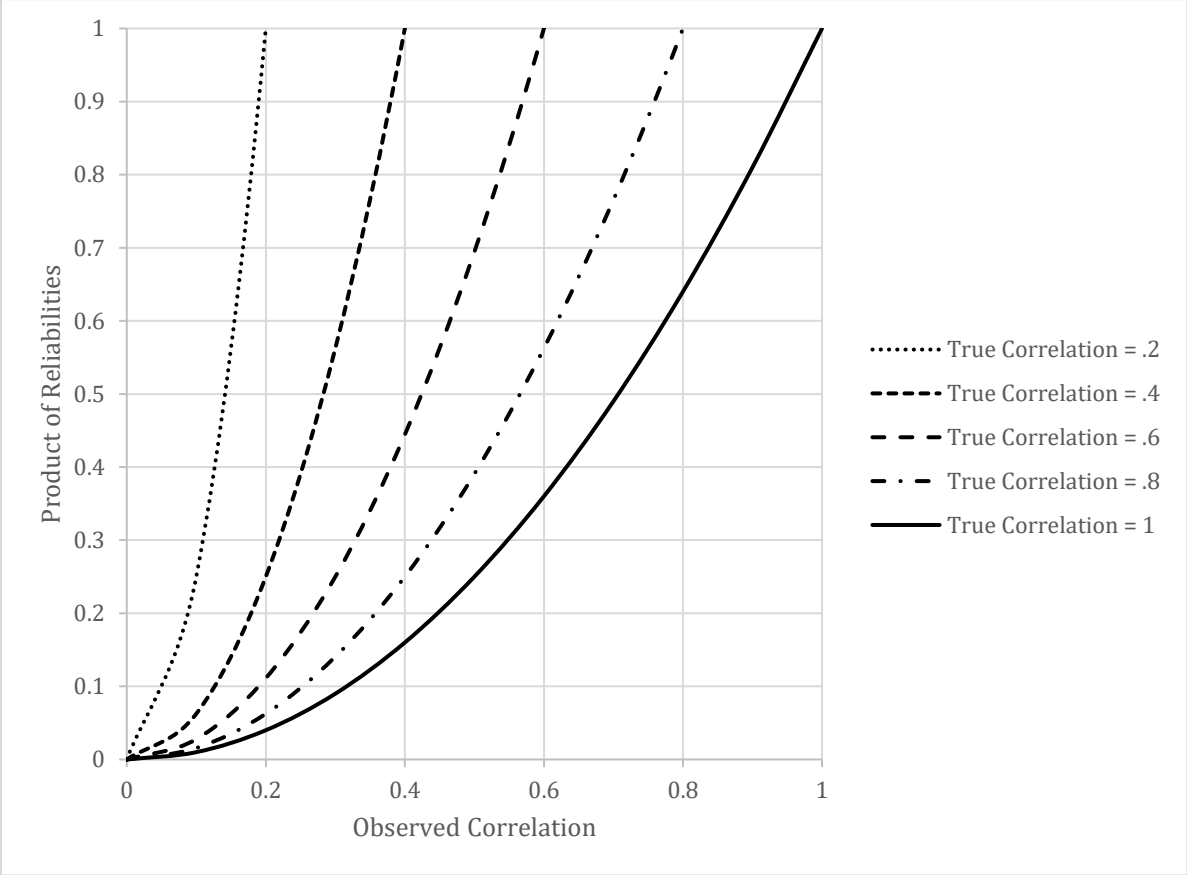

Figure 1. The product of the reliabilities (*PROD*) is expressed along the vertical axis as a function of the observed correlation expressed along the horizontal axis. The five curves illustrate the consequences for *PROD* of different values for the true correlation.

Table 1
Lower limit of the observed correlation necessary to indicate an acceptable value for *PROD* (0.49) as a function of the assumed value for the true correlation coefficient.

Assumed True Correlation	Necessary Observed Correlation for $PROD \geq .49$
0.10	0.07
0.20	0.14
0.30	0.21
0.40	0.28
0.50	0.35
0.60	0.42
0.70	0.49
0.80	0.56
0.90	0.63
1.00	0.70

The previous paragraph used an example where the observed correlation is 0.70, thereby implying that *PROD* meets or exceeds the acceptable value of 0.49 in the worst-case reliability scenario assuming a perfect true correlation. But what if the observed correlation is less than 0.70? Is this fatal for the researcher? Not necessarily, as the four examples in the subsequent section show.

Examples to Illustrate Relevant Arguments

To see the relevance of the forgoing mathematics for drawing scientific conclusions, consider the following examples. In each example, the sample obtained correlation r_{XY} is used to estimate the population observed correlation ρ_{XY}.[4]

Example 1

Suppose that an aviation researcher finds a correlation in fighter pilots between an index of various capabilities and performance on a simulator to be 0.70. The researcher wishes to use this as support that the abilities represented in the index importantly relate to fighter pilot performance. Unfortunately, the researcher did not have necessary funds available to include a second session for each participant; there was no way to calculate the reliabilities of the ability index or of fighter performance; and consequently, the researcher did not report them. A reviewer pounces on the lack of reliability data and recommends rejection: "How can we trust the data if the reliabilities of the measures are unknown, and might be extremely low?" The present mathematics suggests a response. That is, as we saw earlier, even if the true correlation is assumed to be 1.00, *PROD* would be at the acceptable level of .49, thereby indicating acceptable reliabilities of the measures. And this is the worst-case scenario for reliability. As Figure 1 shows, assuming more realistic values for the true correlation implies even better values for *PROD*.

Example 2

Suppose that instead of a correlation of 0.70, the researcher in the foregoing example obtained a correlation of only 0.45? Is she sunk? Not necessarily. One question that would need to be asked is: What true correlation, if it were known, would be necessary for the result to be considered important enough for publication? Suppose, for example, that a true correlation of 0.40 would be considered the lower limit of the necessary effect size to meet the reviewer's or editor's threshold for importance. Well, then, it is obvious that with an obtained correlation coefficient of 0.45, the researcher has exceeded the threshold value of 0.40. Remembering an implication of Equation 4 that all unreliability can do is decrease correlations, it is obvious that the reliabilities of the measures were sufficiently large to allow the researcher to exceed the threshold, thereby addressing the criticism.

Put more quantitatively, we can consider the worst-case scenario for the true correlation, which is that it equals the observed correlation. Remember that, in general,

---

[4] Sometimes statisticians place a "hat" on the estimate of the population observed correlation to indicate that it is a sample statistic used to estimate a population parameter, rather than the population parameter itself: $\hat{\rho}_{XY}$. It seems more straightforward to simply use the standard symbol for the sample correlation: r_{XY}. Of course, $r_{XY} = \hat{\rho}_{XY}$.

true correlations exceed observed correlations because true correlations are not subject to random measurement error whereas observed correlations are. Thus, the true correlation cannot be a lower value than the observed correlation. Well, then, if the true correlation is 0.45, Equation 3 implies the following: $PROD = \left(\frac{0.45}{0.45}\right)^2 = 1.00$. In other words, the worst-case scenario for the true correlation implies the best-case scenario for the reliabilities of the measures, thereby negating the criticism that the reliabilities might be low. Alternatively, if the reviewer wishes to make the worst-case scenario for reliability, it is necessary to make the best-case scenario for the true correlation; namely, the true correlation would be set at 1.00. But note that although this renders *PROD* a poor number $\left(PROD = \left(\frac{0.45}{1.00}\right)^2 = 0.20\right)$, which seems bad for the researcher at first glance because 0.20 is less than our arbitrary value of 0.49; it gives away the ballgame by admitting that the true correlation is a perfect value of 1.00. Thus, either way, the intellectually honest reviewer would have to admit that the researcher wins.

Example 3

Suppose that instead of an observed correlation of 0.70 or 0.45, the obtained value is only 0.29. Remaining with a threshold value of 0.40 for a true correlation to be considered sufficiently important for publication, the argument made in Example 2 will not work. Is the researcher sunk in this case? Here we have more of a judgment call. To see this, let us suppose that the true correlation really is exactly at the importance threshold of 0.40. In that case, Equation 3 implies the following, if we assume that the threshold value really is true: $PROD = \left(\frac{0.29}{0.40}\right)^2 = 0.53$. Clearly, then, the reliabilities of the measures are acceptable under an assumption that the true correlation is 0.40, and if we assume lower values for the true correlation, the *PROD* score would be even better, as Figure 1 illustrates. Ironically, then, the intellectually honest reviewer would have to admit that a complaint about the reliability of the measures does not make sense in this context. This is not to say that the intellectually honest reviewer must recommend acceptance. The reviewer could suggest that, in fact, the measures are very reliable, though unmeasured; but that the foregoing mathematics force, under this assumption, that the true score fails to exceed the importance threshold of 0.40. To see the extreme case of this argument, consider that if *PROD* is set at 1.00 (perfect reliability), the observed correlation of 0.29 would imply that the true correlation also is 0.29, which would be well under the importance threshold of 0.40. Ironically, however, this reviewer criticism would be contrary to the criticism that the reliabilities might be low because it assumes high reliabilities.

Example 4

Let us consider the same values as in Example 3; but change the nature of the argument. That is, instead of wishing to demonstrate that the ability index is importantly related to performance; suppose that the researcher wishes to show that the relationship is

unimportant. In this case, if the measures are unreliable, the implication is that the true correlation would be above the importance threshold. For example, using Equation 2, suppose we assume that $PROD$ is .35. In that case, remaining with the obtained correlation coefficient of 0.29, the true correlation would be estimated to be as follows: $\rho_{T_X T_Y} = \frac{.29}{\sqrt{.35}} = 0.49$. This value is well above the importance threshold of 0.40. Thus, Example 4 provides a valuable contrast to Examples 1-3. That is, in Examples 1-3, where the goal was to show a strong relationship between the two variables, arguments about unreliability merely strengthen that the true correlation must be strong, if the observed correlation coefficient is a reasonable but not necessarily impressive value. Or alternatively, arguments about low true correlations merely strengthen that the reliabilities of the variables must be strong, again depending on a reasonable value for the observed correlation. In contrast, Example 4 illustrates that if the researcher wishes to establish that two measures are not very related, it is entirely reasonable for a critical reviewer to demand evidence that the measures have high reliability. In Example 4, unlike Examples 1-3, the omission of reliability measures is devastating.

Extension to Experimental Research

It is not difficult to extend the foregoing correlational approach to true experiments with experimental and control conditions. Consider that the typical effect size measure in experiments is Cohen's *d*.[5] Once the researcher has obtained Cohen's *d*, it can be converted into a correlation coefficient using Equation 5 below (Rosenthal & Rosnow, 2008).

$$r_{XY} = \sqrt{\frac{d^2}{d^2+4}} \tag{5}$$

In turn, once Equation 5 has been used to convert Cohen's *d* into an observed correlation coefficient, the procedure described earlier applies directly. For example, suppose that the experimenter obtains a value of 0.90 for Cohen's *d*; Equation 5 implies that the observed correlation between the independent variable and the dependent variable is as follows: $r_{XY} = \sqrt{\frac{0.9^2}{0.9^2+4}} = 0.41$. This value can be instantiated into Equation 4.

With the issue of generalization to experimental paradigms addressed by Equation 5, there remains a conceptual asymmetry to be discussed. Imagine that Researcher A obtains a large effect size whereas Researcher B obtains a very small effect size. All else being equal, should Researcher A be given preference over Researcher B in the publication process? There has been considerable discussion recently about precisely this issue; that is, whether the size of the obtained effect should be an important consideration in reviewer recommendations and editorial decisions (e.g., Grice, 2017; Hyman, 2017;

[5] Cohen's *d* is the difference in means divided by the pooled standard deviation: $d = \frac{M_1 - M_2}{\sigma_p}$. Most statistical packages will compute it automatically or with an extra click of the mouse.

Kline, 2017; Locascio, 2017a, 2017b; Marks, 2017). Although there is insufficient space here to discuss the issue properly, the foregoing examples can be argued to militate on the side of those who believe that the obtained effect size should matter in editorial decisions.

To understand why, let us return to Researcher A who obtains a large effect size. Equation 4 proves that this researcher's measures must have been reliable—though there is no way to know their exact reliability—and so lack of reliability cannot reasonably be used as an argument against publication. In contrast, consider Researcher B, who obtained a small effect size. In this case, there is no way to counter the accusation that the measures were unreliable, thereby causing the small obtained effect. Clearly, all else being equal, Researcher A is in a much better position than Researcher B with respect to the reliability issue.

Of course, as was suggested earlier, the effect size also may matter from the point of view of application. For example, if a new pilot training procedure results in a large improvement over the old one, and the finding replicates, that would provide a strong reason for converting to the new procedure. In contrast, if there is little difference between the two conditions, the implications for applications are less clear.

Conclusion

In conclusion, the lack of feasibility of reliability measures in many aviation research paradigms need not be a fatal flaw. We have seen that Equation 4 (augmented by Equation 5 for experimental designs) provides a way for researchers to come to limited conclusions about reliability, even when reliability cannot be directly assessed. The larger the obtained value for the sample correlation or for Cohen's *d*, the larger the lower limit for the reliabilities of the measures. Furthermore, even if the obtained effect size is moderate, rather than large; if expert knowledge in the substantive area implies that the true effect size likely is moderate, Examples 2 and 3 suggest that a reasonable case may nevertheless be made that the measures are sufficiently reliable. The procedure recommended here is not a panacea, as Example 4 illustrates; but it does provide researchers with an ability to come to limited conclusions about reliability, in the absence of reliability data. Depending on the obtained effect size as well as expert knowledge of the substantive domain of interest, the proposed procedure may help justify research that otherwise would be more difficult to justify.

References

Grice, J. W. (2017). Comment on Locascio's results blind manuscript evaluation proposal. *Basic and Applied Social Psychology* 39: 254–255. doi: 10.1080/01973533.2017.1352505

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

Hyman, M. (2017). Can 'results blind manuscript evaluation' assuage 'publication bias'? *Basic and Applied Social Psychology*, *39*(5), 247-251. https://doi.org/10.1080/01973533.2017.1350581

Kline, R. (2017). Comment on Locascio, results blind science publishing. *Basic and Applied Social Psychology*, *39*(5), 256–257. https://doi.org/10.1080/01973533.2017.1355308

Locascio, J. (2017a). Results blind publishing. *Basic and Applied Social Psychology*. *39*(5), 239-246. https://doi.org/10.1080/01973533.2017.1336093

Locascio, J. (2017b). Rejoinder to responses to "results blind publishing." *Basic and Applied Social Psychology*. *39*(5), 258-261. https://doi.org/10.1080/01973533.2017.1356305

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

Marks, M. J. (2017). Commentary on Locascio 2017. *Basic and Applied Social Psychology*. *39*(5), 252–253. https://doi.org/10.1080/01973533.2017.1350580

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis*. Boston, M. A.: McGraw-Hill.

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

Trafimow, D. (2016). The attenuation of correlation coefficients: A statistical literacy issue. *Teaching Statistics*, *38*(1), 25-28. doi: 10.1111/test.12087

Trafimow, D. (2018). The importance of reliability in clinical research. *Timely Topics in Clinical Immunology*, *2*(1), 1. http://www.alliedacademies.org/articles/The%20importance%20of%20reliability%20in%20clinical%20research.pdf