Acta Scientific Neurology (ISSN: 2582-1121)

Volume 4 Issue 8 August 2021

Research Article

Revisiting Old-Fashioned Reliability and Validity Concerns

David Trafimow*

Department of Psychology, New Mexico State University, Mexico

***Corresponding Author:** David Trafimow, Department of Psychology, New Mexico State University, Mexico.

Received: April 09, 2021

Published: July 31, 2021

Abstract

Reported correlations in psychology research tend to be unimpressive. This would not be a problem if the underlying reason were that the phenomena under investigation really were not very related. However, a more troubling explanation pertains to the reliability and validity of the measures. As has been known since the seminal research by Spearman (1904), reliability sets an upper limit on predictive validity; unreliable measures result in unimpressive correlations even if all else is right. The present article briefly reviews the old literature on classical true score theory with an eye towards (a) reiterating long-known but rarely attended to prescriptions for obtaining more impressive correlations, (b) drawing lessons that contradict cliches in the field, and (c) expanding classical true score theory wisdom to cases where there are two predictor variables rather than a single amalgamated variable.

Keywords: Correlation; Multiple Correlation; Classical True Score Theory; Reliability; Validity

To anyone who has read broadly in substantive literatures that feature correlation coefficients, it is impossible to fail to notice that the correlation coefficients are usually low (in the 0.1 to 0.4 range), though there are exceptions. One such exception might be work in the theory of reasoned action tradition [1], where multiple correlations to predict behavior have tended to be around 0.7 or higher, nor is this merely a recent trend [2]. But one of the features separating work in the theory of reasoned action tradition from other correlational work has been the careful attention paid to measurement [3]. This is not to say that all correlational work should be performed like work in the theory of reasoned action tradition; however, that work exemplifies the gains that can be made by careful attention to measurement. The present goal is to consider two old-fashioned measurement concerns: validity and reliability—and recommend how a serious consideration of these issues could result in substantially improved correlations. No new basic material will be presented, though there will be new inferences and demonstrations.

Very quick review of basic psychometrics

The present section briefly reviews validity and reliability. The review is far from exhaustive but serves to remind the reader of two points. First, reliability sets an upper limit on validity. Second, reliability, in turn, is determined by (a) interitem correlations and (b) number of items.

The classic attenuation equation

Charles Spearman [4] derived Equation 1 below that explains how the reliability of the measures sets a limit on predictive validity, the degree to which the two variables can correlate with each other [5,6]:

$$r_{XZ} = r_{T_X T_Z}\sqrt{r_{XX'}r_{ZZ'}}; \text{ ---------(1)}$$

Where r_{XZ} is the correlation between observed scores, $r_{T_XT_Z}$ is the correlation between true scores (or the true correlation or the correlation that would be obtained in the absence of random error), and $r_{XX'}$ and $r_{ZZ'}$ are the reliabilities of the two measures. One way to understand Equation 1 is to consider the extremes. At one extreme, imagine that $r_{XX'} = r_{ZZ'} = 1$. In that case, the observed correlation (predictive validity) would equal the true correlation which is a best-case scenario. At the other extreme, imagine that $r_{XX'} = 0$ or that $r_{ZZ'} = 0$. In that case, it would not matter what the true correlation would be, validity would be 0.

Another way to understand Equation 1 is to combine the reliabilities of both measures into a reliability product: $\text{Prod} = r_{XX'}r_{ZZ'}$. In that case, Equation 1 reduces to Equation 2:

$$r_{XZ} = r_{T_XT_Z}\sqrt{\text{Prod}}. \text{ ----------(2)}$$

Equation 2 is useful for drawing figure 1, where the observed correlation is expressed along the vertical axis, as a function of the product of the reliabilities along the horizontal axis, with different curves corresponding to different true correlations. Both Equation 2 and figure 1 show that (a) validity cannot exceed the true correlation and (b) validity cannot exceed the square root of the product of the reliabilities. Thus, the importance of reliability for validity is obvious, thereby bringing up the issue of how one obtains impressive reliability.

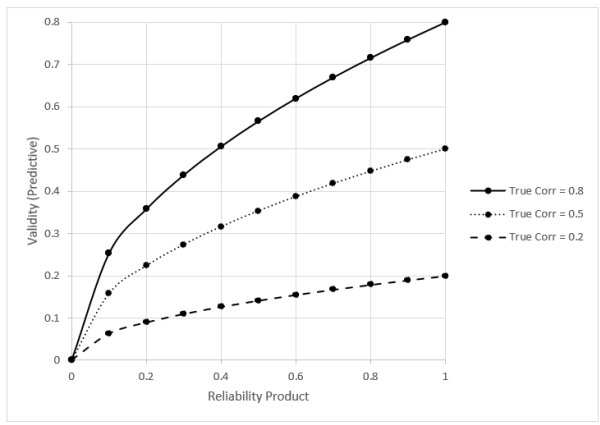

Figure 1: Validity is expressed along the vertical axis as a function of the reliability product along the horizontal axis, with curves representing the true correlation equaling 0.80 (top curve), 0.50 (middle curve), and 0.20 (bottom curve).

Cronbach's alpha

Although there are many reliability formulas, Cronbach's alpha is easily the most common and will be featured here [7], though other reliabilities indices are slightly superior but more complex [8,9]. In its usual form, Cronbach's alpha is expressed as Equation 3 below:

$$\alpha_{\text{standardized}} = \frac{K\bar{r}}{1+(K-1)\bar{r}}; \text{ ----------(3)}$$

Where *K* refers to the number of units (hereafter, these are test items) in the test and $\bar{r}$ refers to the average interitem correlation. Equation 3 shows that reliability can be increased by (a) having more items and (b) increasing the similarity between items to increase interitem correlations. Figure 2 illustrates the consequences of Equation 3, with reliability ranging along the vertical axis as a function of the average interitem correlation ranging across the horizontal axis, with different curves for tests with 2, 4, 8, or 16 items.

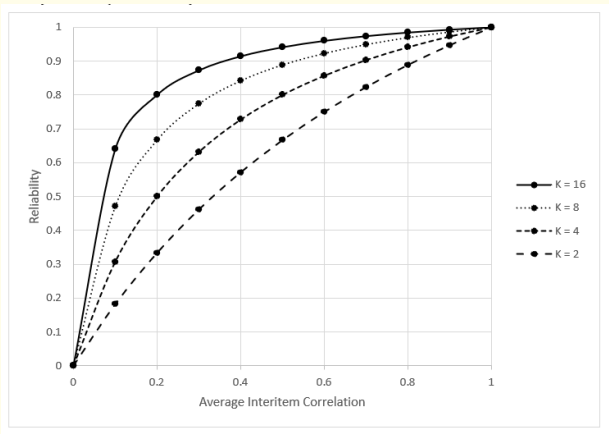

Figure 2: Reliability is expressed along the vertical axis as a function of the average interitem correlation along the horizontal axis, with curves representing when the test comprises 16 items (top curve), 8 items (second curve), 4 items (third curve), or 2 items (bottom curve).

Implications of number of items and interitem correlations for validity

It is not difficult to combine the implications of Spearman's (1904) equation and Cronbach's (1951) equation (Equations 1 and

3). Using Cronbach's alpha as our reliability index, Equation 3 indicates that $r_{XX'} = \frac{K_X \bar{r}_X}{1+(K_X-1)\bar{r}_X}$ and that $r_{ZZ'} = \frac{K_Z \bar{r}_Z}{1+(K_Z-1)\bar{r}_Z}$. Instantiating these equations, in turn, into Equation 1 renders Equation 4:

$$r_{XZ} = r_{T_X T_Z}\sqrt{\frac{K_X \bar{r}_X}{1+(K_X-1)\bar{r}_X} \cdot \frac{K_Z \bar{r}_Z}{1+(K_Z-1)\bar{r}_Z}} \text{ -------- (4)}$$

An advantage of Equation 4 is that it expresses validity (r_{XZ}) directly as a function of the number of items and the average interitem correlation, with respect to each measure. In addition, for the sake of illustration, let us make the simplifying assumptions that the number of items is the same for both tests ($K_X = K_Z = K$) and that the average interitem correlation is the same for both tests ($\bar{r}_X = \bar{r}_Z = \bar{r}$), which renders Equation 5:

$$r_{XZ} = r_{T_X T_Z} \frac{K\bar{r}}{1+(K-1)\bar{r}} \text{ --------- (5)}$$

Although Equation 5 is a blatant oversimplification, an advantage is that it renders figure 3 possible, which illustrates how both the number of items and the average interitem correlation influence validity as a proportion of the true correlation. As the average interitem correlation increases, and as the number of items increases, validity can be expected to be an increasingly larger proportion of the true correlation. And there are implications.

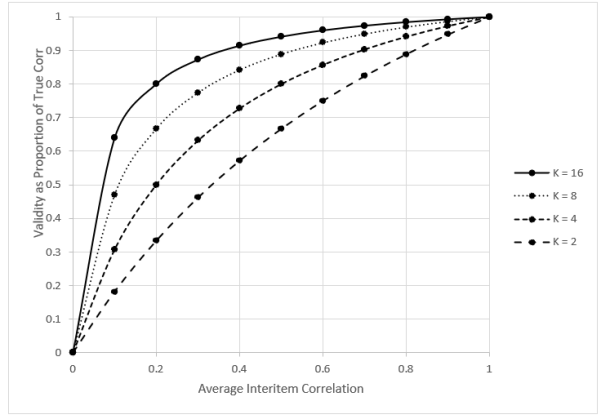

Figure 3: Validity as a proportion of the true correlation is expressed along the vertical axis as a function of the average interitem correlation along the horizontal axis, with curves representing when the test comprises 16 items (top curve), 8 items (second curve), 4 items (third curve), or 2 items (bottom curve).

Implication 1: Increase the number of items

The most obvious implication is that researchers can increase reliability—and validity—simply by having more items, even if the individual items are not very good items. For example, figure 3 shows that even if the average interitem correlation is at the very low level of 0.20, validity will be 80% of the true correlation provided that there are 16 items. Thus, it is possible to substantially improve obtained correlations merely by having longer tests.

To see that this is not trivial, consider the longstanding trait-situation debate in personality psychology. Based on the personality literature in the 1960s, Mischel [10] showed that personality rarely correlates at more than the 0.3 level with behavior, implying a 10% ceiling on the extent to which personality could predict behavior, with the further implication that perhaps personality is not of particular importance. There were many responses to Mischel's critique, but one of the most successful responses was to simply include more items in personality tests, which increased reliability, with a knock-on effect of increasing validity. Instead of obtaining correlations with a ceiling of 0.3, researchers who used longer tests were able to extend to a ceiling of 0.4 and sometimes even more than that [11]. Thus, instead of personality only accounting for 10% of the variance, the value increased to around 15%, an approximately 50% improvement, and rivaling the ability of situations to predict behaviors [12]. To be sure, these researchers did not explicitly use Equation 4 or Equation 5, but the action of increasing test lengths is very consistent with the implications of those equations. It was beneficial for personality psychologists to increase test lengths.

Before proceeding, however, it is necessary to acknowledge limitations. One limitation is that increasing test lengths implies similarly increasing participant time, cost, boredom, exhaustion, and so on. Thus, there often are practical reasons why researchers do not wish to increase test lengths. Another limitation is that the ease with which tests lengths can be increased can sometimes lead to temptation to pay too little attention to the exact nature of the items themselves. An example is the case where a person has measures of two correlated constructs rather than one, but nevertheless obtains an impressive reliability coefficient via the combined test. In this case, it would be better to keep the two tests separate, though paying attention only to reliability might seem to indicate otherwise. Figure 3 illustrates this whereby even if the average in-

teritem correlation only equals 0.20, having 16 items renders the validity coefficient at 80% of the true correlation. We will explore this issue in some detail later.

Implication 2: Have items that are synonyms

As figure 3 shows that having strong interitem correlations increases reliability and validity dramatically, even with short tests, the obvious thing to do is to include items that are synonyms of each other. If the items are synonyms, the result will be strong interitem correlations, and hence, validity will be a correspondingly impressive percentage of the true correlation. Figure 3 shows, for instance, that if the average interitem correlation is 0.80, then even with only two items, validity will be 89% of the true correlation.

The obvious rejoinder to the recommendation to use synonyms is that the researcher risks failing to cover all of the construct. But there is a rejoinder to the rejoinder, which is that if different kinds of items, that are not synonyms, are needed to cover all of the construct, then perhaps the construct is really an amalgamation of two or more constructs and the researcher has failed to see it. More than that, the mere fact that the researcher feels the need to have different kinds of items, that are not synonyms, indicates that it is very likely that the researcher is unknowingly amalgamating different constructs into one construct. There are at least two good reasons for remaining with synonyms as a way to avoid amalgamation and keep separate constructs distinct. The first reason is conceptual clarity. From the reasoned action literature cited earlier, there is a variable termed "perceived behavioral control" with items mentioning capability to perform the behavior, and synonyms; but with items mentioning difficulty in performing the behavior, and synonyms, too. Trafimow., *et al*. [13] suspected that perceived behavioral control was really an amalgamation of two constructs, that they termed "perceived control" and "perceived difficulty." Not only did keeping the constructs separate confer psychometric advantages, but Trafimow., *et al*. showed it was possible to perform experimental manipulations that influenced perceived control without influencing perceived difficulty, and to perform experimental manipulations that influenced perceived difficulty without influencing perceived control. By experimentally demonstrating a double dissociation, Trafimow., *et al*. clarified that there really were two different—though correlated—constructs and distinguishing between them constituted an important contribution to the literature.

Apart from true experimentation, another advantage of keeping separate constructs distinct, even if amalgamation confers acceptable reliability, is that predictive validity can be increased. But this issue deserves its own section, that ensues immediately.

Reliability and validity the multiple correlation way

Consider again two correlated constructs. For the sake of simplification, imagine that the two constructs are related such that the reliability of an amalgamated measure equals the reliability of each of the separate measures. That is, when the measures are kept separate, interitem correlations are stronger and compensate for having fewer items. When the measures are amalgamated, having more items compensates for having smaller average interitem correlations. From a strict reliability standpoint, there is no reason to prefer two constructs to a single amalgamated construct, when all are equally reliable. But of course, from the point of view of conceptual clarity, construct validity, and simply having a correct theory, it would be better to keep the constructs distinct. But suppose we do not care about conceptual clarity, construct validity, or having a correct theory; but simply care about the ability to predict the criterion variable. It might seem that with this restriction, it would be fine to amalgamate and even desirable from a parsimony standpoint. But appearances can be deceiving.

There is a well-known multiple correlation equation that shows the ability to predict a variable from two other variables, expressed below as Equation 6 [14]:

$$R_{y.12} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}}; \text{ ---------(6)}$$

Where $R_{y.12}$ is the multiple correlation for predicting a criterion variable, represented as , from two predictor variables, represented as 1 and 2, respectively. In addition, r_{y1} represents the correlation between the criterion variable and one of the predictor variables, r_{y2} represents the correlation between the criterion variable and the other predictor variable, and r_{12} represents the correlation between the two predictor variables. In turn, these component correlations can be expressed in the form of true correlations and reliabilities consistent with Equation 1:

- $r_{y1} = r_{T_y T_1}\sqrt{r_{yy}r_{11'}}$,
- $r_{y2} = r_{T_y T_2}\sqrt{r_{yy}r_{22'}}$,

- $r_{12} = r_{T_1T_2}\sqrt{r_{11'}r_{22'}}$,
- $r^2_{y1} = r^2_{T_yT_1}r_{yy'}r_{11'}$,
- $r^2_{y2} = r^2_{T_yT_2}r_{yy'}r_{22'}$,
- $r^2_{12} = r^2_{T_1T_2}r_{11'}r_{22'}$

Instantiating the bullet-pointed equations into Equation 6 implies Equation 7:

$$R_{y.12} = \sqrt{\frac{r^2_{T_yT_1}r_{yy'}r_{11'} + r^2_{T_yT_2}r_{yy'}r_{22'} - 2r_{T_yT_1}\sqrt{r_{yy'}r_{11'}}r_{T_yT_2}\sqrt{r_{yy'}r_{22'}}r_{T_1T_2}\sqrt{r_{11'}r_{22'}}}{1 - r^2_{T_1T_2}r_{11'}r_{22'}}} \quad \text{----(7)}$$

Equation 7 has the desirable characteristic of including the reliabilities of all the measures, but it has the undesirable characteristic of being too complex to be susceptible of clear illustration via a figure. At the risk of oversimplification, let us assume that all true correlations equal each other, and all reliabilities equal each other. This renders the following components:

- $r_T = r_{T_yT_1} = r_{T_yT_2} = r_{T_1T_2}$,
- $r_{rel} = r_{yy'} = r_{11'} = r_{22'}$,
- $r^2_T = r^2_{T_yT_1} = r^2_{T_yT_2} = r^2_{T_1T_2}$.

Instantiating the bullet-listed components into Equation 7 and simplifying gives Equation 8, which is sufficiently simple to be susceptible to illustration:

$$R_{y.12} = \sqrt{\frac{2r^2_T r^2_{rel} - 2r^3_T r^3_{rel}}{1 - r^2_T r^2_{rel}}} \text{------ (8)}$$

Figure 4 illustrates the implications of Equation 8, but the Figure is complex and requires explanation. The validity coefficient ranges along the vertical axis as a function of reliability along the horizontal axis, but with six curves. Three of the curves are really straight lines (in gray) whereas the other curves really are curves (in black). The straight lines in gray represent validity in the amalgamated case where the criterion variable is being predicted from an amalgamated predictor variable (see Equation 1). In contrast, the curves in black represent validity in the form of a multiple correlation where the criterion variable is predicted from both predictor variables (see Equation 8). In addition, the curves represent different true correlations in a solid curve, a dotted curve, or a dashed curve. Note that each curve, in black, illustrates an increase in validity over the corresponding (same true correlation) curve, in gray. Thus, figure 4 illustrates that even if a researcher is willing to ignore issues such as conceptual clarity, construct validity, and simply having a correct theory; even from a restricted point of view only concerned with predictive validity, it still is sensible to un-amalgamate the constructs. And note that the differences between the black and gray curves would be greater still under a more realistic assumption about the true correlation between the two predictor variables. Specifically, to simplify Equation 7 to the point of obtaining Equation 8, it was necessary to make the unrealistic assumption that the true correlation between the predictor variables equals the true correlation between either of them with the criterion variable. A more realistic assumption, that the true correlation between the predictor variables is somewhat less than that, would increase the differences between the black and corresponding gray curves in figure 4, thereby accentuating the importance of un-amalgamating constructs, even from a standpoint strictly concerned only with prediction.

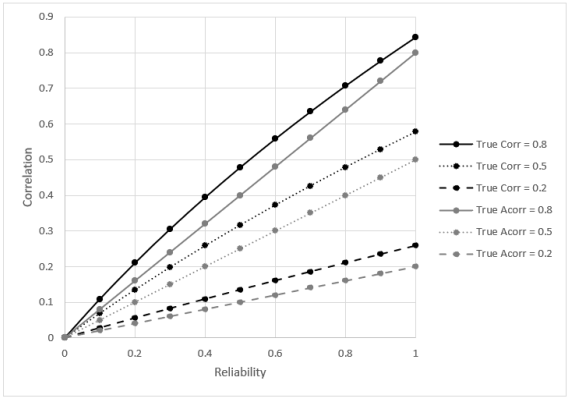

Figure 4: Predictive validity is expressed along the vertical axis as a function of reliability along the horizontal axis, with six curves representing different true correlations. The gray curves (straight lines) represent various states of the true correlation in the context of an amalgamated predictor variable (Acorr), so that predictive validity refers to a bivariate correlation. In contrast, the black curves represent various states of the true correlation in the context of two separate, though correlated, predictor variables, so that predictive validity refers to a multiple correlation.

Discussion and Conclusion

The unimpressive correlations psychologists typically obtain provided the original stimuli for the present work. Given the psychometric advances that were already in place half a century ago [5], and clear demonstrations in some empirical literatures mentioned here of the benefits a few researchers have enjoyed by taking them onboard, it is mystifying that unimpressive correlations continue to be the rule rather than the exception (e.g., Jussim [15], Table 6-1). Of course, psychometric advances continue to be made, and advances such as Equation 1 and Equation 3 are considered extremely old-fashioned by contemporary psychometricians. The point of using old-fashioned psychometric advances was not to advocate for their use, but to render salient that the advanced mathematical or computer skills for understanding recent advances in psychometrics are unnecessary for dramatically improving the quality of research. If research could be improved dramatically by careful attention to the psychometric lessons already in place half a century ago, how much more could research be improved by attending to modern psychometric advances too?

One explanation for unimpressive correlations might be the advices researchers get, that arguably contradict the psychometric lessons discussed here. For example, it is a cliché that reliability at the level of 0.70 or higher is acceptable. But consider two points. First, if reliability is 0.70, then squaring that value results in the measure only accounting for 49% of the variance in itself. This can hardly be considered impressive. Second, if two variables are measured, both with reliability at the 0.70 level, then the obtained correlation will only be 70% of the true correlation. Suppose that the true correlation is 0.5. Taking 70% of that will result in an obtained correlation of 0.35. And if the true correlation is the more realistic value of 0.40, the obtained correlation drops to 0.28. Worse yet, researchers fairly often report reliability values in the range between 0.6 and 0.7, and evaluate these as being close to the conventional 0.70 level of acceptability, thereby resulting in even more of a discrepancy between true and observed correlations. For example, 60% of a true correlation of 0.40 would be an observed correlation of only 0.24. In contrast, suppose that both measures are reliable at the 0.8 or 0.9 levels, so that the observed correlation is 80% or 90% of the true correlation. In that case observed correlations would be 0.32 or 0.36, respectively.

And there is no reason to settle for low reliabilities. The typical justification—nay, advice—that low reliability is a necessary consequence of covering all of the construct, has been an important detriment to psychological research. As the Trafimow., *et al*. [13] case described earlier demonstrates, when a researcher has to struggle to capture all of a construct, it is tantamount to certain that there is more than one construct and that amalgamation has occurred. The onus is on the researcher to carefully think through the construct and distinguish exactly the constructs that he or she is amalgamating, albeit unintentionally, to result in decreased reliability. Not only will un-amalgamating result in better interitem correlations, thereby resulting in increased reliability and validity; but doing so has the added benefits of introducing increased conceptual clarity, an increased likelihood of having a correct theory, and better prediction of the criterion variable.

This last point deserves amplification. In figure 4, we saw that even under the assumption that the amalgamated measure has the same reliability as the un-amalgamated measures, prediction of the criterion variable is better with the un-amalgamated measures than with the amalgamated measure. However, figure 4 provided an advantage to the amalgamated measure of increased items that compensated for the better interitem correlations in the un-amalgamated measures, to result in equivalent reliabilities. Thus, figure 4 shows that the un-amalgamated measures perform better than the amalgamated measure, even when subjected to an unfair disadvantage. Under the fairer condition that each of the un-amalgamated measures has the same number of items as the amalgamated measure, the un-amalgamated measures would have a reliability advantage, and the resulting increased ability to predict the criterion variable would be further enhanced, above and beyond the effect illustrated in figure 4.

In summary, impressive reliability is not a luxury; it is a necessity. The two cliches, that reliability at the 0.70 level is acceptable, and that researchers should make sure to cover all of the construct, both contradict the implications of classical true score theory, exemplified by Equations 1 and 3. Furthermore, both cliches are simply wrong. Researchers should insist on reliability at the 0.80, or even 0.90 level, thereby ensuring that obtained correlations are a higher percentage of true correlations. And researchers who struggle to capture all of the construct should carefully consider that the reason for the struggle is that the construct under consideration is really an amalgamation of two or more other constructs, and that un-amalgamation is necessary. If contemporary researchers would

take onboard the old-fashioned lessons that have existed for well over half a century, their obtained correlations would better represent the underlying true correlations. More generally, empirical correlations would better correspond to reality. In turn, having empirical correlations that better correspond to reality would aid both in theory-building and in theory-testing, two goals to which researchers ought to aspire.

Bibliography

1. Fishbein M and Ajzen I. "Predicting and changing behavior: The reasoned action approach". Psychology Press, New York, NY (2010).
2. Kraus SJ. "Attitudes and the prediction of behavior: A meta-analysis of the empirical literature". *Personality and Social Psychology Bulletin* 21 (1995): 58-75.
3. Fishbein M and Ajzen I. "Belief, attitude, intention and behavior: An introduction to theory and research". Addison-Wesley, Reading, MA (1975).
4. Spearman C. "The proof and measurement of association between two things". *American Journal of Psychology* 15.1 (1904): 72-101.
5. Lord F M and Novick M R. "Statistical theories of mental test scores". Reading: Addison-Wesley (1968).
6. Gulliksen H. "Theory of mental tests". Hillsdale: Erlbaum (1987).
7. Cronbach L J. "Coefficient alpha and the internal structure of tests". *Psychometrika* 16.3 (1951): 297-334.
8. Morera O F and Stokes M A. "Coefficient α as a measure of test score reliability: Review of 3 popular misconceptions". *AJPH Methods* 106.3 (1963): 458-461.
9. Trizano-Hermosilla I and Alvarado J M. "Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements". *Frontiers in Psychology* 7 (2016): 769.
10. Mischel W. "Personality and assessment". Mahwah, N. J.: Lawrence Erlbaum Associates (1968).
11. Epstein S. "The stability of behavior: I. On predicting most of the people much of the time". *Journal of Personality and Social Psychology* 37.7 (1979): 1097-1126.
12. Funder DC and Ozer DJ. "Behavior as a function of the situation". *Journal of Personality and Social Psychology* 44.1 (1983): 107-112.
13. Trafimow D., *et al*. "Evidence that perceived behavioral control is a multidimensional construct: Perceived control and perceived difficulty". *British Journal of Social Psychology* 41.1 (2002): 101-121.
14. Pedhazur E J. "Multiple regression in behavioral research: Explanation and prediction (3rd edition)". United States: Wadsworth (1997).
15. Jussim L. "Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy". Oxford University Press (2012).

Volume 4 Issue 8 August 2021