Running Head: Unamalgamating Multi-item Scales

Enhancing Predictive Power by Unamalgamating Multi-item Scales

David Trafimow (Corresponding author)
Distinguished Achievement Professor of Psychology
New Mexico State University
Department of Psychology, MSC 3452
Box 30001
Las Cruces, NM 88003
Voice Phone: 575-646-4023
Email: mailto:dtrafimo@nmsu.edu

Michael R. Hyman
Founder and President, Institute for Marketing Futurology and Philosophy
5260 Redman Road
Las Cruces, NM 88011-7556
Voice Phone: 575-522-8463
Email: michaelrhyman88011@gmail.com

Alena Kostyk
Senior Lecturer in Marketing
University of Glasgow
Adam Smith Business School
University Avenue, Glasgow G12 8QQ
Email: Alena.Kostyk@glasgow.ac.uk

Word count: 10317

Enhancing Predictive Power by Unamalgamating Multi-item Scales

Abstract

The generally small but touted as 'statistically significant' correlation coefficients in the social sciences jeopardize theory testing and prediction. To investigate these small coefficients' underlying causes, traditional equations such as Spearman's (1904) classic attenuation formula, Cronbach's (1951) α, and Guilford and Fruchter's (1973) equation for the effect of additional items on a scale's predictive power are considered. These equations' implications differ regarding large inter-item correlations enhancing or diminishing predictive power. Contrary to conventional practice, such correlations decrease predictive power when treating items as multi-item scale components but can increase predictive power when treating items separately. The implications are wide-ranging.

Keywords: correlations; Spearman; Guilford; multi-item scales; single scale items; effect size

Even when theory suggests otherwise, small correlation coefficients (*r*s) pervade the social sciences (Hofmann, 2005; Smedslund, 2016). Unfortunately, such *r*s imply limited predictive and theoretical power. Because many social scientists believe *r*s, R^2s, and path coefficients reflect the likely effect of interventions or policy changes on dependent variables, predictive limits imply corresponding application limits (Pearl & Mackenzie, 2018). From a theory-testing perspective, smaller effects are more susceptible to alternative explanations because minor study confounds and imperfections can more plausibly explain smaller effects (Trafimow, 2022).

Social scientists cope with small *r*s by either mitigating the associated problems (e.g., a *preponderance of the evidence approach* in which multiple theoretically related *r*s that differ significantly from zero confirm the underlying theory; see Trafimow et al., 2022) or determining why they exist as a prelude to formulating subsequent empirical studies. The former way is problematic because it assumes abundant weak evidence sufficiently tests and confirms an associated theory or application. However, philosophers of science have undercut this argument (Duhem, 1954; Quine, 1951; Spirtes, Glymour, & Scheines, 2000). Data generally underdetermine theories, and this problem worsens when data, regardless of the quantity, are weak. The latter way focuses on classical concepts—mainly classical test theory, aka classical true score theory or the classical theory—to revisit construct measurements' tenets and familiar postulates, such as increasing R^2 by adding items to multi-item scales. Contrary to intuition, large *r*s among a multi-item scale's items worsen criterion prediction. In addition, small *r*s among such scale items can worsen criterion prediction when item-criterion *r*s differ substantially. These outcomes contravene the entrenched practice of creating multi-item scales to

reveal latent construct structures. Instead, treating construct-related items separately improves criterion prediction, remedies the small *r*s issue, and encourages better theory development.

Classical equations and multi-item scales

Classical test theory remains foundational in psychology and related fields. By subsuming rather than refuting it, newer theories like generalizability theory (see Brennan, 2001 for a review) and item response theory (see Hulin et al., 1983 for a review) superseded classical test theory. Essentially, the classical theory is a special case of newer and stronger measurement theories. However, the increased power of newer theories derives from 'stronger but less likely to be unequivocally true' assumptions (Gulliksen, 1987; Lord & Novick, 1968; Trafimow, 2021b).

In his well-known formula, Spearman (1904) showed that reliability sets an upper limit on prediction, rendered as Equation 1:

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

Equation 1 contains the following components:

- r_{XY} is the observed correlation between *X* and *Y*;
- $r_{T_X T_Y}$ is the true correlation between *X* and *Y*, or the *r* obtained without random measurement error; and
- $r_{XX'}$ and $r_{YY'}$ are the reliabilities of the *X* and *Y* measures.

If the reliability of *X* or *Y* = 0, the observed *r* = 0 regardless of the true *r*'s magnitude. In contrast, if both reliability coefficients equal 1, the observed *r* = the true *r* (i.e., the best-case reliability scenario). For intermediate cases, which take the product of the reliabilities as a single reliability product (*RP*), $RP = r_{XX'}r_{YY'}$ and Equation 1 reduces to $r_{XY} = r_{T_X T_Y}\sqrt{RP}$. By showing

how observed *r*s increase as reliability products increase, with different curves representing different true *r*s, Figure 1 shows reliability is a prerequisite for high *r*s.

----- Place Figure 1 here -----

How can social scientists ensure high reliability? First, consider the typical respondent data collected to measure X and Y. Researchers assuming multiple questionnaire items can measure the same variable (Shevlin et al., 1997; Spector, 1992) often use "a collection of items, the responses to which are…combined to yield a scale score" (Dawis, 1987, p.481). Marketing scholars, for example, rely on an average of four items per scale (Bruner et al., 1993). Such scales are popular due to their construction and administration ease, formulaic statistical analyses, intuitive appeal, and flexibility (Drewes, 2009; Hyman & Sierra, 2010).

To calculate scale scores, researchers use equally or unequally weighted combinations of people's responses to the scale comprising items (Drewes, 2009; Shevlin et al., 1997). Commonly, they compute scale scores by 'averaging' equally weighted items (hereafter called 'multi-item scales' regardless of amalgamation method; Marsh & Hocevar, 1988; Perloff & Persons, 1988).

Although more favored reliability metrics for multi-item scales exist (Zinbarg et al., 2005), Cronbach's α remains the most popular and simplest to understand (Cronbach, 1951; see Crocker & Algina, 1986 for an accessible review).[1] Equation 2 makes the necessary conceptual points sufficiently.

$$\text{Reliability as indexed by } \alpha_{standardized} = \frac{K\bar{r}}{1+(K-1)\bar{r}}. \qquad (2)$$

Equation 2 includes the following:

- $\alpha_{standardized}$ is the reliability coefficient,

[1] There is no assumption that standardized and unstandardized Cronbach alpha are mutually interchangeable.

- $\bar{r}$ is the mean inter-unit (inter-item for present purposes) coefficient, and
- K is the number of units (items for present purposes).

Based on Equation 2, Figure 2 shows reliability increases as the mean inter-item *r* or number of items increases. If the mean inter-item *r* is low (e.g., 0.2) but the number of items is high (e.g., 32), reliability can be high (e.g., 0.89). Alternatively, if the mean inter-item *r* is high (e.g., 0.9) but the number of items is low (e.g., two), reliability also can be high (e.g., 0.95). Thus, *scales with many items and high inter-item rs can be highly reliable regardless of theoretical justification*.

----- Place Figure 2 here -----

The strange implications of Guilford and Fruchter (1973)

In their classic text entitled *Fundamental statistics in psychology and education* (1973, p.386), Guilford and Fruchter show that adding items to one scale affects its ability to predict scores on another scale. Although they provided equations for weighting items differently, their simpler unweighted equations are sufficient here. In its most general form, Equation 3 indicates the ability of a scale comprising any number of unweighted items to predict a criterion.

$$r_{cs} = \frac{\sum r_{ci}\sigma_i}{\sqrt{\sum \sigma_i^2 + 2\sum r_{ij}\sigma_i\sigma_j}}. \qquad (3)$$

Equation 3 has the following components:

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

Equation 3 is expandable for one, two, three, four, or five items. These expanded equations appear in Table 1.

----- Place Table 1 here -----

Figure 3, which derives from Table 1, relates scale-criterion *r*s ranging from 0.3 to 0.8 along the vertical axis to inter-item *r*s ranging from 0.1 to 0.9 along the horizontal axis. Adding more items increases the number of inter-item *r*s, item-criterion *r*s, and standard deviations. For simplicity's sake, the inter-item *r*s for a given number of items vary consistently (i.e., all inter-item correlations are 0.1, 0.2, …, 0.9, along the horizontal axis), all item-criterion *r*s equal 0.4, and the standard deviations equal 1.0 in Figure 3 (and all subsequent figures). The five curves represent 5 (top curve), 4, 3, 2, or 1 (bottom curve) items. For only one item, the scale-criterion *r* equals the item-criterion *r*; there are no inter-item *r*s, so the bottom curve in Figure 3 is a straight line representing an *r* of 0.4.

----- Place Figure 3 here -----

Figure 3's implications are straightforward yet subtle. The straightforward implication is more items improve criterion prediction. The subtle implication is a scale's predictive power decreases as the inter-item *r*s increase. In addition, an interaction exists whereby this latter effect is more pronounced as the number of scale items increases.

A scale's predictive power decreases as the inter-item *r*s increase. For simplicity's sake, equalizing the standard deviations does not imply equal standard deviations are necessary for predictive power to decrease as inter-item *r*s increase. Equation 3 and any expansion in Table 1 show that each inter-item *r* is in the denominator and connects to the denominator's other terms by plus signs. Therefore, larger denominators imply smaller overall values because it is mathematically necessary—ceteris paribus and whether equal or not—that larger inter-item *r*s

imply poorer prediction.[2] Although larger or smaller item-criterion *r*s and standard deviations can influence the extent large inter-item *r*s decrease predictive power, large inter-item *r*s will never increase predictive power over that engendered by small inter-item *r*s.[3]

Although Equation 2 indicates higher inter-item *r*s cause increased test reliability, Equation 3 shows decreased predictive power (i.e., each monotonically increasing curve in Figure 1), which contradicts the conventional wisdom that higher reliability improves a scale's predictive power. In essence, higher reliability decreases the predictive power of scales with a given number of items (i.e., each curve in Figure 3).

To address this provocative implication, consider social scientists' loose yet standard usage of reliability coefficients. Typically used reliability metrics—such as split-half, unidimensional, and multidimensional—represent single-administration estimates (i.e., reliability assessment via one once-administered test; Revelle & Condon, 2019; Subkoviak, 1976). Despite their popularity and seeming simplicity, social scientists often misunderstand these metrics (Dunn et al., 2014; Lee and Hooley, 2005). For example, fewer than half of psychology program administrators indicated most Ph.D. students could assess reliability correctly (Aiken et al., 2008).

The most popular single-administration reliability coefficient, Cronbach's α, depends on the number of items and their mean *r* (see Equation 2; Dunn et al., 2014). Consistent with Equation 1 and Figure 1, social scientists often insert Cronbach's α into a disattenuation formula to correct *r*s for unreliability-induced attenuation (see Hunter & Schmidt, 1990 for a review).

[2] As in any mathematical equation with multiple variables, researchers can counteract the effects of varying one variable with variations on one or more other variables, as shown subsequently. However, such cases are not germane here.

[3] We thank an anonymous reviewer for the opportunity to clarify that large inter-item *r*s decrease predictive power ceteris paribus, even if they are not equal.

Nonetheless, Cronbach's α is "most definitely not an actual measure of reliability" (Lee & Hooley, 2005, p.370) and neither a measure of internal consistency nor a function of test unidimensionality (Drewes, 2009; Revelle & Condon, 2019). Cronbach's α and other single-administration reliability coefficients indicate almost nothing about reliability under classical test theory (Revelle & Condon, 2019).

Under classical test theory, reliability is the correlation between parallel scales, i.e., scales with identical means and standard deviations. (Note: An alternative classical test theory definition is true score variance divided by observed score variance.) Assuming this definition, consider this fanciful example: a 'personal profile' scale with one item asking respondents the number of digits in their street address, a second item asking the number of letters in their surname, and a third item asking their birth month (coded 1 through 12). This scale's inter-item *r*s (and related reliability coefficients) should be minimal. However, classical reliability should not be low, as the test-retest reliability for each respondent's answers to these items on successive months should be high. Relative to single-administration reliability coefficients, test-retest reliability is more consistent with the classical theory assumption of infinite independent scale administrations (e.g., see Lord & Novick, 1968; Gulliksen, 1987 for well-cited reviews). Likewise, rephrased items from a parallel scale would correlate highly with the original items. This fanciful example shows high classical scale reliability with a low single-administration reliability coefficient is possible. Classical reliability assessed via test-retest or parallel tests remains a prerequisite for large *r*s.

Unfortunately, Cronbach's α confounds inter-item *r*s and the number of items. In addition, the present demonstration indicates, ceteris paribus, that larger inter-item *r*s imply poorer prediction. Hence, researchers should ensure that the inter-item *r*s and the number of

items are unconfounded, doable by reporting those coefficients in a table, possibly augmented by mean or median inter-item *r*s. They also should report item-criterion *r*s, allowing other researchers to conduct independent analyses, decide the extent particular items are acceptable indicators of the focal construct, and determine whether the test predicts better or worse than single items.

An implication of Figure 3 is that more items increase a scale's predictive power, which is consistent with the standard practice of creating multi-item scales. However, Figure 3 assumes equal item-criterion *r*s. What if non-equality is assumed? Figure 4 addresses this question. Unlike Figure 3, the *r* between the first item and the criterion is 0.5, and the *r* between the other items and the criterion is 0.1. Thus, adding items in Figure 4 means adding items that correlate poorly with the criterion. Although Figures 3 and 4 show increasing the inter-item *r* decreases the scale-criterion *r*, these figures differ meaningfully. In Figure 3, adding items increases the scale-criterion *r*, but in Figure 4, adding items decreases the scale-criterion *r*. Thus, the maxim 'more items are better' requires qualification; more items are better if they predict the criterion equally well, but more items are worse if the added items correlate poorly with the criterion. Striving for impressive single-administration reliability and adding items that correlate poorly with the criterion are two standard research practices that can worsen prediction.

----- Place Figure 4 here -----

When all else is not constant[4]

Ceteris paribus and relative to smaller inter-item *r*s, larger inter-item *r*s worsen criterion prediction. However, ceteris paribus may not pertain. In such cases, larger inter-item *r*s may out-predict smaller ones. Consider Equation 3. In addition to r_{ij} influencing r_{cs}, variables such as r_{ci},

[4] We thank an anonymous reviewer for suggesting this section.

σ_i, and σ_j can influence r_{cs}. Although increasing r_{ij} causes r_{cs} to decrease, ceteris paribus, a net increase in r_{cs} can occur if r_{ci}, σ_i, or σ_j vary while r_{ij} decreases. Modifying Equation 3 to derive an expression for change can show this effect. Imagine a context that considers values $r_{cs\prime}$, $r_{ij\prime}$, $r_{ci\prime}$, $\sigma_{i\prime}$, and $\sigma_{j\prime}$ rather than the original values r_{cs}, r_{ij}, r_{ci}, σ_i, and σ_j. A change in the multi-item scale's ability to predict a criterion is expressable as $r_{cs\prime} - r_{cs}$. A positive (negative) difference represents an item's increased (decreased) ability to predict the criterion. Rewriting Equation 3 with $r_{cs\prime}$, $r_{ci\prime}$, $\sigma_{i\prime}$, and $\sigma_{j\prime}$ produces Equation 3':

$$r_{cs\prime} = \frac{\sum r_{ci\prime}\sigma_{i\prime}}{\sqrt{\sum \sigma_{i\prime}^2 + 2\sum r_{ij\prime}\sigma_{i\prime}\sigma_{j\prime}}}. \quad (3')$$

Subtracting Equation 3 from Equation 3' yields Equation 4:

$$r_{cs\prime} - r_{cs} = \frac{\sum r_{ci\prime}\sigma_{i\prime}}{\sqrt{\sum \sigma_{i\prime}^2 + 2\sum r_{ij\prime}\sigma_{i\prime}\sigma_{j\prime}}} - \frac{\sum r_{ci}\sigma_i}{\sqrt{\sum \sigma_i^2 + 2\sum r_{ij}\sigma_i\sigma_j}}. \quad (4)$$

Equation 4 permits any variable that influences $r_{cs\prime} - r_{cs}$ to remain constant or vary.

Figure 5 shows for the two-item case how simultaneously varying r_{12} and r_{c1} affects $r_{cs\prime} - r_{cs}$ (labeled DIFF along the vertical axis). In Figure 5, $\sigma_{1\prime} = \sigma_1 = \sigma_{2\prime} = \sigma_2 = 1.0$, and $r_{c2\prime} = r_{c2} = 0.5$. Although r_{c1} was set at 0.5 (like r_{c2}), $r_{c1\prime}$ was set at 0.8 (top curve) or 0.5 (middle curve) or 0.2 (bottom curve), $r_{12\prime}$ ranged along the horizontal axis from 0 to 0.9, and r_{12} was set at 0.5. All curves monotonically decline as $r_{12\prime}$ increases, consistent with increasing inter-item *r*s worsening a scale's predictive accuracy, ceteris paribus. However, all else is not constant in Figure 5, so many positive values for $r_{cs\prime} - r_{cs}$ exist.

Consider the top curve ($r_{c1\prime} = 0.8$ and $r_{c1} = 0.5$). Although the 0.3 difference confers a substantial advantage for criterion prediction, large values for $r_{12\prime}$ mitigate that advantage, which is insufficient for the curve to dip into negative territory. That $r_{cs\prime} - r_{cs} > 0$ for all cases exemplified by the curve, even when $r_{12\prime} > r_{12}$, shows improved criterion prediction is possible

if an item-criterion *r* increases despite the generally deleterious effect of larger inter-item *r*s on prediction.

----- Place Figure 5 here -----

The middle curve is the most interesting because $r_{cs\prime} - r_{cs}$ values can be greater or less than zero. In this case, $r_{c1\prime} = r_{c1} = r_{c2\prime} = r_{c2} = 0.5$, so item-criterion *r*s are neither advantaged nor disadvantaged. Thus, all depends on $r_{12\prime}$. When $r_{12\prime} > r_{12}$, $r_{cs\prime} - r_{cs} < 0$; when $r_{12\prime} = r_{12}$, $r_{cs\prime} - r_{cs} = 0$; and when $r_{12\prime} < r_{12}$, $r_{cs\prime} - r_{cs} > 0$. The bottom curve is completely within negative territory because $r_{c1\prime} = 0.2$, which comports with the low value for substantially disadvantaging criterion prediction relative to $r_{c1} = 0.5$. As $r_{12\prime}$ increases, $r_{cs\prime} - r_{cs}$ decreases.

When inter-item and item-criterion *r*s vary simultaneously, $r_{cs\prime} - r_{cs} > 0$ is possible even when inter-item *r*s increase. In contrast, Figure 6 shows the effect of keeping item-criterion *r*s constant while varying standard deviations and inter-item *r*s. Specifically, $r_{c1\prime} = r_{c1} = r_{c2\prime} = r_{c2} = 0.5$ $\sigma_{2\prime} = \sigma_2 = 1.0$, and $\sigma_{1\prime}$ varies from 1 (top curve) to 0.1 (middle curve) to 0.001 (bottom curve).[5] As in Figure 5, $r_{12\prime}$ varies along the horizontal axis.

----- Place Figure 6 here -----

The top curve exemplifies an ideal situation because $\sigma_{1\prime} = \sigma_{2\prime} = 1$, maximizing prediction accuracy. Both this curve and the middle curve are in positive territory. In contrast, the bottom curve, with $\sigma_{1\prime} = \sigma_1 = 0.001$, differs markedly. All the values are so near zero that the trend is difficult to discern. Again, when $r_{12\prime} > r_{12}$, $r_{cs\prime} - r_{cs} < 0$; when $r_{12\prime} = r_{12}$, $r_{cs\prime} - r_{cs}$ is zero; and when $r_{12\prime} < r_{12}$, $r_{cs\prime} - r_{cs} > 0$. Like the other curves, lower values for $r_{12\prime}$ imply better criterion prediction than do higher values for $r_{12\prime}$. Hence, larger inter-item *r*s always worsen criterion prediction, ceteris paribus. Although Figures 5 and 6 show $r_{cs\prime} - r_{cs}$ can exceed zero

[5] We also set $\sigma_{1\prime}$ to 0.01, but that curve's proximity to the 0.001 bottom curve made viewing challenging.

even with larger inter-item *r*s, this requires simultaneously varying item-criterion *r*s or standard deviations to compensate for those larger inter-item *r*s.

Now consider the implications of negative inter-item *r*s by revisiting two-item criterion prediction and simultaneously varying inter-item *r*s (including negative values) and item-criterion *r*s. Set $\sigma_1, = \sigma_1 = \sigma_2, = \sigma_2 = 1.0$, as in Figure 5, $r_{c1} = r_{c2} = r_{c2}, = 0.1$, and $r_{12} = 0$. Create simultaneous variations by setting r_{c1}, to 0.1, 0.2, or 0.3 while letting r_{12}, range from -0.90 and 0.90. The top curve in Figure 7 shows highly negative inter-item *r*s markedly improve criterion prediction even when r_{c1}, = 0.3. At the extreme, where r_{12}, = -0.9, $r_{cs}, - r_{cs} = 0.75$. The middle curve, where $r_{c1}, = 0.2$, also shows more negative inter-item *r*s imply better criterion prediction. However, the bottom curve illustrates the same point more interestingly. When r_{12}, < 0, $r_{cs}, - r_{cs}$ > 0; when $r_{12}, = 0$, $r_{cs}, - r_{cs} = 0$; and when r_{12}, > 0, $r_{cs}, - r_{cs}$ < 0.

----- Place Figure 7 here -----

To conclude, more negative inter-item *r*s imply more accurate criterion prediction. However, two caveats pertain. First, the implications of negative inter-item *r*s can depend on whether all else is constant or another variable can vary. Second, researchers might bristle at negative inter-item *r*s because they prefer all scale items assess the same construct.

Multi-item scales versus multiple single-item components

A scale with more items predicts a criterion better than a scale with fewer items if the item-criterion *r*s are similar but not if the added items have low item-criterion *r*s. Does this result favor amalgamating items into a multi-item scale or treating them separately? Should social scientists predict the criterion from single items entered separately in a multiple regression equation or with a multi-item scale? In essence, does multiple regression with single items outperform bivariate correlations? The answer: Sometimes.

Consider a simple two-item case. Figure 8 contains curves reflecting a large (0.5 and 0.1) or a small (0.5 and 0.4) discrepancy between the two item-criterion *r*s. A multi-item scale with both items, resulting in a bivariate *r*, or each item as a separate measure, resulting in a multiple *r*, can predict a criterion. For small discrepancies, the multiple *r* and the bivariate *r* perform similarly regardless of the inter-item *r*s, as indicated by the proximity of the black-solid and black-dashed curves in Figure 8. Although the curves diverge slightly when the inter-item *r* approaches 1.0, the multiple *r* is much larger than the bivariate *r* for large discrepancies in the item-criterion *r*s. The gap increases meaningfully when the inter-item *r* increases, as indicated by contrasting the gray-solid and gray-dashed curves in Figure 8. An error-suppression increase may cause a substantial rise in the gray-solid curve with an increasing inter-item *r*.

----- Place Figure 8 here -----

Hence, a multi-item scale predicts the criterion worse than treating each item separately. However, the difference is trivial when the item-criterion discrepancy is small and substantial when the discrepancy is large, with these effects qualified by the inter-item *r*s.

Discussion with implications

This exposition began with traditional psychometric equations and the social sciences' *small r* scourge. Although the classical attenuation formula indicates reliability increases a scale's predictive power (Figure 1), the equation in Guilford and Fruchter (1973) indicates the opposite (Figure 3) when the number of items is held constant, thereby creating a paradox. Acknowledging researchers' loose use of reliability can avoid this enigma. Whereas large inter-item *r*s worsen predictions, classical reliability improves them. Although this resolution provides a critical foundation, it does not fully address the problem of small *r*s in the social sciences. Figure 3 shows large inter-item *r*s worsen prediction and more items outperform fewer items

when item-criterion *r*s are similar. Figures 3 and 4 have negative implications for large inter-item *r*s, and Figure 4 indicates adding items worsens prediction when those items have substantially smaller item-criterion *r*s than the original item. Both figures suggest researchers can incur low scale-criterion *r*s in several ways.

Synonymous versus non-synonymous scale items

A longstanding psychometric controversy entails whether to use (1) non-synonymous items to reflect a complex-yet-univariate construct's various aspects or (2) synonymous items to enhance inter-item *r*s and single-administration reliability indices (e.g., Trafimow, 2021). Social scientists reporting single-administration reliability metrics like Cronbach's α are motivated to use synonymous items. Researchers justify such metrics as an alternative to test-retest reliability when they prefer administering similar items consecutively to administering identical items repeatedly (Revelle & Condon, 2019). As shown previously, large inter-item *r*s reduce predictive power, so synonymous items (i.e., items with likely large inter-item correlations) yield smaller scale-criterion *r*s. Hence, researchers interested in larger *r*s should reject single-administration reliability and instead rely on measures with smaller inter-item *r*s. Although Figure 2 shows following this advice would worsen single-administration reliability, Figures 3 and 4 show it would enhance predictive power.

The difference between Figures 3 and 4 suggests a second controversy that can yield smaller *r*s. Similar item-criterion *r*s imply using more items enhances prediction; in contrast, different item-criterion *r*s imply using more items worsens prediction when the added items have small item-criterion *r*s. Hence, researchers should not add items that worsen a multi-item scale's predictive power relative to the most predictive item(s). Figure 4 shows that researchers trying to reflect a complex-yet-univariate construct's various aspects by adding items that correlate poorly

with the criterion compromise predictive power. Although challenging in practice, adding items with moderate item-criterion correlations could sidestep this problem.

Although scientifically ill-advised, researchers can obtain smaller *r*s using items that poorly predict the criterion. Causes of this unpreferred approach include the construct is spurious or unimportant (Mischel, 1968), exists and is important but poorly understood (Stroebe et al., 2018), requires additional high-quality auxiliary assumptions linking a non-observational term to an observable attribute (Trafimow, 2012a), is not quantitative (Michell, 1999), fails to specify appropriate measurement units (Trafimow, 2012b), and violates the qualitative homogeneity assumption (Richters, 2021).

Enhancing scale items' predictive power

How can researchers address low *r*s and enhance predictive power if neither dissimilar scale items with widely varying item-criterion *r*s that capture a construct's complete domain nor synonymous scale items are suitable? The answer can be easy or challenging.

Figure 8 suggests the easy answer is to treat each scale item separately and use multiple regression for criterion prediction. Although a multiple regression approach adds little to predictive power when item-criterion *r*s are similar, it outperforms the bivariate *r* when item-criterion *r*s differ meaningfully. Figures 3 and 4 show a multi-item scale's predictive power decreases as its inter-item *r*s increase. However, Figure 8 suggests that multiple regression's error variance suppression (e.g., see the gray-solid curve) somewhat or substantially boosts the predictive power of unamalgamated scale items, which counters previous advice favoring multi-items scales with small inter-item *r*s. Thus, small inter-item *r*s are superior when amalgamating items pre-analysis, but large inter-item *r*s can be superior when keeping items separate. However, separate items increase analytical and reporting complexity by requiring each item's

justification as a valid construct gauge. These considerations indicate a challenging answer that builds on the simple answer.

Consider the concept of perceived behavioral control (e.g., Ajzen, 1988), often assessed by items like ‘doing *X* is under/not under my control’ and ‘doing *X* is easy/difficult’. As perceived control and perceived difficulty are separable constructs—i.e., researchers can manipulate both independently (Trafimow et al., 2002)—merging them into a *perceived behavioral control* construct is problematic. This double dissociation would be improbable with a unidimensional construct, reinforcing the notion that domain-spanning items only seemingly assess the same construct.

Scales with items mapping onto different constructs

A general and underappreciated social science problem is multi-item scales with items mapping onto different constructs. For example, psychologists often treat the 21-item Beck Depression Inventory (see ismanet.org) as a unidimensional scale. Whereas *item 13* asks about people’s decision-making ease, *item 16* asks about sleep. Although both items tap into different aspects of depression, the inventory’s overall single-administration reliability is reasonable. Now consider Figure 2, which shows conventionally acceptable single-administration reliability is possible for multi-item scales with some minimally related items. The item-weighting approach for calculating scale scores can exacerbate this problem; equal-weighting (e.g., when a researcher averages responses to all items) is rarely theory-driven, and unequal-weighting can bias predictions and lower explanatory power (Perloff & Persons, 1988).

Can exploratory factor analysis resolve this issue by guaranteeing that items loading onto the same factor measure the same latent construct? No social science domain has been more subjected to such analysis than the so-called Big 5 personality traits (see John, Naumann, &

Soto, 2008 for a review). Many psychologists argue that the Big 5 personality traits exemplify the triumph of exploratory factor analytic methods in uncovering and measuring latent constructs. But consider two level-of-agreement items from the extraversion trait, perhaps the most popular Big 5 traits: 'Person X is talkative' and 'Person X generates a lot of enthusiasm'. Although these items correlate positively, people can be enthusiastic without being talkative, and vice versa. Upon reflection, the two items assess different constructs despite numerous factor analytic studies indicating otherwise. A similar argument pertains to other extraversion items and items ostensibly measuring other Big 5 traits.

Rather than denigrating exploratory factor analysis per se, the extraversion trait example counters beliefs about factor analyses' sufficiency for ensuring all items loading onto the same factor assess the same construct. Despite robustness claims, exploratory factor analysis is prudent only when scale items load highly onto a single factor and responses to each item are roughly normally distributed and continuous, which frequently is false (Blanca et al., 2013; Hyman, 1996; Marsh et al., 1994; Micceri, 1989; Shevlin et al., 1997). Constructing a multi-item scale with non-synonymous items is theoretically and practically daunting (Churchill, 1979; DeVellis, 2017; Furr, 2011; Hyman & Sierra, 2010). Moreover, wholly and precisely defining a Big 5 trait, depression, love, intelligence, or most other psychological constructs to create valid domain-spanning items is unlikely. However, definition imperfectability justifies a more sophisticated conceptual approach to measurement.

Undefined primitive constructs and competing theories

Interminable debates about defining psychological constructs are feckless absent theoretical contexts. Consider the longstanding debates about the meanings of intelligence and various psychological disorders (e.g., schizophrenia, depression, alcoholism), or the labels

assigned to the Big 5 personality constructs. Because psychologists often underappreciate theory's crucial role in providing meaning, consider this brief history of physics example.

Neither Newton's [1642-1720] nor Einstein's [1879-1955] theories define *mass*. As a dictionary-type definition requires words that themselves require definition, ad infinitum (i.e., an infinite regress problem; see Skipper & Hyman, 1995), both scientists treated mass as an undefined primitive construct (i.e., the intrinsic nature problem; see Goff, 2017, 2019; Harris, 2021) and relied on a comprehensive theory to provide its meaning. However, their theories necessitated different meanings. Algebraic manipulations of Newton's equation $force = mass \cdot acceleration$—the most important one in physics' history (Lederman, 1993)—means $mass = \frac{force}{acceleration}$ for Newton. In contrast to this velocity-independent meaning, an object's relativistic mass for Einstein increases as its relative velocity increases. Because extensive data support Einstein's theory over Newton's theory, Einstein's meaning prevails over Newton's meaning.

Researchers could proceed similarly. If they ponder and propose comprehensive theories anchored by a focal construct, inter-relationships between other constructs and the focal construct should clarify the focal construct's meaning, thereby facilitating its assessment. Hence, researchers could resolve disagreements about construct meaning via empirical contests between competing theories with different focal construct usages and set the winning theory's usage as the construct's de facto meaning. Of course, new theories can alter accepted construct meaning, as Einstein's conceptualization of mass replaced Newton's conceptualization.

Using a focal personality trait like extraversion to denote enthusiasm and talkativeness implies an underlying theory. In the reflective case, the assumption is extraversion causes responses to each scale item. However, this seeming simplicity is an unintended but deceptive artifact of models depicted with arrows between the focal trait and the items (sometimes called

indicators). Although many psychologists insist that extraversion causes enthusiasm and talkativeness, that causation is indirect and questionable at best. In turn, these notions are supposed to induce responses to talkativeness, enthusiasm, and similar questionnaire items, which implies a complex model given the criterion.

The criterion

Researchers rarely consider criterion constructs when devising predictive scales. As reflected by the social science literature and multi-item scale compendia in psychology, marketing, and other social sciences (e.g., Bearden et al., 2011; Bruner II, 2021; Milhausen et al., 2020; Ostrow, 1996; Tate, 2010; Waters & Stephane, 2015), they typically assess targeted constructs with either new or previously developed scales. However, two considerations oppose using such scales if they ignore criterion constructs: (1) item-criterion *r*s are crucial to an item set's predictive power, and (2) a construct's meaning often is best informed by the construct's place within a larger theory. Ignoring the increased clarity and likelihood of creating a valid scale by embedding a focal construct into a comprehensive theory is ill-advised.

For example, Wicker's (1969) famous review precipitated a research crisis by showing attitudes predict behaviors poorly. To resolve the quandary, Fishbein proposed the *principle of correspondence* (aka *principle of compatibility*; Fishbein, 1963, 1967, 1980; Fishbein & Ajzen, 1975, 2010; see Ajzen & Fishbein, 1980 for an accessible description). This principle explicitly considers attitudinal and criterion constructs like behavioral intentions and behaviors. Fishbein argued that attitudes adequately predict behavioral intentions or behaviors when scales of the focal (e.g., attitude) and criterion (e.g., behavioral intentions) constructs correspond to action, target, time, and context. Once researchers started complying with this principle, criterion

predictions based on attitudinal constructs improved markedly (e.g., see Kraus, 1995 for a meta-analysis).

A theory need not include all possible criterion constructs; instead, it needs only specify sufficient connections to inform the focal construct's meaning. Once a theory imbues meaning into a focal construct, it is easier to understand that meaning and identify high-quality auxiliary assumptions connecting that construct to items measuring it (Trafimow, 2012).

A strong assumption undergirding this argument is researchers can devise a theory that includes the focal construct. Again, consider extraversion—a label applied to a mathematically generated factor of seemingly relevant items. Because enthusiasm and talkativeness items load highly on this factor, extraversion's meaning is equivocal. Perhaps neither item captures an aspect of extraversion. Merely assuming extraversion exists is insufficient for specifying the auxiliary assumptions for generating valid scale items.

Single-item scales

The typical argument for multi-item and against single-item scales is 'a single item cannot cover a construct's complete domain'. Hence, survey researchers often create large item pools and use exploratory factor analysis to uncover latent constructs (Churchill, 1979; DeVellis, 2017). However, as already noted, their approach is problematic because having large inter-item *r*s worsens criterion prediction, but small inter-item *r*s imply that not all the items map well onto the same construct.

Researchers distinguish between formative and reflective measurement models (Diamantopoulos & Winklhofer, 2001; Edwards & Bagozzi, 2000; Hyman et al., 2002). The formative case assumes the items define and cause the construct. In contrast, the reflective case assumes people's stance on a construct causes their responses to predictive items. Modeled

pictorially, arrows representing causality between constructs and items point from the items to the construct in the formative case (see Figure 9) and vice versa in the reflective case (see Figure 10).

----- Place Figures 9 and 10 here -----

Formative measurement models entail multiple indicators; for example, socioeconomic status includes income, education, occupational status, and homeownership (*American Psychological Association*, 2007). In contrast, reflective measurement models assume latent variables cause scale item responses; for example, extroversion causes people's responses to items about talkativeness and enthusiasm. Hence, the reflective assertion is extroverted people endorse being talkative and enthusiastic. However, an alternative and more probable assertion is talkative people endorse talkativeness and enthusiastic people endorse enthusiasm. If believers in traits prefer reflective assertions, why attribute responses about talkativeness and enthusiasm to an extraversion trait when they are more directly attributable to a talkativeness or enthusiasm trait? This rhetorical question suggests that researchers reject improbable reflective measurement models like 'extroversion causes responses to items about talkativeness and enthusiasm' rather than 'talkative and enthusiastic people endorse items about talkativeness and enthusiasm'. In essence, should researchers use a single talkativeness item to assess talkativeness?

The predictive validity of single-item and multi-item scales for simple, concrete, and intuitively accessible constructs (e.g., talkativeness) are similar (Begrkvist & Rossiter, 2007). Many comparative studies indicate single and multi-item scales for the same variables perform similarly (e.g., Abdel-Khalek, 1998; Cheah et al., 2018; Graf et al., 2018). Although the intuitive accessibility of some items may vary by culture, multi-item scales are prone to similar challenges and provide fewer advantages. Ensuring each scale item is as concrete as possible may require

several simplifying iterations. Instead of using exploratory factor analysis to create a multi-item scale that covers a construct's complete domain, as it will perpetuate the problem, researchers could treat each item separately.

Because constructs always have surplus meaning and are thus non-observational (MacCorquodale & Meehl, 1948), efforts to create conceptually exhaustive multi-item scales are doomed. Instead, researchers could either (1) identify a construct's single underlying essence or (2) create a subconstruct amalgam from multiple unidimensional constructs (Calder et al., 2021). Under approach (1), researchers with a perfect understanding of extroversion's essence could ensure its unidimensional assessment with a near-perfect single-item scale, i.e., not rely on conceptually different notions like talkativeness or enthusiasm. Under approach (2), extraversion is a multidimensional construct comprising aspects like talkativeness and enthusiasm assessable with single talkativeness and enthusiasm items. Regardless, single-item scales are not inherently problematic (Bergkvist & Rossiter, 2007; Diamantopoulos et al., 2012; Gardner et al., 1998).

In contrast, researchers could use exploratory factor analyses to uncover latent dimensions (e.g., Big 5 research). However, this process tacitly concedes a posited construct may be a multidimensional amalgam. Hence, researchers must carefully conceptualize and analyze each of those dimensions. For example, merely positing five personality traits (factors) leaves many important questions unanswered, such as:

- Does each factor correspond to a single construct?
- Assuming each factor represents a single construct, what theory explains the inter-construct connections?
- What causes the factor-defined constructs?
- What do the factor-defined constructs cause?

Researchers interested in Big 5 subfactors could ask similar questions and theorize accordingly. Such thinking might yield a theory with sufficient clarity to inform opinions about whether talkativeness, enthusiasm, both, or neither best assess extraversion (assuming it exists).

Inadequate initial theorizing does not justify using multi-item scales. Consider many researchers' lackadaisical theorizing. Treating talkativeness and enthusiasm independently is preferable sans theory. Figure 5 shows that treating scale items separately improves prediction. Hence, researchers should reject multi-item scales that create paradoxes and treat each scale item as a separate predictor. Multiple regression analyses with separate items can suppress error, offer excellent criterion prediction when inter-item *r*s are large, and address the small *r*s problem, as the gray-solid curve in Figure 5 shows. This curve counters the conventional wisdom that multi-item scales generally are superior. Alternatively, moderate inter-item *r*s and moderate or weak criterion predictive power may signal multidimensional predictive constructs.

Focusing on the predictive power of single items could help refine constructs. Imagine construct θ is assessed with five items (X_1, X_2, X_3, X_4, and X_5). In addition, suppose researchers use those items to predict a criterion with a multi-item scale. If the typical *r*s are small, little meaningful knowledge is acquired. In contrast, suppose researchers use each item to predict a criterion and X_3 always performs best. If a criterion connects to θ theoretically, even if the reasons are unknown, X_3's consistent predictive superiority provides essential information. Perhaps X_3, relative to the other four items, better approximates θ's essence. That knowledge could help refine thinking about θ and eventually yield a well-specified theory that better indicates its essence, includes an improved set of constructs with clear linkages to θ, and improves criterion prediction.

Canceling random response errors to items in multi-item scales

Assume factor θ with sub-factors θ_1, θ_2, …, θ_k and random response errors for each item that cancel one another (i.e., the items collectively reflect θ accurately). From a classical true score theory perspective, θ need not exist. A person's true score is the expectation across infinite test administrations (i.e., the observed score is the true score plus random error: $O = T + E$). Hence, a person's true score and θ score can differ for many reasons, such as non-continuous (i.e., discrete) and truncated (due to endpoints) measures (Lord & Novick, 1968).

If $T \neq \theta$, the error terms contradict. From a classical true score theory perspective, $E = O - T$. From a θ perspective, $E = O - \theta$, but if $T \neq \theta$, it follows that $E \neq E$! The solution, however, is to assume two different meanings for *E*; under classical theory, *E* equals random error only, whereas, under a θ perspective, *E* equals random and systematic error. In essence, under the θ perspective, errors typically are not entirely random; hence, a belief in inter-item error cancellation may be problematic.

Multicollinearity

Social scientists often consider the effect of multicollinearity on least-squares-estimated coefficients in multiple regression analyses. Because multicollinearity poses a problem when it is high relative to the overall multiple correlation, they use variance inflation factors (VIFs) to estimate this magnitude (Haitovsky, 1969; Mansfield & Helms, 1982). The approach proposed here (i.e., entering multiple single-item predictors into a multiple regression model to predict a criterion) may seem undesirable due to possible multicollinearity issues.

Fortunately, multicollinearity is only consequential for regression coefficient interpretation. When focusing on multiple correlations rather than regression coefficients, intercorrelation can increase the scale items' ability to suppress error collectively, which is

desirable. Zero-order correlations rather than regression weights can indicate item-criterion relationships. Ultimately, multicollinearity is acceptable when focusing on prediction—the primary purpose of most measures (Perloff & Persons, 1988).

Structural equation modeling

Although the present conclusions derive primarily from multiple regression, structural equation modeling (SEM) yields similar conclusions. The simulation was conducted for covariance-based maximum likelihood SEM instead of PLS-SEM because the latter is not a latent-variable method (Goodhue et al., 2012; Rönkkö et al., 2023). For this simulation, an independent variable measured on a 3-item scale (reflective model) predicted a criterion assessed via a single item with no measurement error[6] (see Figure 11).

----- Place Figure 11 here -----

Consistent with the simulation presented in Figure 3, each scale item was set to correlate with the criterion at the 0.4 level, and all standard deviations were set to equal 1.0. Per Figure 3, the simulation was run for four levels of inter-item *r*s from low to high: 0.1, 0.3, 0.8, and 0.9. Table 2 summarizes these simulation parameters as four different covariance matrices. Assuming a sample size of 250 and using these covariance matrices as input in AMOS, standardized path coefficients from the predictor to criterion are 1.26, 0.73, 0.45, and 0.42, respectively, for each inter-item correlation level. Thus, the prior conclusion that increased inter-item *r*s worsen criterion prediction, ceteris paribus, generalizes to SEM.

Empirical examples

Hyman et al. (2002) conducted a consumer lifestyle study based on two pilot studies and a large-sample main study (N = 725). The study focused on consumer affluence and included

[6] The SPSS AMOS syntax for this simulation is as follows: CriterionMeasure = (1) Criterion + (1) CriterionError1; Criterion = (1) r1 + Scale; Item1 = (1) Scale + (1) Error1; Item2 = Scale + (1) Error2; Item3 = Scale + (1) Error3.

questions about consumer-related constructs like fashion consciousness. It conceptualized consumer affluence as a complex and formative construct and compared a formative affluence scale to a single attitudinal affluence item (i.e., 'I think I have an affluent lifestyle').

One of the hypothesized constructs comprising affluence is materialism, assessed with the following six agree-disagree items obtained from Richins (1987), with the last item reverse-scored (Cronbach's $\alpha = 0.71$).

- It's really true that money can buy happiness.
- It is important to me to have really nice things.
- I would like to be rich enough to buy anything I want.
- I'd be happier if I could afford to buy more things.
- It sometimes bothers me quite a bit that I can't afford to buy all the things I want.
- People place too much emphasis on material things. (*reverse-scored*)

The items arguably assess different although related notions; for example, whether 'money can buy happiness' differs from 'owning nice things'. The former item is broader than the latter item. Whereas *nice things* represent various physical objects (i.e., goods), consumers can *buy happiness* by purchasing services (e.g., dry cleaning) or experiences (e.g., exotic vacations). The third *rich enough* item implies consumers' buying ability rather than possessing and consuming things and experiences relates to materialism. Similar comments about the remaining three items are possible.

How well does materialism predict affluence? As the preceding commentary suggests, it depends. Statistically significant due to the large sample size, the bivariate *r* based on the materialism scale is only 0.12. Although this coefficient rises to 0.26 for the single best item (i.e., 'It is important to me to have really nice things'), new data is needed to cross-validate this

finding. Adjusting for the number of items, the multiple *r* for all six unamalgamated items jumps to 0.38. Thus, the multi-item scale predicts poorly, the best single item predicts better, and separately treated items predict best.

However, the single best item will not always outperform the related multi-item scale. Figure 3 shows the opposite can occur when the item-criterion *r*s are somewhat similar across items. Figure 8 also shows that separately treated items may not outperform multi-item scales. However, the materialism-affluence association shows separately treated items' predictive superiority.

The preceding example concerns a formative scheme under which materialism is one of several constructs that define affluence. Now consider a reflective scheme; specifically, predict an item termed *fashion innovativeness* (i.e., 'I am among the first to try a new fashion') from *fashion consciousness* (Darden & Perreault, 1976; Lumpkin & Darden, 1982), with the fashion consciousness items bullet-listed below (Cronbach's $\alpha = 0.86$):

- I usually have one or more outfits that are of the very latest style.
- A person should try to dress in style.
- When I must choose between the two I usually dress for fashion, not for comfort.
- An important part of my life and activities is dressing smartly.
- I often try the latest hairdo styles when they change.
- It is important that my clothes be of the latest style.

As with materialism, it is unclear whether the fashion consciousness items assess the same construct despite the high Cronbach's α; for example, the first item is about owned clothing and the second item is about prescribed attire. Similar comments could apply to the other items. However, the similar meanings between the fashion consciousness and fashion innovativeness

items suggest the former will predict the latter better than the materialism items predict affluence.

The analysis confirms this expectation. The multi-item fashion consciousness scale yields a bivariate *r* of 0.66. This coefficient rises to 0.7 for the best fashion consciousness item (i.e., 'It is important that my clothes be of the latest style'). Adjusting for the number of items, the multiple *r* for all separately treated fashion consciousness items jumps to 0.75. Again, the multi-item scale predicts well but worst, the best single item predicts better, and the separately treated items predict best.

The previous formative and reflective examples indicate that the best single or separately treated items can outperform multi-item scales. To reinforce the latter point, consider the criterion's variance explained by the multi-item scale versus separately treated items. For materialism-affluence adjusted for the number of separately treated items, these values are 0.01 and 0.14, respectively, which differ by roughly 13%. For fashion, these values are 0.44 and 0.57, which also differ by roughly 13%. Thus, the present recommendations can confer meaningful explained variance gains. Moreover, these examples underestimate the possible gains because the multi-item scales were empirically (factor analysis) rather than theoretically based. A strong theory connecting materialism to affluence or fashion consciousness to fashion innovativeness would yield improved items that better capture the constructs and correlate with the criterion.

Conclusion

How can researchers obtain larger *r*s between predictive measures and criteria? The answer is multifaceted. One way is to use single items that better capture the construct, which, in turn, implies its embeddedness in a comprehensive theory and robust nomological network that maximizes meaning clarity. A second but related way is to use single items that better correlate

with the criterion and rely on a sound theory that includes the criterion construct and its connections with other constructs. A third way is to contemplate inter-item *r*s under amalgamation or separate treatment. Under amalgamation, large inter-item *r*s worsen criterion prediction, which may account for the small *r*s that pervade the social science literature.

Error suppression can cause large inter-item *r*s to enhance criterion prediction for non-multi-item analyses. Hence, amalgams offer no statistical advantage over separate treatments because the former never yields better criterion prediction. Although amalgams seem more straightforward, such apparent simplicity is highly deceptive and obscures so-called factors often including items mapping onto multiple constructs. If added items, because they map onto other constructs, have low item-criterion *r*s, their amalgamation may lower scale-criterion *r*s. Determining the extent to which single items predict a criterion may help researchers identify, refine, and theorize about the best focal constructs.

Researchers could substantially mitigate the problem of small *r*s by adhering to the preceding strictures: better theorizing with explicit connections between constructs, better item-criterion *r*s, small inter-item *r*s with amalgamation, and large inter-item *r*s sans amalgamation in error suppression contexts. Larger scale-criterion *r*s imply better practice and theory testing less subject to alternative explanations. Researchers need not settle for small scale-criterion *r*s.

References

Abdel-Khalek, A. M. (1998). Single-versus multi-item scales in measuring death anxiety. *Death Studies*, *22*(8), 763-772. doi: 10.1080/074811898201254

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of Ph.D. programs in North America. *American Psychologist*, *63*(1), 32-50. doi: 10.1037/0003-066X.63.1.32

Ajzen, I. (1988). *Attitudes, personality, and behavior*. Chicago, IL: Dorsey.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall.

American Psychological Association (2007). *Report of the APA task force on socioeconomic status*. Washington, DC: American Psychological Association.

Bearden, W. O., Netemeyer, R. G., & Haws, K. L. (eds.) (2011). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research, 3rd ed.* Thousand Oaks, CA: Sage Publications, Inc.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*(2), 175-184. doi: 10.1509/jmkr.44.2.175

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*, 78-84. doi: 10.1027/1614-2241/a000057

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071. doi: 10.1037/0033-295X.111.4.1061

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Bruner, G. C., & Hensel, P. J. (1993). Multi-item scale usage in marketing journals: 1980 to 1989. *Journal of the Academy of Marketing Science, 21*(4), 339-344. doi: 10.1007/BF02894526

Bruner II, G. C. (2021). *Marketing scales handbook: Multi-item measures for consumer insight research, Vol. 11*. Fort Worth, TX: GCBII Productions, LLC.

Calder, B. J., Brendl, C. M., Tybout, A. M., & Sternthal, B. (2021). Distinguishing constructs from variables in designing research. *Journal of Consumer Psychology*, *31*(1), 188-208. doi: 10.1002/jcpy.1204

Cheah, J. H., Sarstedt, M., Ringle, C. M., Ramayah, T., & Ting, H. (2018). Convergent validity assessment of formatively measured constructs in PLS-SEM: On using single-item versus multi-item measures in redundancy analyses. *International Journal of Contemporary Hospitality Management*, *30*(11), 3192-3210. doi: 10.1108/IJCHM-10-2017-0649

Churchill, Jr., G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*(1), 64-73. doi: 10.1177/002224377901600110

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

Darden, W. R., & Perreault, Jr., W. D. (1976). Identifying interurban shoppers: Multiproduct purchase patterns and segmentation profiles. *Journal of Marketing Research, 13*(1), 51-60. doi: 10.1177/002224377601300107

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*(4), 481-489. doi: 10.1037/0022-0167.34.4.481

DeVellis, R. F. (2017). *Scale development: Theory and applications (4th ed)*. Los Angeles, CA: Sage.

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science, 40*(3), 434–449. doi: 10.1007/s11747-011-0300-3

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269-277. doi: 10.1509/jmkr.38.2.269.18845

Drewes, D. W. (2009). Subject-centered scalability: The sine qua non of summated ratings. *Psychological Methods*, *14*(3), 258-274. doi: 10.1037/a0016621

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

Edwards J., & Bagozzi R. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155-174. doi: 10.1037/1082-989X.5.2.155

Fishbein, M. (1963). An investigation of the relationships between beliefs about an object and the attitude toward that object. *Human Relations*, *16*(3), 233-239. doi: 10.1177/001872676301600302

Fishbein, M. (1967). Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement* (pp. 477-492). New York, NY: Wiley.

Fishbein, M. (1980). Theory of reasoned action: Some applications and implications. In H. Howe & M. Page (Eds.), *Nebraska symposium on motivation, 1979* (pp. 65-116). Lincoln, NB: University of Nebraska Press.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York, NY: Psychology Press.

Furr, R. M. (2011). *Scale construction and psychometrics for social and personality psychology.* Thousand Oaks, CA: Sage Publications Inc.

Gardner, D. G., Cummings, L. L., Dunham, R. B., & Pierce, J. L. (1998). Single-item versus multiple-item measurement scales: An empirical comparison. *Educational and Psychological Measurement, 58*(6), 898-915. doi: 10.1177/0013164498058006003

Goff, P. (2019). *Galileo's error: Foundations for a new science of consciousness*. New York, NY: Pantheon Books.

Goff, P. (2017). *Consciousness and fundamental reality*. New York, NY: Oxford University Press.

Goodhue, D. L., Lewis, W., & Thompson, R. (2012). Does PLS have advantages for small sample size or non-normal data? *MIS Quarterly*, *36*(3), 981-1001.

Graf, L. K., Mayer, S., & Landwehr, J. R. (2018). Measuring processing fluency: One versus five items. *Journal of Consumer Psychology*, *28*(3), 393-411. doi: 10.1002/jcpy.1021

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.

Haitovsky, Y. (1969). Multicollinearity in regression analysis: Comment. *The Review of Economics and Statistics*, *51*(4), 486-489.

Harris, A. (2021). *Conscious: A brief guide to the fundamental mystery of the mind*. New York, NY: Harper Collins.

Hofmann, W., Gawronski, B. Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369-1385. doi: 10.1177/0146167205275613

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. New York, NY: Dow Jones-Irwin.

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

Hyman, M. R. (1996). A critique and revision of the Multidimensional Ethics Scale. *Journal of Empirical Generalisations in Marketing Science*, 1, 1-35, http://www.empgens.com/ArticlesHome/Volume1/MultidimensionalEthics.html.

Hyman, M. R., Ganesh, G., & McQuitty, S. (2002). Augmenting the household affluence construct. *Journal of Marketing Theory and Practice, 10*(3), 13-31. doi: 10.1080/10696679.2002.11501917

Hyman, M. R., & Sierra, J. J. (2010). *Marketing research kit for dummies*. Hoboken, NJ: Wiley Publishing, Inc.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big-five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L.

A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.

Kraus, S. J. (1995). Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin*, *21*(1), 58-75. doi: 10.1177/0146167295211007

Lederman, L. (1993). *The God particle: If the universe is the answer, what is the question?* New York, NY: Houghton Mifflin.

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

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

Lumpkin, J. R., & Darden, W. R. (1982). Relating television preference viewing to shopping orientations, lifestyles, and demographics. *Journal of Advertising, 11*(4), 56-67. doi: 10.1080/00913367.1982.10672822

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55*(2), 95-107. doi: 10.1037/h0056029

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician, 36*(3a), 158-160. doi: 10.1080/00031305.1982.10482818

Marsh, H., Hau, K. T., Roche, L. A., Craven, R., Balla, J. R., & McInerney, V. (1994). Problems in the application of structural equation modeling: Comment on Randhawa, Beamer, and Lundberg (1993). *Journal of Educational Psychology*, *86*(3), 457-462. doi: 10.1037/0022-0663.86.3.457

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73*(1), 107-117. doi: 10.1037/0021-9010.73.1.107

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi: 10.1037/0033-2909.105.1.156

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* New York, NY: Cambridge University Press.

Milhausen, R. R., Sakaluk, J. K., Fisher, T. D., Davis, C. M., & Yarber, W. L. (2020). *Handbook of sexuality-related measures*. New York, NY: Routledge.

Mischel, W. (1968). *Personality and assessment*. New York, NY: Wiley.

Ostrow, A. C. (1996). *Directory of psychological tests in the sport and exercise sciences.* Morgantown, WV: Fitness Information Technology.

Pearl, J., & Mackenzie, D. (2018). *The book of why*. New York, NY: Basic Books.

Perloff, J. M., & Persons, J. B. (1988). Biases resulting from the use of indexes: An application to attributional style and depression. *Psychological Bulletin, 103*(1), 95-104. doi: 10.1037/0033-2909.103.1.95

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review, 60*(1), 20-43. doi: 10.2307/2181906

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

Richins, M. L. (1987). Media, materialism, and human happiness. In M. Wallendorf & P. Anderson (Eds.), *Advances in consumer research, Vol. 14* (pp.352-356). Provo, UT: Association for Consumer Research.

Richters, J. E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology, 43*(6), 366-405. doi: 10.1080/01973533.2021.1979003

Rönkkö, M., Lee, N., Evermann, J., McIntosh, C., & Antonakis, J. (2023). Marketing or methodology? Exposing the fallacies of PLS with simple demonstrations. *European Journal of Marketing*, ahead of print. https://doi.org/10.1108/EJM-02-2021-0099

Shevlin, M., Miles, J. N., & Bunting, B. P. (1997). Summated rating scales. A Monte Carlo investigation of the effects of reliability and collinearity in regression models. *Personality and Individual Differences*, *23*(4), 665-676. doi: 10.1016/S0191-8869(97)00088-3

Skipper, R. B., & Hyman, M. R. (1995). On foundations research in the social sciences. *The International Journal of Applied Philosophy, 10*(2), 23-38. doi: 10.5840/ijap19951019

Smedslund, J. (2016). Practicing psychology without an empirical evidence-base: The bricoleur model. *New Ideas in Psychology, 43*, 50-56. doi: 10.1016/j.newideapsych.2016.06.001

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

Spector, P. E. (1992). *Summated rating scale construction*: *An introduction.* Newbury Park, CA: Sage.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: The MIT Press.

Stroebe, W., Gadenne, V., & Nijstad, B. A. (2018). Do our psychological laws apply only to college students?: External validity revisited. *Basic and Applied Social Psychology, 40*(6), 384-395. doi: 10.1080/01973533.2018.1513362

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

Tate, R. L. (2010). *A compendium of tests, scales and questionnaires: The practitioner's guide to measuring outcomes after acquired brain impairment*. New York, NY: Psychology Press.

Trafimow, D. (2012a). The role of auxiliary assumptions for the validity of manipulations and measures. *Theory & Psychology*, *22*(4), 486-498. doi: 10.1177/0959354311429996

Trafimow, D. (2012b). The concept of unit coherence and its application to psychology theories. *The Journal for the Theory of Social Behaviour*, *42*(2), 131-154. doi: 10.1111/j.1468-5914.2011.00483.x

Trafimow, D. (2021a). Revisiting old-fashioned reliability and validity concerns. *Acta Scientific Neurology*, *4*(8), 81-87. https://www.actascientific.com/ASNE/pdf/ASNE-04-0409.pdf

Trafimow, D. (2021b). The underappreciated effects of unreliability on multiple regression and mediation. *Applied Finance and Accounting, 7*(2), 14-30. doi:10.11114/afa.v7i2.5292

Trafimow, D. (2022). The power of directional predictions in psychology. *Journal for the Theory of Social Behaviour*. doi: 10.1111/jtsb.12343

Trafimow, D., Hyman, M. R., & Kostyk, A. (2022). Are structural equation models theories and does it matter? *Journal of Global Scholars of Marketing Science*. doi: 10.1080/21639159.2022.2048960

Trafimow, D., Sheeran, P., Conner, M., & Finlay, K. A. (2002). Evidence that perceived behavioral control is a multidimensional construct: Perceived control and perceived difficulty. *British Journal of Social Psychology, 41*(1), 101-121. doi: 10.1348/014466602165081

Waters, F., & Stephane, M. (eds.) (2015). *The assessment of psychosis: A reference book and rating scales for research and practice*. New York, NY: Routledge.

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*(4), 41-78. doi: 10.1111/j.1540-4560.1969.tb00619.x

Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133. doi: 10.1007/s11336-003-0974-7

Number of Components	**Equation**
1	$r_{cs} = r_{c1}$
1 and 2	$r_{cs} = \frac{r_{c1}\sigma_1 + r_{c2}\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2r_{12}\sigma_1\sigma_2}}$
1, 2, and 3	$r_{cs} = \frac{r_{c1}\sigma_1 + r_{c2}\sigma_2 + r_{c3}\sigma_3}{\sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + 2(r_{12}\sigma_1\sigma_2 + r_{13}\sigma_1\sigma_3 + r_{23}\sigma_2\sigma_3)}}$
1, 2, 3, and 4	$r_{cs} = \frac{r_{c1}\sigma_1 + r_{c2}\sigma_2 + r_{c3}\sigma_3 + r_{c4}\sigma_4}{\sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + 2(r_{12}\sigma_1\sigma_2 + r_{13}\sigma_1\sigma_3 + r_{14}\sigma_1\sigma_4 + r_{23}\sigma_2\sigma_3 + r_{24}\sigma_2\sigma_4 + r_{34}\sigma_3\sigma_4)}}$
1, 2, 3, 4, and 5	$r_{cs} = \frac{r_{c1}\sigma_1 + r_{c2}\sigma_2 + r_{c3}\sigma_3 + r_{c4}\sigma_4 + r_{c5}\sigma_5}{\sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2 + 2(r_{12}\sigma_1\sigma_2 + r_{13}\sigma_1\sigma_3 + r_{14}\sigma_1\sigma_4 + r_{15}\sigma_1\sigma_5 + r_{23}\sigma_2\sigma_3 + r_{24}\sigma_2\sigma_4 + r_{25}\sigma_2\sigma_5 + r_{34}\sigma_3\sigma_4 + r_{35}\sigma_3\sigma_5 + r_{45}\sigma_4\sigma_5)}}$

Table 1. Expansions of Equation 3 for tests with one, two, three, four, or five components.

	Criterion item	**Item 1**	**Item 2**	**Item 3**
Inter-item correlation coefficients equal 0.1				
Criterion item	1			
Item 1	0.4	1		
Item 2	0.4	0.1	1	
Item 3	0.4	0.1	0.1	1
N	250	250	250	250
Inter-item correlation coefficients equal 0.3				
Criterion item	1			
Item 1	0.4	1		
Item 2	0.4	0.3	1	
Item 3	0.4	0.3	0.3	1
N	250	250	250	250
Inter-item correlation coefficients equal 0.8				
Criterion item	1			
Item 1	0.4	1		
Item 2	0.4	0.8	1	
Item 3	0.4	0.8	0.8	1
N	250	250	250	250
Inter-item correlation coefficients equal 0.9				
Criterion item	1			
Item 1	0.4	1		
Item 2	0.4	0.9	1	
Item 3	0.4	0.9	0.9	1
N	250	250	250	250

Table 2. Covariance matrices for SEM simulation.

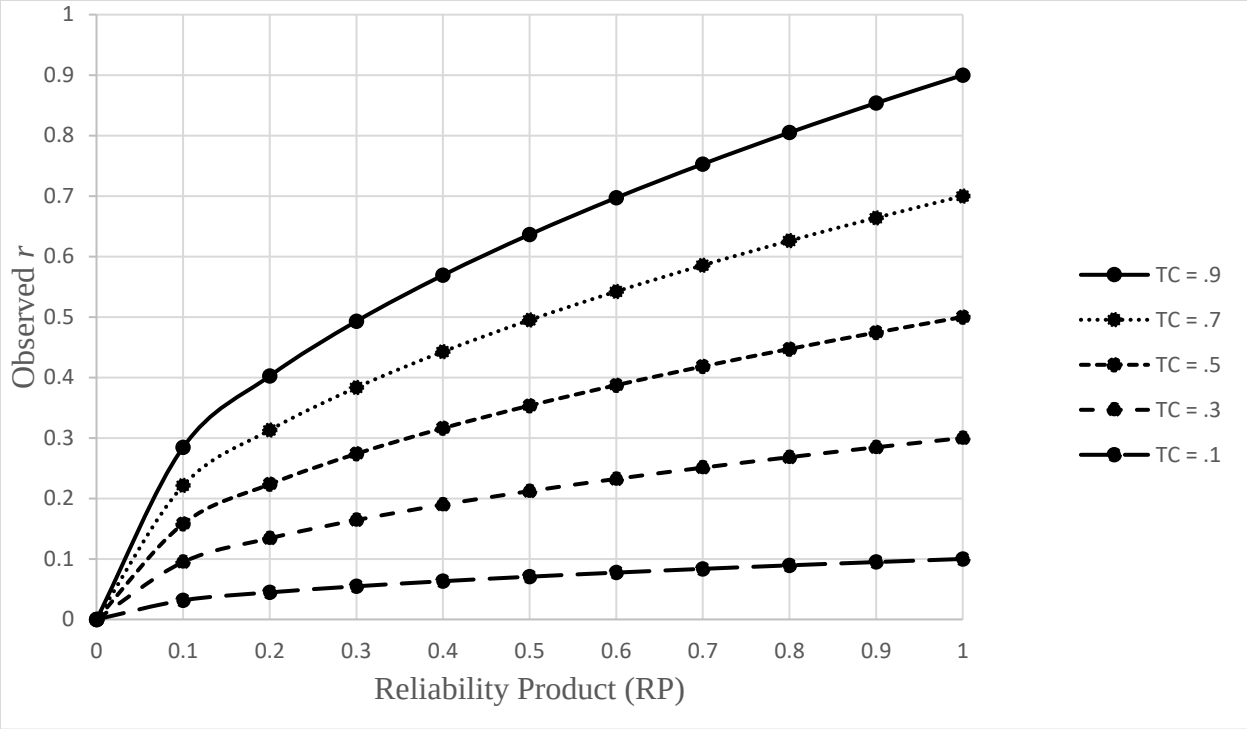

Figure 1. Along the vertical axis, the observed *r* is a function of the reliability product (RP) along the horizontal axis. Different curves represent a true *r* (TC) of 0.9 (top curve), 0.7, 0.5, 0.3, or 0.1 (bottom curve).

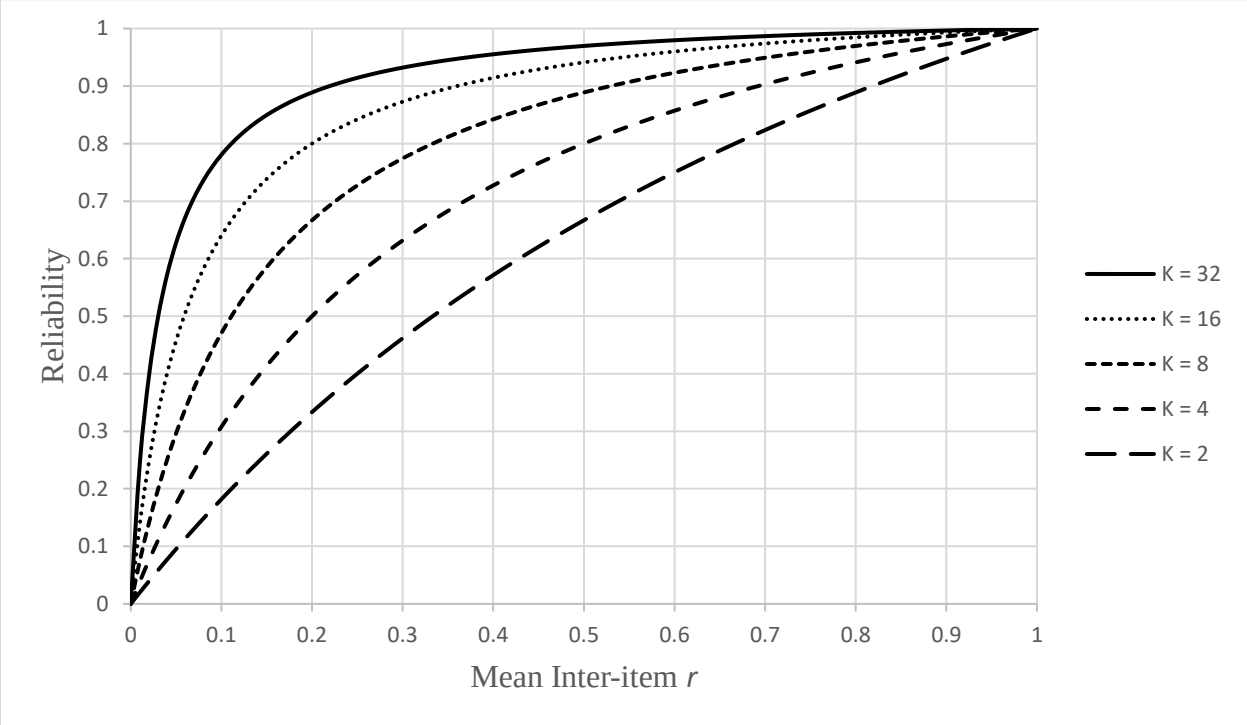

Figure 2. Along the vertical axis, reliability is a function of the mean inter-item *r* along the horizontal axis. The curves represent 32 items (top curve), 16 items, 8 items, 4 items, or 2 items (bottom curve).

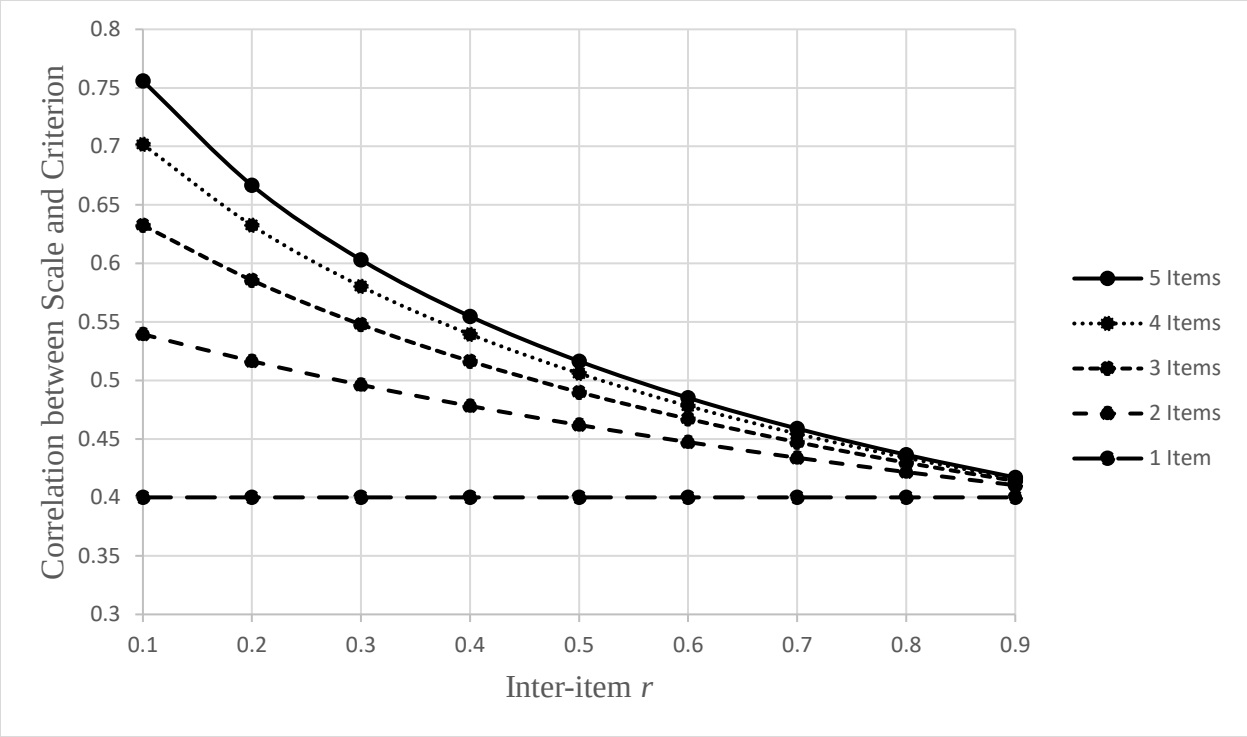

Figure 3. Along the vertical axis, the *r* between the scale and the criterion is a function of the inter-item *r* along the horizontal axis. Each item's *r* with the criterion is constant at 0.4 for all items regardless of number, i.e., 5 (top curve), 4, 3, 2, or 1 (bottom curve).

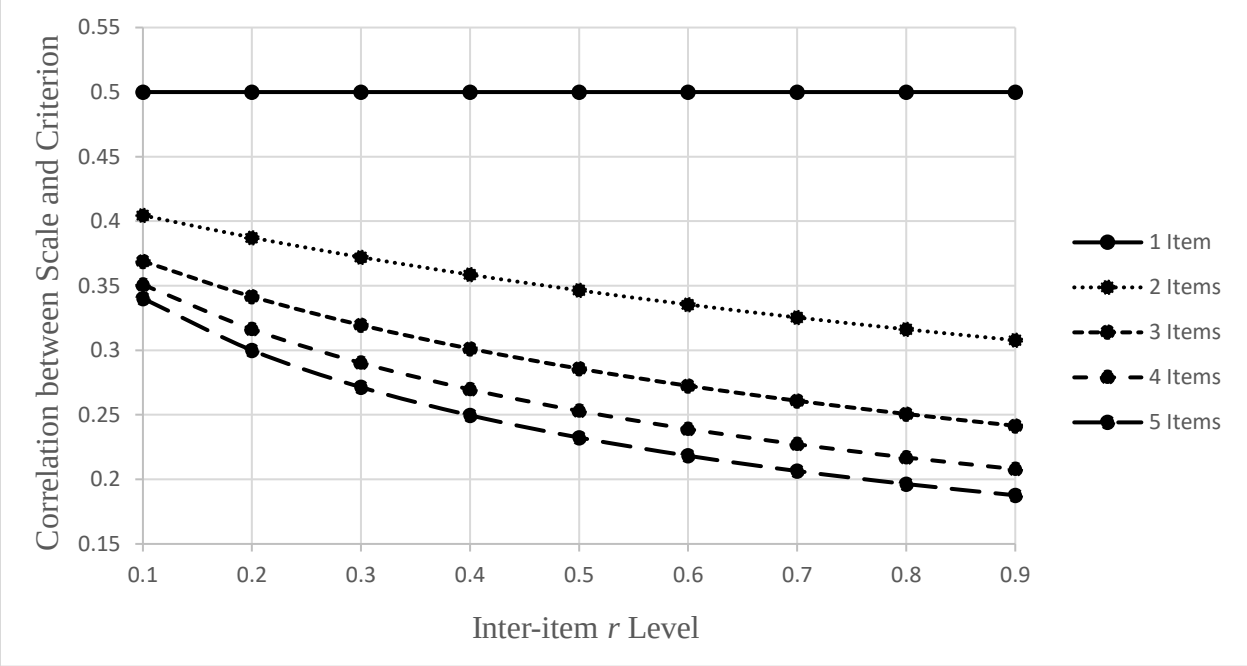

Figure 4. Along the vertical axis, the *r* between the test and the criterion is a function of the inter-item *r* level along the horizontal axis. The first component's *r* with the criterion is 0.50, and the other *r*s are 0.10. The number of items is set at 1 (top curve), 2, 3, 4, or 5 (bottom curve).

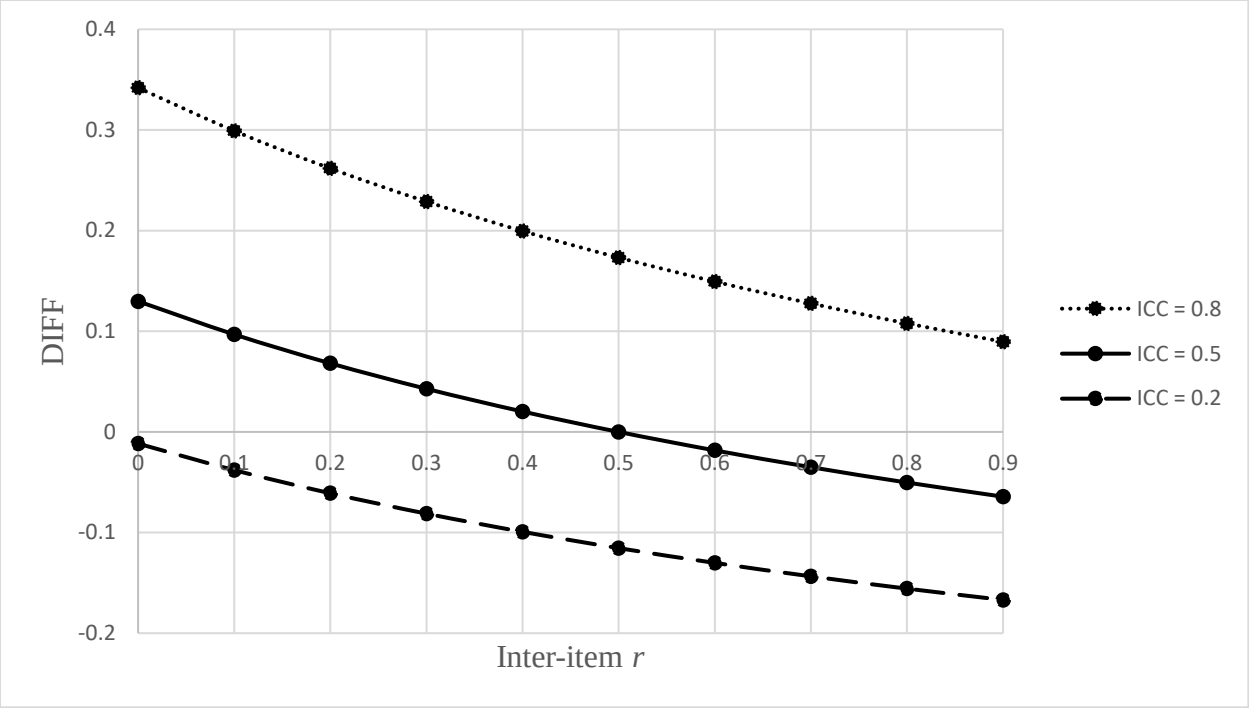

Figure 5. $r_{cs\prime} - r_{cs}$ (DIFF) is expressed along the vertical axis as a function of the inter-item r ($r_{12\prime}$) along the horizontal axis, with curves representing when $r_{c1\prime}$ (the inter-item r or ICC) is set at 0.8, 0.5, or 0.2. For all curves, $\sigma_1 = \sigma_{1\prime} = \sigma_2 = \sigma_{2\prime} = 1.0$, $r_{c2\prime} = r_{c2} = 0.50$, and $r_{c1} = 0.5$.

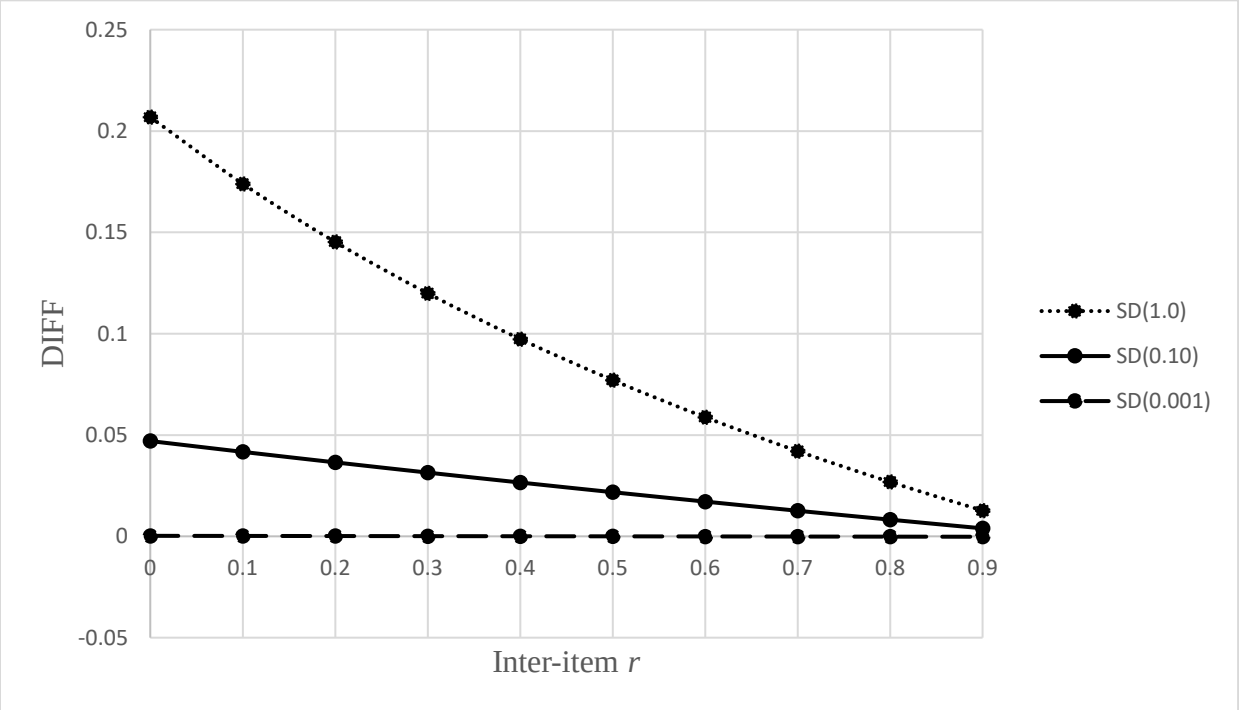

Figure 6. $r_{cs\prime} - r_{cs}$ (DIFF) is expressed along the vertical axis as a function of the inter-item r ($r_{12\prime}$) along the horizontal axis, with curves representing when $\sigma_{1\prime}$ (the standard deviation or SD) is set at 1.0, 0.10, or 0.001. For all curves, $r_{c1\prime} = r_{c1} = r_{c2\prime} = r_{c2} = 0.50$, $\sigma_{2\prime} = \sigma_2 = 1.0$, and $\sigma_1 = 0.001$.

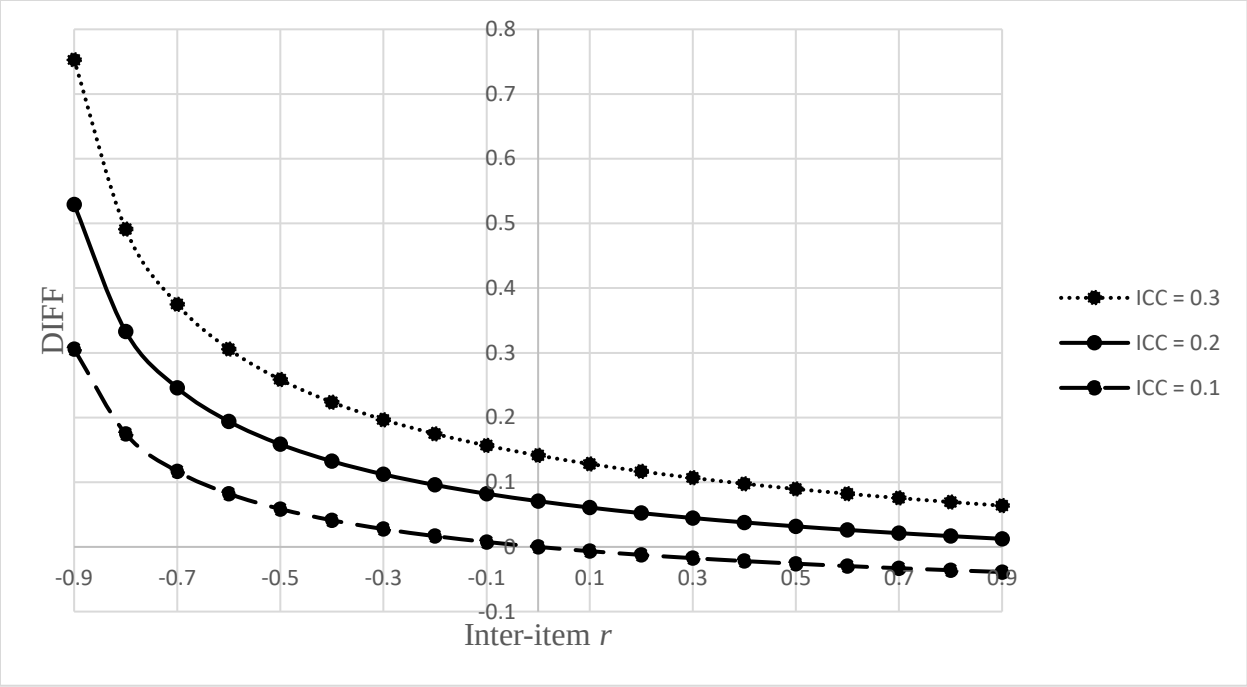

Figure 7. $r_{cs'} - r_{cs}$ (DIFF) is expressed along the vertical axis as a function of the inter-item *r* ($r_{12'}$) along the horizontal axis, with curves representing when $r_{c1'}$ (the item-criterion *r* or ICC) is set at 0.3, 0.2, or 0.1. For all curves, $\sigma_1 = \sigma_{1'} = \sigma_2 = \sigma_{2'} = 1.0$, $r_{c2'} = r_{c2} = 0.1$, and $r_{c1} = 0.1$.

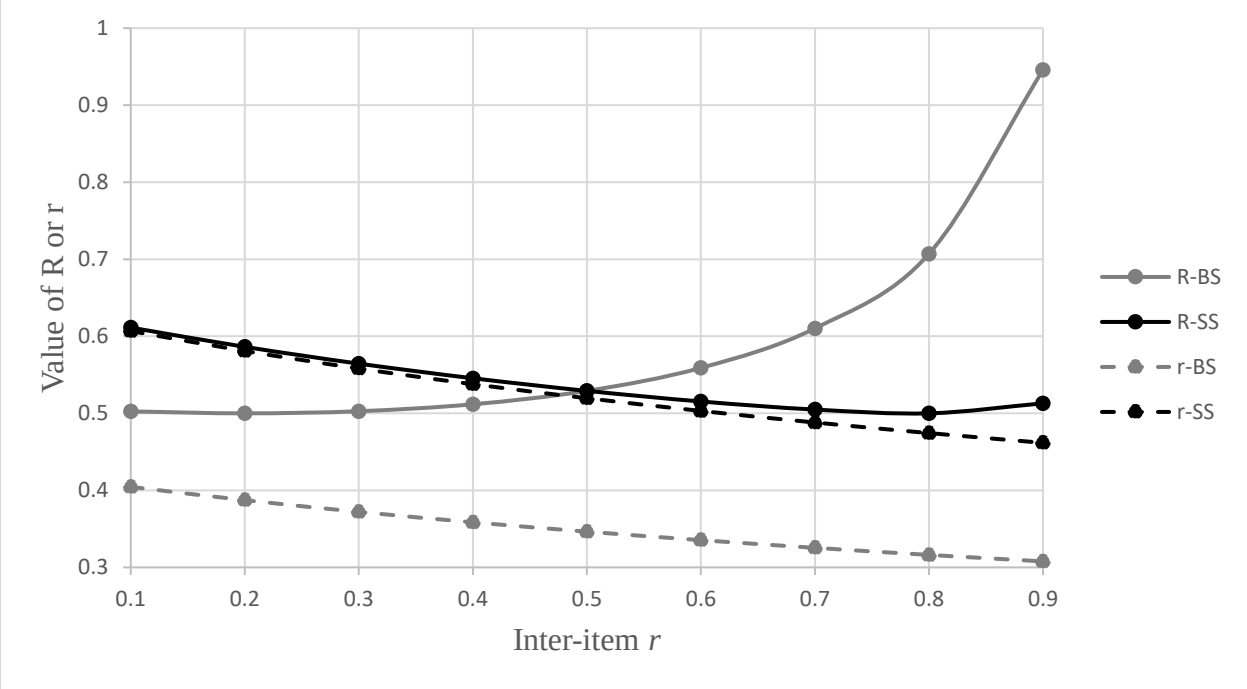

Figure 8. The value of the multiple *r* or bivariate *r* ranges along the vertical axis as a function of the inter-item *r* along the horizontal axis. Curves are in gray for a big spread (BS) in item-criterion *r*s (0.5 versus 0.1) or a small spread (SS) in item-criterion *r*s (0.5 versus 0.4). Solid curves represent multiple *r*s (R), and dashed curves represent bivariate *r*s (r).

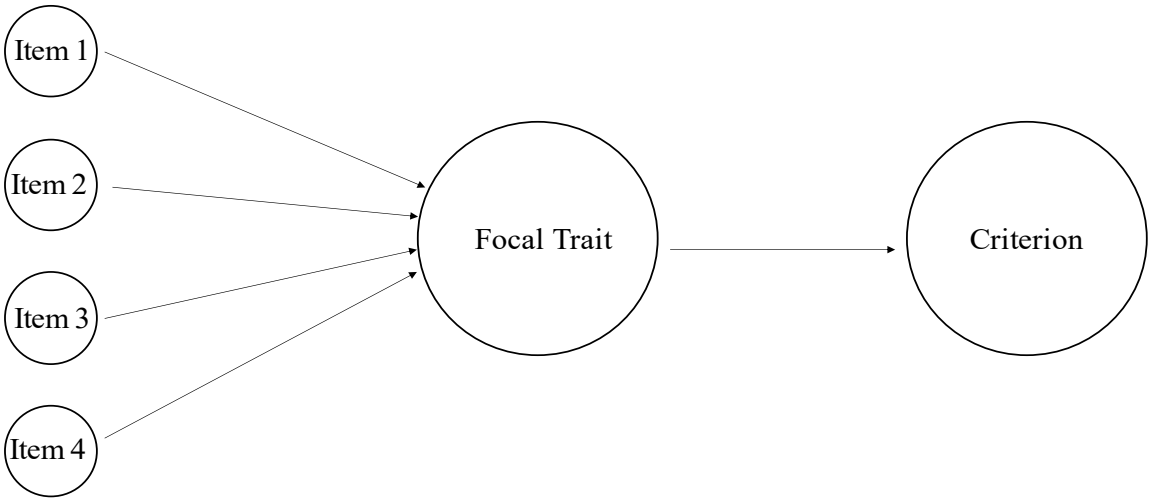

Figure 9. Model with the focal trait caused by the items and, in turn, causing the criterion

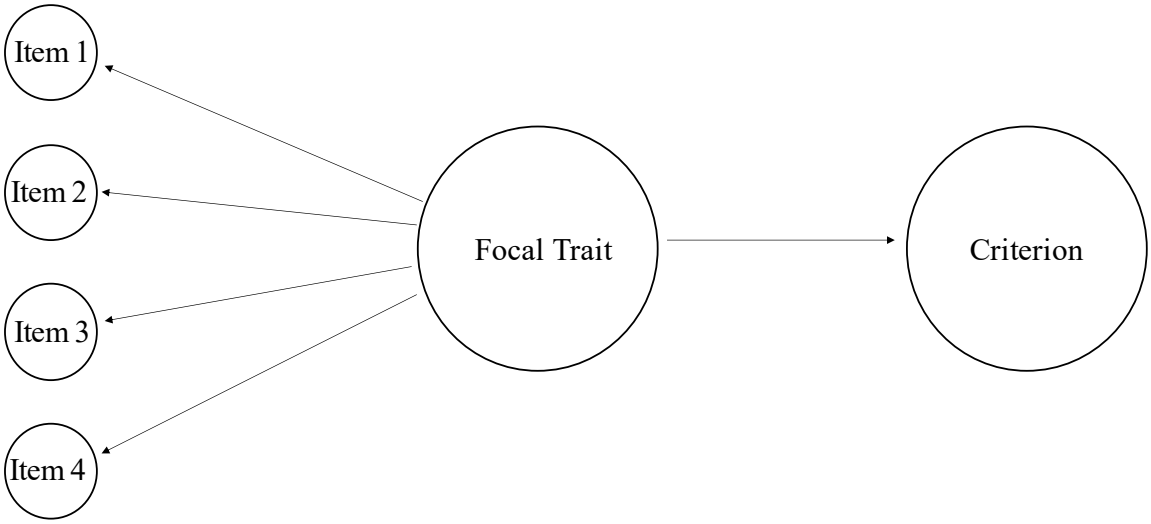

Figure 10. Model with the focal trait causing both the items and the criterion

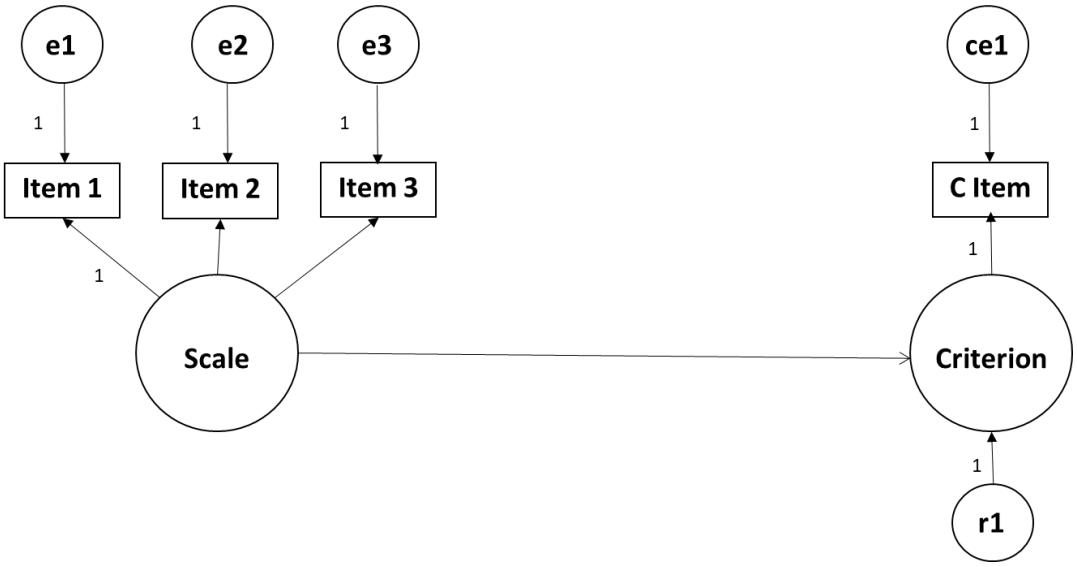

Figure 11. Model for SEM simulation