# Assessing measure congruence in nomological networks

**George R. Franke**
University of Alabama, Culverhouse College of Business, Tuscaloosa, Alabama, USA
(emeritus)
gfranke@cba.ua.edu

**Marko Sarstedt***
Otto-von-Guericke-University Magdeburg, Germany
Universitaetsplatz 2
39106 Magdeburg, Germany
*and*
Babeș-Bolyai University
Faculty of Economics and Business Administration
Strada Teodor Mihali 58-60
Cluj-Napoca 400591, Romania
marko.sarstedt@ovgu.de

**Nicholas P. Danks**
Trinity College Dublin, Ireland
nicholas.danks@tcd.ie

*Corresponding author

**Abstract**

Evaluations of convergent and discriminant validity are generally conducted by analyzing constructs in isolation or by comparing pairs of latent variables. These approaches ignore the broader nomological network that is intrinsic to a measure's construct validity, and fail to test the implications of either perfect correlations (convergence) or imperfect correlations (divergence). This paper proposes *congruence* assessment as a useful approach to simultaneously examining the relationships between multiple latent variables within nomological networks. Two measures are congruent if they have proportionally equal correlations with other constructs. We present measures for quantifying congruence within nomological networks, discuss statistical tests of significance, and demonstrate their performance in simulation studies. We reanalyze three published studies to contrast findings from congruence assessment versus traditional criteria for convergent and discriminant validity. We also discuss methodological and theoretical implications of congruence assessment, and suggest future research directions for both covariance- and composite-based structural equation modeling.

**Keywords:** Nomological network, structural equation modeling, discriminant validity, convergent validity, congruence

**Assessing measure congruence in nomological networks**

## 1. Introduction

Authors have no control over what parts of their work readers will use, or whether a citation implies actual readership. In a retrospective on the famous coefficient alpha (Cronbach, 1951), Cronbach opined that many citations to his paper were from people who had not read or even looked at it. Many authors who use the term "Cronbach's alpha" may not realize that, as the 1951 paper documents, equivalent formulas had previously been presented "numerous times in the psychological literature" (Cronbach & Shavelson, 2004, p. 396). An approach for assessing discriminant validity commonly attributed to Anderson and Gerbing (1988), which tests whether two latent variables correlate significantly less than unity, was previously described in multiple papers that they cite. Another criterion for discriminant validity is that the squared correlation between latent variables should be greater (not necessarily significantly) than the average squared standardized loadings of the variables' indicators (average variance extracted or AVE) (Fornell & Larcker 1981, pp. 45-46). Despite conceptual and empirical limitations (Franke & Sarstedt, 2019; Rönkkö & Cho, 2021), this heuristic has become a standard in measure validation (Voorhees et al., 2016). Fornell and Larcker (1981, p. 41) also build on Campbell and Fiske (1959) to describe an alternative but long-overlooked test of discriminant validity: whether the correlations *within* indicators of two constructs are greater than the correlations of the indicators *between* the constructs. Henseler et al. (2015) combine these sets of correlations in the heterotrait-monotrait ratio of correlations, or HTMT. This has become a popular criterion (as either a heuristic or test statistic) for assessing discriminant validity in structural equation modeling (SEM) (Franke & Sarstedt, 2019; Voorhees et al., 2016).

Another influential study also has an important neglected aspect. Cronbach and Meehl (1955) focused attention on the concept of construct validity, originally described as *congruent validity* (APA Committee on Test Standards, 1952). A core of their perspective is that understanding what traits or qualities a test measures depends on examining a network of relationships, rather than considering only bivariate relationships between latent variables: "Scientifically speaking, to 'make clear what something *is*' means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a *nomological network*" (Cronbach & Meehl, 1955, p. 290, italics in original). Though construct validity quickly became a key concept in measurement assessment, researchers generally focus on pairwise relationships and only rarely explicitly consider broader nomological networks.

Campbell and Fiske (1959) show the importance of considering multiple variables simultaneously in measure validation, though not explicitly within nomological networks. Instead, they emphasize the use of multiple independent measurement procedures, summarized in multitrait-multimethod (MTMM) matrices. A "trait is a response tendency which can be observed under more than one experimental condition and … can be meaningfully differentiated from other traits. The testing of these two propositions must be prior to the testing of other propositions to prevent the acceptance of erroneous conclusions…. [MTMM procedures] are intended to be as appropriate to the relatively atheoretical efforts typical of the tests and measurement field as to more theoretical efforts" (Campbell & Fiske, 1959, p. 100). MTMM matrices are used to assess convergent and discriminant validity, with convergence established by high correlations between related variables, and divergence by correlations that are not "too high" between measures of different theoretical concepts.

Campbell and Fiske (1959, p. 102) emphasize the use of multiple measurement procedures,

because for convergent validity, “with only one method, one has no way of distinguishing trait variance from unwanted method variance.” However, studies that use a single measurement procedure for related sets of items far outnumber those that explicitly consider MTMM correlations. Convergent validity is normally treated simply as a question of whether two latent variables X and Y correlate as expected, or whether indicators of items load on the expected constructs in conformance to a confirmatory factor analysis (CFA) model (e.g., Anderson & Gerbing, 1988).

Researchers have generally considered discriminant validity as a more complex question that can be addressed as described by Anderson and Gerbing (1988), Fornell and Larcker (1981), Henseler et al. (2015), and other researchers. Except for Campbell and Fiske’s (1959) MTMM guidelines, something these procedures have in common is ignoring patterns of similarity and dissimilarity across multiple variables simultaneously. This practice can, however, produce misleading results. For example, latent variables X, Y, and Z may correlate perfectly, yet Y and Z do not correlate at all. This pattern creates a contradiction: If X and Y are the same, and X and Z are the same, how can Y be completely different from Z? The correlation patterns in this hypothetical example are extreme (and inconsistent with an underlying common-factor model, even if the model fits perfectly), but less extreme examples are common. Researchers will frequently encounter situations in which two latent variables are highly correlated, yet have meaningfully different relationships with other latent variables. For example, X and Y can have a very high correlation of r = .9, but if X correlates say r = .5 with Z, then Y’s correlation with Z can range from as high as r = .83 to as low as r = .07 (Carlson & Herdman, 2012). This range of possible correlations illustrates that the relationship between X and Y can only be interpreted in light of their relationships with other variables that form a broader nomological network.

Addressing the limitations of conventional pairwise approaches for assessing convergent and discriminant validity, we argue theoretically and demonstrate empirically how nomological networks can be used to assess to what extent pairs of latent variables both converge and diverge as expected—that is, are *congruent*. To do so, the next section shows how congruence across nomological networks can be summarized in terms of overall effect sizes and tested for significance. Simulations demonstrate the relative performance of two procedures for testing the significance of congruence between variables. Example analyses illustrate the performance of the tests with real data. The final section discusses theoretical and methodological implications, and suggests future research directions.

While our descriptions primarily relate to factor-based modeling, as executed in CFA and covariance structure analysis (Jöreskog, 1970), the principles underlying congruence also apply to composite-based methods such as generalized structured component analysis (Hwang & Takane, 2004) and partial least squares (Wold, 1982). In the following we distinguish between congruence testing in factor-based versus composite-based methods, where relevant.

## 2. Nomological networks and congruence

Multiple terms in the literature express the concept that two instruments are measures of the same construct, including collapsible, collinear, confounded, congruent, equivalent, interchangeable, parallel, proxy, and redundant. In conventional language, *congruence* has the usual English meaning based on the underlying Latin concept of agreement and lack of conflict. In geometry, *congruent* describes figures that have the same form. Similarity of factor loadings is often described using a *congruence coefficient* (e.g., Lorenzo-Seva & ten Berge, 2006), as shown below for variable correlations (Eq. 2). Raykov et al. (2015) consider congruence in terms of whether indicators of different latent variables can be combined as indicators of a general overall

construct. Thus, *congruence* and *congruent* appear to be useful terms to use in examining relationships between latent variables within nomological networks. Importantly, congruence as developed here is not simply a novel test of either convergent or discriminant validity. Instead, it directly addresses questions of construct validity as described by Cronbach and Meehl (1955).

Congruence implies that the focal latent or observed variables have the same *proportional* correlations with other variables:

$$r_{XZ_i} = \rho\, r_{YZ_i} \tag{1}$$

for X and Y relative to a network of variables $Z_i$ ($i > 1$), where $\rho$ is a nonzero constant (Hunter, 1973). The special case of $\rho = 1$ has been widely studied (e.g., Anderson & Gerbing, 1988; Jöreskog, 1971; Lord, 1957; Raykov, West & Trayner, 2015; Rönkkö & Cho, 2021; van der Sluis et al., 2005). However, this case is often unrealistic. Various authors have pointed out that common (but not necessarily equal) patterns of correlations with other latent variables are necessary for two latent variables to have convergent validity (e.g., Anderson & Gerbing, 1982; Hunter, 1973; Messick, 1989). Conversely, differing relationships with other latent variables imply a lack of convergence, thereby supporting discriminant validity (e.g., Brooke et al., 1988; Schwab, 1980; Zeller & Carmines, 1980). As an example, Kroger (1968) suggests that height and weight are correlated but conceptually distinct, and may have differing correlations with other variables such as skills in various sports.

### *2.1. Measuring congruence*

An effect size for congruence, known as the congruence coefficient $r_c$, is calculated as

$$r_c = \sum r_{XZ_i} r_{YZ_i} / (\sum {r_{XZ_i}}^2 \sum {r_{YZ_i}}^2)^{.5} \tag{2}$$

where the summations are over X, Y, and $Z_i$, including $r_{XY}$ and the reliabilities $r_{XX}$ and $r_{YY}$ on the diagonal. If the reliabilities are unknown, 1 or other plausible values can be used (Hunter, 1973).

This formula dates as far back as Burt (1937), and has been revived or rediscovered by other authors in succeeding decades (e.g., Anderson & Gerbing, 1982; Hunter, 1973; Steenbergen, 2000; Tryon, 1958). Alternative labels for Eq. 2 include the proportionality or similarity coefficient, cosine correlation, cosine distance, and cosine similarity; when applied to comparing factor loadings, it is often called Tucker's phi. Tryon (1958) omits $r_{XY}$ and the reliabilities from the calculation and calls the square of the resulting value the *index of proportionality*.

Multiple resources facilitate $r_c$ calculations, including straightforward spreadsheets and SAS PROC DISTANCE (see Appendix 1). R and Matlab offer built-in functions, and websites give online calculators (e.g., https://scistatcalc.blogspot.com/2015/11/cosine-similarity-calculator.html#). Because Eq. 2 can be expressed in terms of scalar products and vector norms (e.g., Steenbergen, 2000), it can also be calculated using matrix algebra procedures. Regression software that allows for suppression of the intercept gives $r_c$ as the standardized effect of X on Y or vice versa, or as the (correctly signed) square root of the variance explained.

The congruence coefficient $r_c$ equals the cosine of the angle between two vectors X and Y. Cosines are not very intuitive as effect sizes, especially as $r_c$ approaches 1. The actual angle, given by the inverse or arc cosine of $r_c$, is a more interpretable indicator of similarity:

$$\theta = cos^{-1}(r_c) \tag{3}$$

To convert from cosines to angles, using the absolute value of $r_c$ and taking the angles in degrees rather than radians provides a 90-degree range of possible values ($0 \leq \theta \leq 90$). $r_c = 1$ ($\theta = 0$) implies complete congruence, but useful standards for small, medium, and large effects are less obvious. One approach is to "put into the diagonal whatever [reliability] value seems appropriate and ask if the similarity is 'high'" (Hunter, 1973, p. 60). Anderson and Gerbing (1982, p. 458) suggest that a cutoff of $r_c = .80$ ($\theta = 37$) "is not meant to be interpreted as a rigid

value, but may serve as a useful guideline for model construction." They do not give a rationale for this value, but other authors have accepted it as "the conventional cutoff" (Steenbergen, 2000, p. 273). However, as shown next, simulations indicate that observed values of $r_c$ and $\theta$ strongly depend on the number of Z-variables in the nomological network.

### *2.2. Congruence values in random correlation matrices*

To explore the distribution of $r_c$ and $\theta$, we conduct a simulation experiment. The approach of the simulation experiment is to identify at which thresholds, for varying sizes (complexity) of nomological network, a value for $r_c$ or $\theta$ could be said to be significantly different from that of a random correlation matrix.

We consider nine nomological networks of varying complexity: X, Y, and two to ten variables $Z_i$ (i.e., max(i) = I = 2,...,10) and k being the total number of variables (i.e., k = 4,…,12). For each network, we generated 10,000 random correlation matrices (i.e., square positive semidefinite matrices with 1s on the diagonal and values from -1 to +1 elsewhere, rather than random samples from a specified population correlation matrix). In each replication, the first two columns serve as the X and Y variables of interest, providing a total of 90,000 $r_c$ and $\theta$ values. The box plots in Fig. 1 summarize the distributions of the absolute values of the coefficients (which would otherwise range from -1 to +1 with means and medians near 0) and corresponding angles $\theta$. Each shaded box gives the interquartile range of the estimates (bottom $25^{th}$ to top $25^{th}$ percentiles). The horizontal line within each box is the distribution median, and the diamond figures show the means. The lengths of the lines extending from the boxes are 1.5 times the interquartile range, or less if constrained by the range of the data. Circles depict outliers beyond these extremes. Line graphs quantify high degrees of congruence at the $90^{th}$, $95^{th}$, and $99^{th}$ percentiles of $r_c$, and the $1^{st}$, $5^{th}$, and $10^{th}$ percentiles of $\theta$.

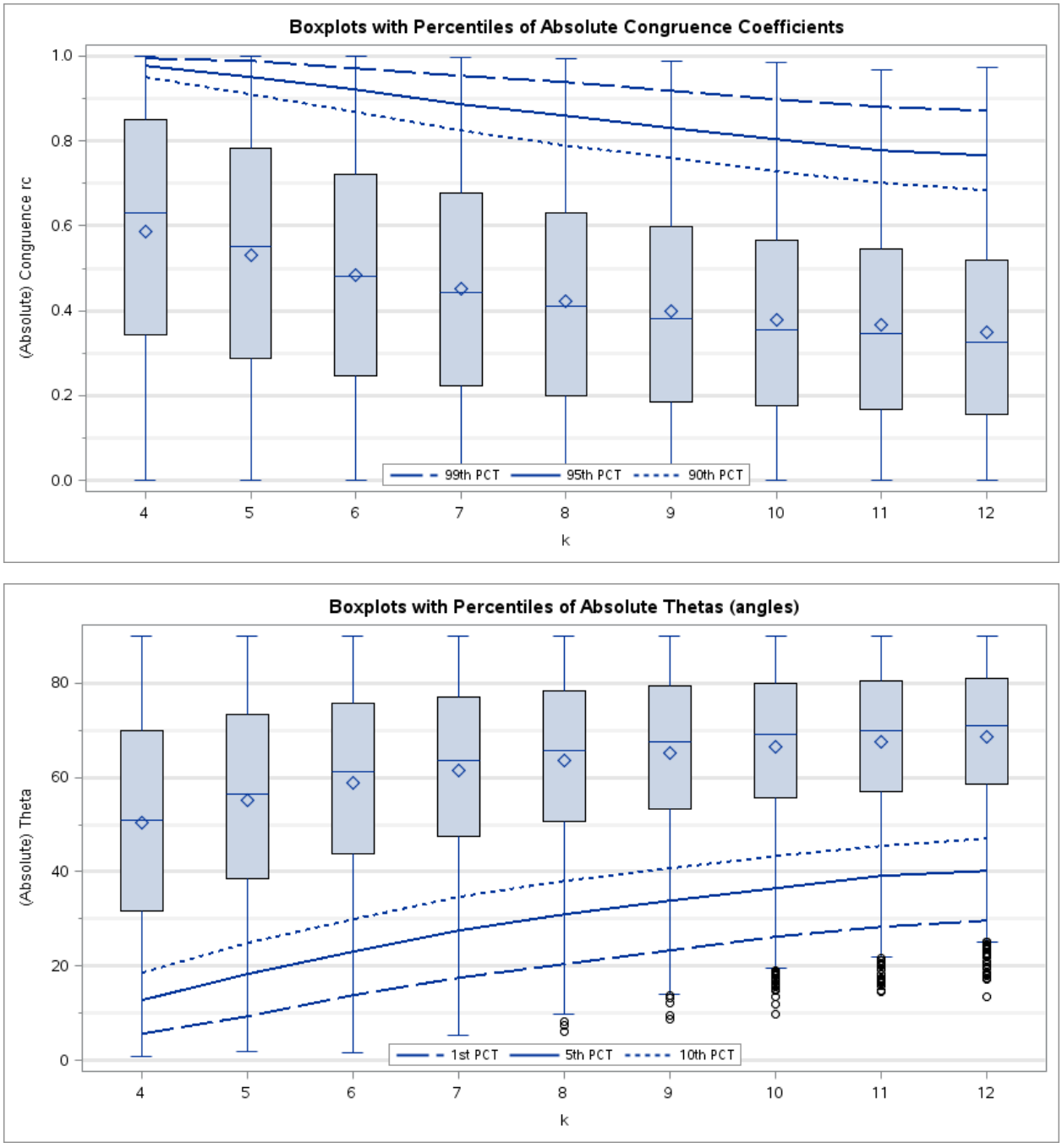


**Fig. 1**. Congruence values in random correlation matrices

**Notes:** Box plots show the distribution of the absolute value of $r_c$ (top panel) and its corresponding θ (bottom panel) for 10,000 random correlation matrices for k = 4 to 12 variables (X, Y, and 2 to 10 $Z_i$ variables). After taking absolute values, the possible range of $r_c$ is 0 to 1 and of θ is 0 to 90. Each box spans the 25$^{th}$ to 75$^{th}$ percentiles; horizontal lines and diamonds are the median and mean of the distribution respectively. For $r_c$, lines connecting the distributions, from lowest to highest, are the 90$^{th}$, 95$^{th}$, and 99$^{th}$ percentiles; for θ, the lines are the 1$^{st}$, 5$^{th}$, and 10$^{th}$ percentiles.

The leftmost boxes in Fig. 1 are based on just four random variables, the smallest nomological network for which $r_c$ can be meaningfully calculated (representing X, Y, $Z_1$, and $Z_2$). The mean and median $r_c$ (upper panel) in this case are approximately .60. The 75$^{th}$ percentile is about .85, and the 90$^{th}$, 95$^{th}$, and 99$^{th}$ percentiles range from .95 to 1. Average percentiles decrease with each increment in k, but observed values often remain high. For example, $r_c$ = .80 is the 90$^{th}$ percentile with k = 8 and the 95$^{th}$ percentile with k=10. Therefore, a .80 cutoff could be a reasonable criterion for larger nomological networks, but it would give many false positive signals of congruence for smaller numbers of variables.

The θ distributions (lower panel) in Fig. 1 are similar to $r_c$ except for the conceptual reversal of the vertical axis (e.g., as mentioned above, $r_c$ = 1 implies θ = 0). The extreme values are clearly greater than zero as k increases, and the gaps between the 10$^{th}$, 5$^{th}$, and 1$^{st}$ percentiles are larger for θ than for $r_c$. Therefore, θ is less likely than $r_c$ to falsely signal a very high degree of congruence.

These results are based on simulated variables in random correlation matrices. Measure validation more commonly examines relationships between latent variables with multiple indicators and with differing degrees of latent correlations, sample sizes, and measure reliability. The next section presents conceptual and applied approaches for assessing congruence with latent variables, then summarizes empirical findings for congruence effect sizes in a more complex simulation design.

## 3. Testing for congruence: Concepts and simulations

### *3.1. Conceptual view of congruence testing*

The Anderson-Gerbing criterion for discriminant validity applies procedures for testing

whether two observed or latent variables correlate perfectly when accounting for measure reliability. However, in terms of perfect correlation in nomological networks, these procedures do not actually test the underlying model (Rönkkö & Cho, 2021; van der Sluis et al., 2005). If the disattenuated correlation $\phi_{XY} = 1$, then $\phi_{XZi}$ should equal $\phi_{YZi}$ for each $Z_i$. This pattern implies that $\rho = 1$ as shown in Eq. (1), which Hunter (1973) calls *exact equivalence*. Comparing two models, one with $\phi_{XZi}$ and $\phi_{YZi}$ free to vary and another where they are constrained to be equal (along with other constraints, such as $\phi_X$ and $\phi_Y$ having unit variances) is straightforward; the chi-square difference between the model fits tests the significance of the constraints. Allowing for proportional rather than equal correlations ($\rho \neq 1$), which Hunter (1973) calls *weak equivalence*, requires another approach.

Fig. 2 illustrates why measures of the same construct have model-implied proportional correlations with other measures. Fig. 2 (upper panel) shows a CFA model with four latent variables: $\xi_1$, $\xi_2$, $\xi_i$ (representing constructs $\xi_3$, ..., $\xi_{k-1}$), and $\xi_k$. For simplicity, only three indicators are shown for just two constructs, $\xi_1$ and $\xi_k$. With standardized latent variables and indicators, the products of the loadings ($\lambda$) equal the correlations between the indicators of a given construct; for example, for $\xi_1$, $\lambda_{11} * \lambda_{21} = r_{x11x21}$ and $\lambda_{11} * \lambda_{31} = r_{x11x31}$. Given these relationships, the ratio between the loadings $\lambda_{21}$ and $\lambda_{31}$ is the same as the ratio between the correlations $r_{x11x21}$ and $r_{x11x31}$. Similar results hold for the correlation between the indicators of two different constructs, which is a function of their loadings and the correlation between the constructs ($\phi$, not labeled in Fig. 2 but represented by the curved double-headed arrows); for example, for $\xi_1$ and $\xi_k$, $\lambda_{11} * \phi_{\xi 1 \xi k} * \lambda_{1\xi} = r_{x11x1\xi}$. The model-implied proportional relationships give internal consistency within constructs and external consistency between constructs. To the

extent that the observed correlations do not match the model-implied proportional correlations, the model does not fit the data and therefore does not indicate congruence.

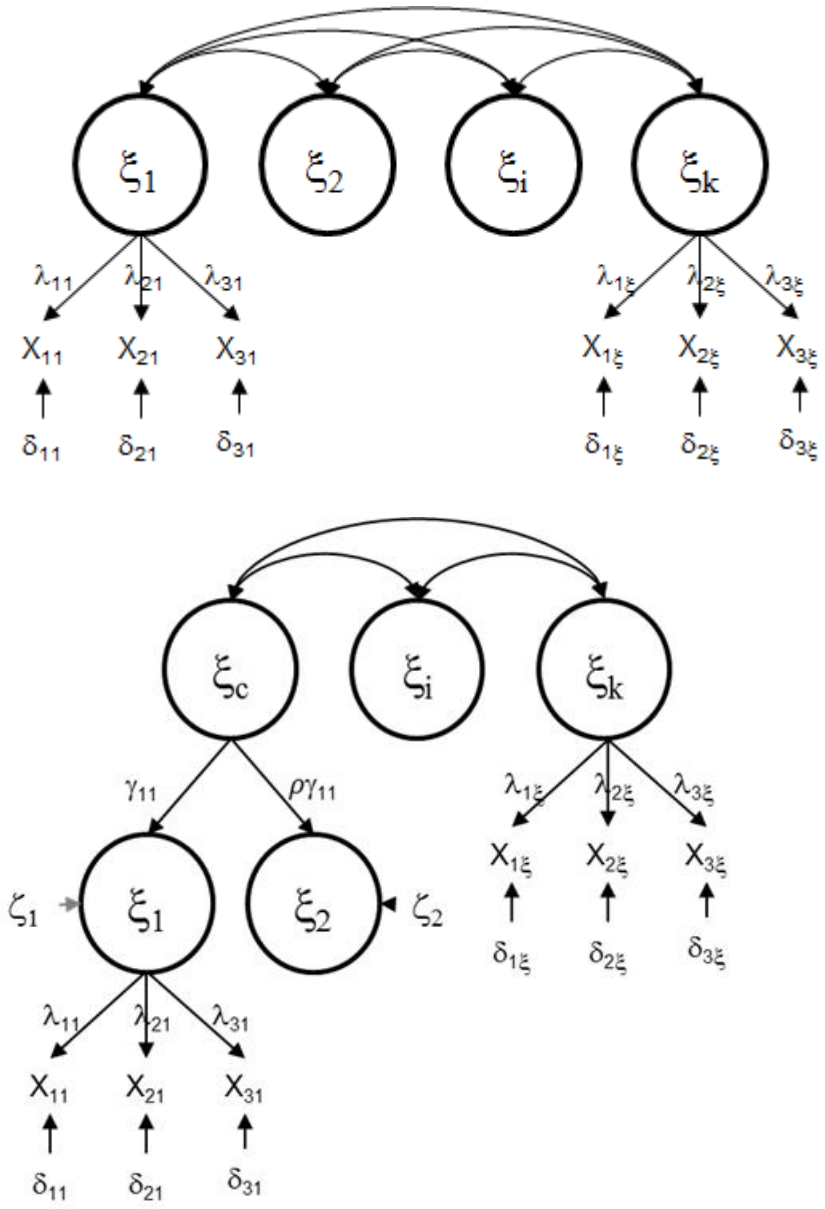


**Fig. 2**. CFA model

**Notes:** The upper panel is a confirmatory factor analysis with four latent variables. The lower panel shows two

lower-order indicators of one higher-order construct, plus two other latent variables. In each panel, three observed indicators are shown for just two of the latent variables.

The CFA model implies proportional relationships between indicators, but not between constructs. Using two latent constructs as indicators of a higher-order construct, illustrated in Fig. 2 (lower panel) by $\xi_1$, $\xi_2$, and $\xi_c$, does impose proportional but not necessarily equal relationships between the lower-order latent variables $\xi_1$ and $\xi_2$ and the other latent variables, $\xi_i$ and $\xi_k$ in the figure. The approaches discussed by van der Sluis et al. (2005) and Rönkkö and Cho (2021) for testing perfect correlations between latent variables correspond to setting $\gamma_{11} = \rho = 1$ in the loadings of $\xi_1$ and $\xi_2$ on $\xi_c$, if $\xi_1$ and $\xi_2$ are standardized to unit variances. Equal fit within sampling error for the models in Fig. 2 supports the proportionality constraints and implies congruence between $\xi_1$ and $\xi_2$.

Fig. 2 shows reflective models with latent variables having multiple indicators, but the concepts illustrated also apply to assessing the congruence of individual measures (or in analyzing published results that show correlations between latent variables). In these applications, $\xi_c$ in Fig. 2 (lower panel) would be the only latent variable. The higher-order model fit would directly test for congruence because the correlation model underlying Fig. 2 (upper panel) would be saturated and fit perfectly. Therefore, like HTMT (Henseler et al., 2015), congruence analysis can be applied to observed correlations without depending on a reflective model of underlying latent variables.

### *3.2. Procedures for congruence testing*

Fig. 2 is a graphical illustration of model-implied congruence, but in practice it is not a very useful test procedure. If the indicators of the higher-order construct are actually not close to

congruent, the model estimates will often fail to converge and produce usable results.[1] Fortunately, more robust test procedures exist, as in the Mplus software. One is to use MODEL CONSTRAINT statements to create new variables representing ratios of the latent-variable correlations. For example, if X and Y are the test variables and $Z_i$ are the other constructs in the nomological network, then the latent correlations in the model specification could be labeled, say, XZ1, XZ2, … YZi (see Appendix 1 for example code). Then new variables could be calculated as RATIO1 = XZ1 / YZ1, etc. Finally, MODEL TEST statements would produce a simultaneous test of equality across the ratios (i.e., whether they all equal the ratio shown as $\rho$ in Eq. 1). The program output gives the test result as a Wald statistic with I – 1 (equals k – 3) degrees of freedom.

Another approach in Mplus is to create a new variable, say rho, then use it in model constraint statements (e.g., XZ1 = RHO*YZ1, XZ2 = RHO*YZ2, etc. Again, see Appendix 1 for example code). Then comparing chi-square values of model fit for the constrained and unconstrained models tests for significance. Though conceptually this is exactly what the higher-order factor specification in Fig. 2 (lower panel) imposes on the estimates, in practice the ratio-constrained model is far more likely to converge. Both procedures, the Wald test (called WALD) and model comparisons (called DIFF), are used in the following simulations.

Unfortunately, the WALD and DIFF tests cannot be readily transferred to composite-based SEM methods. In Appendix 2, we discuss details and suggest an alternative approach.

### *3.3. Simulation design*

[1] The same is true of the equality constraints discussed by van der Sluis et al. (2005), which perhaps helps account for why their approach is not more widely used in assessing discriminant validity than the simpler alternative popularized by Anderson and Gerbing (1988).

The simulation results in Fig. 1 give effect sizes $r_c$ and $\theta$ for random correlation matrices. In the following, we examine the performance and relative usefulness of the WALD and DIFF tests under conditions of variable congruence (null hypothesis) and incongruence (alternative hypothesis). We incorporate situations commonly encountered in empirical research by systematically varying factors such as sample size, number of latent variables, and correlations between the pairs of focal variables ($\phi_{XY}$). We also examine whether differing reliabilities ($r_{XX}$ and $r_{YY}$) affect congruence inferences, as has been found in testing discriminant validity (Franke & Sarstedt, 2019; McDonald, 1999).

Each latent variable is measured by three indicators. All Y and $Z_i$ indicators have population correlations of .5, giving loadings of .707 and reliabilities of $\alpha$ = CR = .75. All Z-construct correlations are $\phi$ = .3. The five manipulated factors are:

- k: 5 or 7 total variables (such that I = 3 or 5 latent variables $Z_i$).
- n: 100, 250, and 500.
- $\phi_{XY}$: .707 (giving a minimum 50% X-Y shared variance across conditions), .80, and .90.
- Congruent/Incongruent: $XZ_i$ and $YZ_i$ correlations range from relatively small, .1, to relatively large, .5. Correlations are equal or proportional in the congruent conditions, and the other two are somewhat higher (.3 versus .1) or lower (.3 versus .5) in the incongruent conditions:
    - Congruent: $\phi_{XZi} = \phi_{YZi}$ = .1, .3, and .5 for i = 1 to 3, and .1, .2, .3, .4, and .5 for i = 1 to 5, respectively (giving population $r_c$ values ranging from .985 to 1.000 and $\theta$ values from 1.1 to 9.9, depending on $\phi_{XY}$ and CR).
    - Incongruent: $\phi_{XZi}$ = .3 for all $Z_i$ when k = 5, and .3, .2, .3, .4, and .3 for $Z_i$ = 1 to 5 when k = 7 (giving population $r_c$ values ranging from .967 to .973 and $\theta$ values from

11.0 to 14.8, depending on $\phi_{XY}$ and CR). $\phi_{YZi}$ values are the same as in the congruent condition.

- Equal/Unequal X reliabilities: Without changing congruence or incongruence, the higher X-indicator correlations in the unequal-reliability condition increase X's reliability relative to Y, and therefore $\phi_{XZi}$, by $(.85 / .75)^{.5} = 1.065 = 6.5\%$ (Nunnally, 1978, p. 239).
    - Equal: X-indicator correlations = .5 ($\lambda = .707$ and reliability = $\alpha$ = CR = .75).
    - Unequal: X-indicator correlations = .654 ($\lambda = .809$ and reliability = $\alpha$ = CR = .85).
- Analysis methods: The WALD and DIFF tests described above are different approaches to imposing proportionality between the $XZ_i$ and $YZ_i$ correlations with d.f. = 2 or 4 for both tests, depending on whether k = 5 or 7 (I = 3 or 5).
    - WALD: Creates new variables equal to the $XZ_i$ / $YZ_i$ ratios and tests them for equality. d.f. = I - 1 because only two or four comparisons are needed to test equality of all three or five ratios. The latent correlations in the CFA model are unconstrained and unaffected by the new ratio variables; lack of fit is due entirely to random sampling in the measurement model.
    - DIFF: Compares the CFA fit in the WALD analysis with that of a constrained model (on the same simulation data) with $XZ_i$ = RHO*$YZ_i$, with RHO calculated to maximize model fit. d.f. = I - 1 because I correlations and one RHO value (3 + 1 = 4 or 5 + 1 = 6) are estimated in the constrained condition rather than I XZ plus I YZ correlations (3 + 3 = 6 or 5 + 5 = 10) distinct correlations in the unconstrained CFA.

The specifications of measurement models, constructs, and construct correlations are typical of other simulation designs in applied research (e.g., Paxton et al., 2001). We used Mplus

capabilities for Monte Carlo designs to generate 1000 samples for each of the 72 experimental conditions (3 x $\phi$, 3 x n, 2 x I, 2 x congruence/incongruence, and 2 x reliability; the WALD and DIFF tests apply to the same data rather than distinct samples). Mplus saved the output for the total of 72,000 experimental replications in separate files, which we merged for further analyses using SAS software.

## 4. Simulation results

Some (306 or 0.4% of the total) of the replications did not provide usable results, primarily in conditions with I = 3 or n = 100 or both, but the effective sample size is almost or exactly 1,000 replications for each combination of design factors. Therefore, power to detect all but the smallest effects is very high. However, other than the congruence manipulation, most effects of the simulation factors are small to nearly zero.

### *4.1. GLM and effect sizes for simulation manipulations*

Using PROC GLM (general linear model) in SAS gives main effects and interactions for four outcome variables: the congruence coefficient $r_c$ and the corresponding angle $\theta$, plus the chi-square values from the WALD and DIFF tests. Table 1 shows selected results, with the semipartial omega-squared ($\omega^2$) as the effect size indicator. Very similar relative magnitudes of effects are found across other effects that PROC GLM reports, such as noncentrality parameters and partial eta-squares.

**Table 1.** Effect sizes of the GLM on Congruence Coefficients $r_c$ and Theta, WALD Test results, and DIFF Test results

| Effects | df | Congruence Coefficient | | | Theta | | | WALD Test | | | DIFF Test | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F Value | Pr > F | Semipartial Omega-Square | F Value | Pr > F | Semipartial Omega-Square | F Value | Pr > F | Semipartial Omega-Square | F Value | Pr > F | Semipartial Omega-Square |
| Congruent | 1 | **40,698.90** | **0.00** | **0.314** | **71,476.00** | **0.00** | **0.421** | **50,243.20** | **0.00** | **0.236** | **171,665.00** | **0.00** | **0.420** |
| Z | 1 | **12.25** | **0.00** | **0.000** | **105.59** | **0.00** | **0.001** | **17,531.40** | **0.00** | **0.082** | **4,177.02** | **0.00** | **0.010** |
| Congruent*Z | 1 | **142.31** | **0.00** | **0.001** | **326.08** | **0.00** | **0.002** | **3,074.01** | **0.00** | **0.014** | **452.94** | **0.00** | **0.001** |
| Equal | 1 | **1,048.51** | **0.00** | **0.008** | **1,716.28** | **0.00** | **0.010** | **474.67** | **0.00** | **0.002** | **4,124.57** | **0.00** | **0.010** |
| Congruent*Equal | 1 | 1.96 | 0.16 | 0.000 | **79.08** | **0.00** | **0.001** | **357.65** | **0.00** | **0.002** | **4,094.68** | **0.00** | **0.010** |
| Z*Equal | 1 | 0.17 | 0.68 | 0.000 | 0.14 | 0.71 | 0.000 | **44.85** | **0.00** | **0.000** | 0.47 | 0.49 | 0.000 |
| Congruent*Z*Equal | 1 | 0.32 | 0.57 | 0.000 | 0.02 | 0.88 | 0.000 | **50.53** | **0.00** | **0.000** | 0.89 | 0.34 | 0.000 |
| n | 2 | **5,687.51** | **0.00** | **0.088** | **7,594.13** | **0.00** | **0.090** | **17,698.4** | **0.00** | **0.166** | **33,612.5** | **0.00** | **0.165** |
| Congruent*n | 2 | 0.92 | 0.40 | 0.000 | **697.68** | **0.00** | **0.008** | **12,722.4** | **0.00** | **0.119** | **34,340.3** | **0.00** | **0.168** |
| n*Z | 2 | **65.17** | **0.00** | **0.001** | **84.97** | **0.00** | **0.001** | **1,370.87** | **0.00** | **0.013** | **35.35** | **0.00** | **0.000** |
| Congruent*n*Z | 2 | 0.04 | 0.96 | 0.000 | 0.03 | 0.97 | 0.000 | **856.36** | **0.00** | **0.008** | **48.1** | **0.00** | **0.000** |
| n*Equal | 2 | **99.42** | **0.00** | **0.002** | **59.49** | **0.00** | **0.001** | **109.70** | **0.00** | **0.001** | **855.08** | **0.00** | **0.004** |
| Congruent*n*Equal | 2 | 2.82 | 0.06 | 0.000 | 2.13 | 0.12 | 0.000 | **113.38** | **0.00** | **0.001** | **861.76** | **0.00** | **0.004** |
| n*Z*Equal | 2 | 0.50 | 0.60 | 0.000 | 0.68 | 0.51 | 0.000 | **18.83** | **0.00** | **0.000** | 0.65 | 0.52 | 0.000 |
| Congruent*N*Z*Equal | 2 | 0.08 | 0.92 | 0.000 | 0.07 | 0.93 | 0.000 | **17.41** | **0.00** | **0.000** | 2.18 | 0.11 | 0.000 |
| Phi | 2 | **1,084.13** | **0.00** | **0.017** | **1,819.88** | **0.00** | **0.021** | **515.41** | **0.00** | **0.005** | **1,910.90** | **0.00** | **0.009** |
| Congruent*Phi | 2 | **134.51** | **0.00** | **0.002** | **261.74** | **0.00** | **0.003** | **657.18** | **0.00** | **0.006** | **1,954.11** | **0.00** | **0.010** |
| Z*Phi | 2 | 4.37 | 0.01 | 0.000 | **18.96** | **0.00** | **0.000** | **19.42** | **0.00** | **0.000** | **12.44** | **0.00** | **0.000** |
| Congruent*Z*Phi | 2 | 4.64 | 0.01 | 0.000 | **10.04** | **0.00** | **0.000** | **28.24** | **0.00** | **0.000** | **14.60** | **0.00** | **0.000** |
| Equal*Phi | 2 | **537.56** | **0.00** | **0.008** | **1,019.72** | **0.00** | **0.012** | **18.57** | **0.00** | **0.000** | **143.50** | **0.00** | **0.001** |
| Congruent*Equal*Phi | 2 | 0.71 | 0.49 | 0.000 | **67.51** | **0.00** | **0.001** | **14.40** | **0.00** | **0.000** | **139.41** | **0.00** | **0.001** |
| Z*Equal*Phi | 2 | 0.89 | 0.41 | 0.000 | 1.65 | 0.19 | 0.000 | 1.56 | 0.21 | 0.000 | 0.48 | 0.62 | 0.000 |
| Congruent*Z*Equal*Phi | 2 | 1.16 | 0.31 | 0.000 | 2.52 | 0.08 | 0.000 | **7.20** | **0.00** | **0.000** | 1.89 | 0.15 | 0.000 |
| n*Phi | 4 | **201.43** | **0.00** | **0.006** | **220.18** | **0.00** | **0.001** | **94.45** | **0.00** | **0.002** | **395.32** | **0.00** | **0.004** |
| Congruent*n*Phi | 4 | **3.75** | **0.00** | **0.000** | **39.52** | **0.00** | **0.001** | **104.21** | **0.00** | **0.002** | **409.06** | **0.00** | **0.004** |
| n*Z*Phi | 4 | 0.95 | 0.43 | 0.000 | 1.67 | 0.15 | 0.000 | **6.76** | **0.00** | **0.000** | 3.55 | 0.01 | 0.000 |
| Congruent*n*Z*Phi | 4 | 2.49 | 0.04 | 0.000 | 1.08 | 0.36 | 0.000 | **5.19** | **0.00** | **0.000** | 2.27 | 0.06 | 0.000 |
| n*Equal*Phi | 4 | **4.20** | **0.00** | **0.000** | **26.82** | **0.00** | **0.001** | 2.40 | 0.05 | 0.000 | **27.87** | **0.00** | **0.000** |
| Congruent*N*Equal*Phi | 4 | 1.14 | 0.33 | 0.000 | **4.36** | **0.00** | **0.000** | **3.90** | **0.00** | **0.000** | **32.64** | **0.00** | **0.000** |
| n*Z*Equal*Phi | 4 | 0.44 | 0.78 | 0.000 | 0.24 | 0.92 | 0.000 | 0.74 | 0.56 | 0.000 | 0.22 | 0.93 | 0.000 |
| Congruent*n*Z*Equal*Phi | 4 | 0.60 | 0.67 | 0.000 | 0.33 | 0.85 | 0.000 | 0.97 | 0.42 | 0.000 | 0.43 | 0.78 | 0.000 |

**Notes:** Significant effects are shown in bold type. Model R-squares are .45, .58, .66, and .82 for $r_c$, theta (θ), WALD, and DIFF respectively.

Key findings from Table 1 are straightforward. With the effect sizes, the factors account for more variance in the model for θ than for $r_c$, R-square = .58 versus .45 respectively. Across the model conditions, the congruence manipulation has by far the strongest effect: $\omega^2 = .42$ for θ and .31 for $r_c$. The next highest effects are $\omega^2 = .09$ for the sample size across both $r_c$ and θ. The remaining effect sizes are .02 or less for all main effects and interactions. Therefore, the congruence manipulation works as intended, and the sample size influences $r_c$ and θ values in ways not shown by the random correlations summarized in Fig. 1.

The results show that the model terms account for more variance in DIFF outcomes than for WALD tests (R-square = .82 versus .66). The congruence manipulation again has the largest effects, both overall and in combination with the sample size. Congruence, sample size, and congruence x sample size $\omega^2$ values are respectively .42, .16, and .17 for DIFF and .24, .17, and .12 for WALD. The number of Z-variables I has an $\omega^2$ value .08 for WALD versus .01 for DIFF. The other main effects and interactions are all minor, $\omega^2 \leq .014$.

### *4.2. Data distributions and box plots*

To see the data distributions underlying the GLM results in Table 1, the graphs in Figs. 3-6 show box plots for almost all combinations of design factors. The exception is the equal/unequal reliability manipulation, which explains almost no variance in the outcome variables (maximum $\omega^2 = .012$) and is therefore omitted for graphical parsimony. The most obvious patterns in the figures are from the congruence manipulation, with congruent results in the upper graph and incongruent results in the lower. Within each of these sets, the first nine box plots are for I = 3 and the second nine represent I = 5. Further subdivisions include the three levels of sample sizes, and finally across each value of n are the levels of ϕ. The axis labels reflect this ordering, with the first letter representing Congruence vs. Incongruence, then 3 or 5 Z-variables, then sample

sizes of 100, 250, and 500 (shown as 1, 2, and 5), and finally ϕ of .71 (shown as .7), .8, and .9.

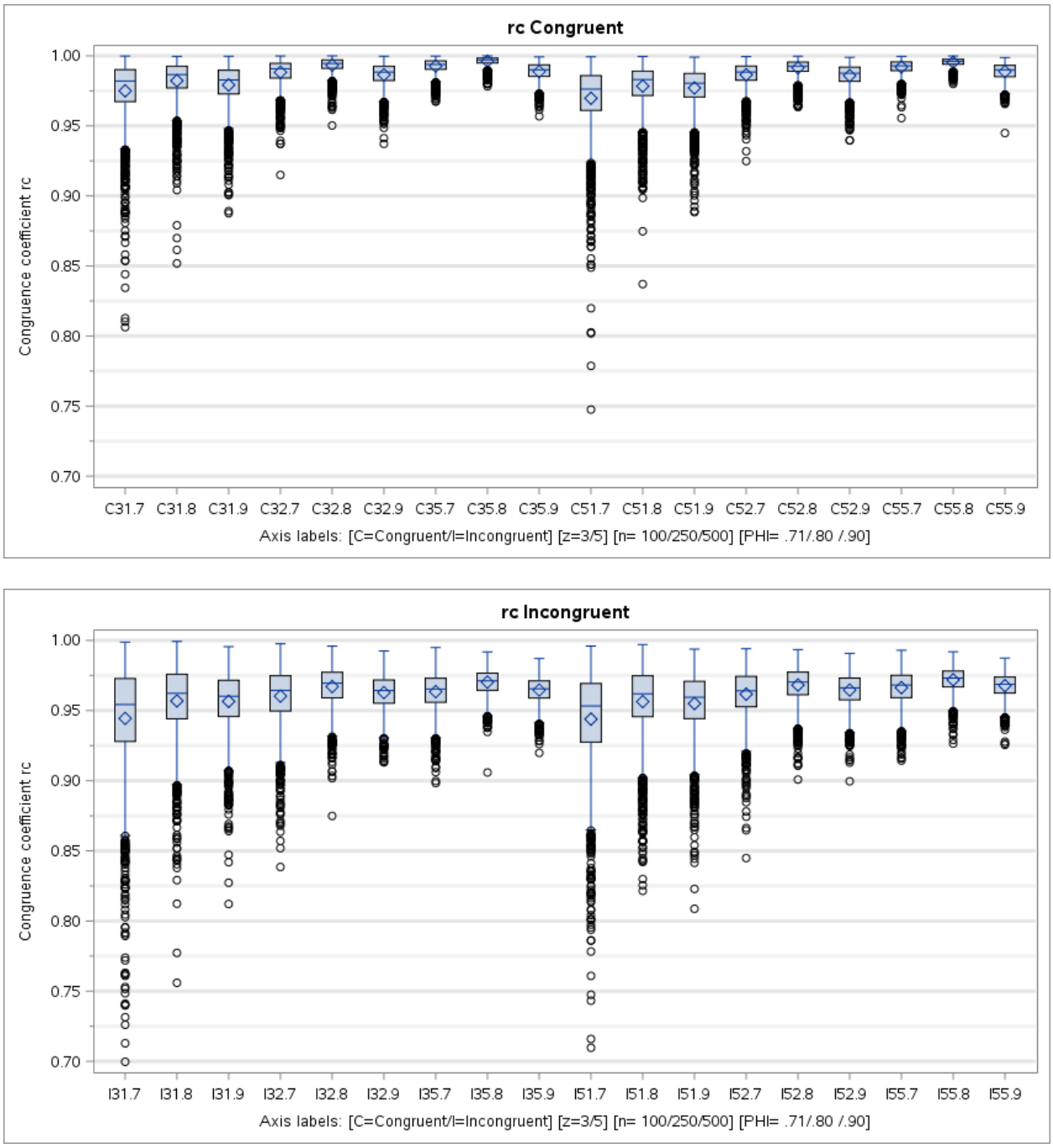


**Fig. 3**. $r_c$ simulation results

**Notes:** $r_c$ distributions for the congruent simulations are in the upper panel; incongruent simulations are in the lower panel. In both panels, k = 5 (three Z-variables) results are on the left, k = 7 (five Z-variables) results are on the right. Results are combined across levels of the Equal/Unequal reliability manipulation.

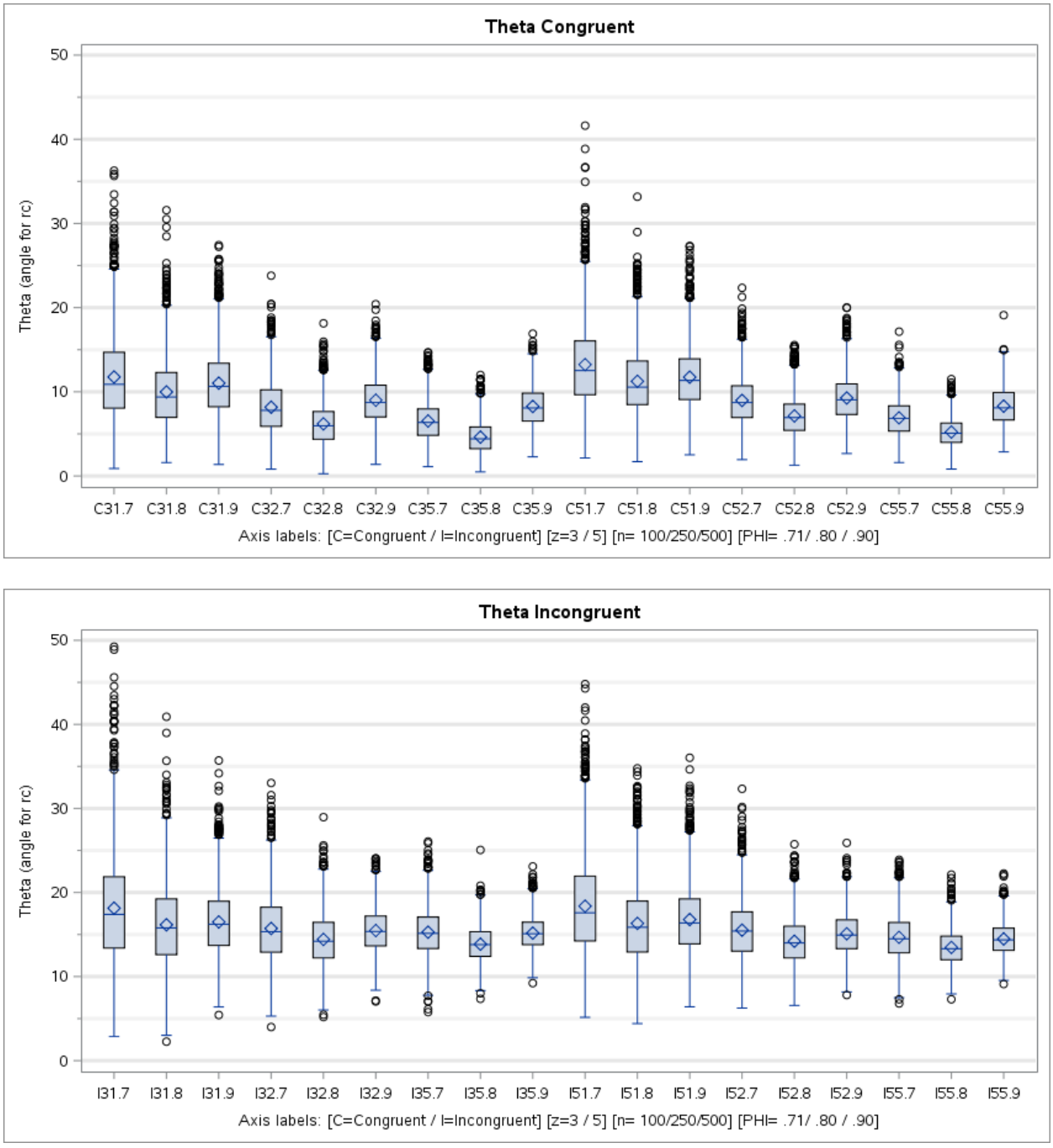


**Fig. 4**. Theta simulation results

**Notes:** θ distributions for the congruent simulations are in the upper panel; incongruent simulations are in the lower panel. In both panels, k = 5 (three Z-variables) results are on the left, k = 7 (five Z-variables) results are on the right. Results are combined across levels of the Equal/Unequal reliability manipulation.

In Figs. 3 and 4, in every case, the averages are higher for $r_c$ and lower for $\theta$ when X and Y are congruent rather than incongruent. The differences are not large, but as discussed above, neither is the incongruence manipulation. The results are less variable for congruent samples, with smaller interquartile ranges and fewer outliers than in incongruent samples. Variability also decreases as sample sizes increase. The figures give a hint of $\phi_{XY}$ effects, with slightly greater congruence (higher $r_c$, lower $\theta$) for $\phi_{XY} = .8$ than for .71 and .9. However, the distributions overlap for all three levels, consistent with the minor main effects and interactions shown in Table 1.

If the WALD and DIFF tests follow a chi-square distribution, as expected, their average values should equal their degrees of freedom when the null hypothesis is true (i.e., in the congruent condition). They should also have the correct size, meaning they reject the true null in the same percent of random samples as the chosen alpha level (e.g., five percent when $\alpha = 0.05$). When the null is false (the incongruent simulations), the tests should have good power to produce significant results relative to alpha. Low power leads to conservative inferences, and the wrong size can be either too conservative or liberal. For example, the Fornell-Larcker (1981) test is normally applied as a directional comparison of shared variance versus AVE, giving it a size of .50 in random samples when its null hypothesis is actually true (Franke & Sarstedt, 2019).

By these standards, the WALD test underperforms. In the congruent condition (upper panel, Fig. 5), when the null hypothesis is true, the average test value is always lower than the expected value shown by the straight horizontal lines at 2 and 4 (left axis). The average percent of significant results is lower than the expected $p = .05$ (right axis). Therefore, the size of the WALD test is lower than the .05 alpha level, meaning it is unnecessarily conservative.

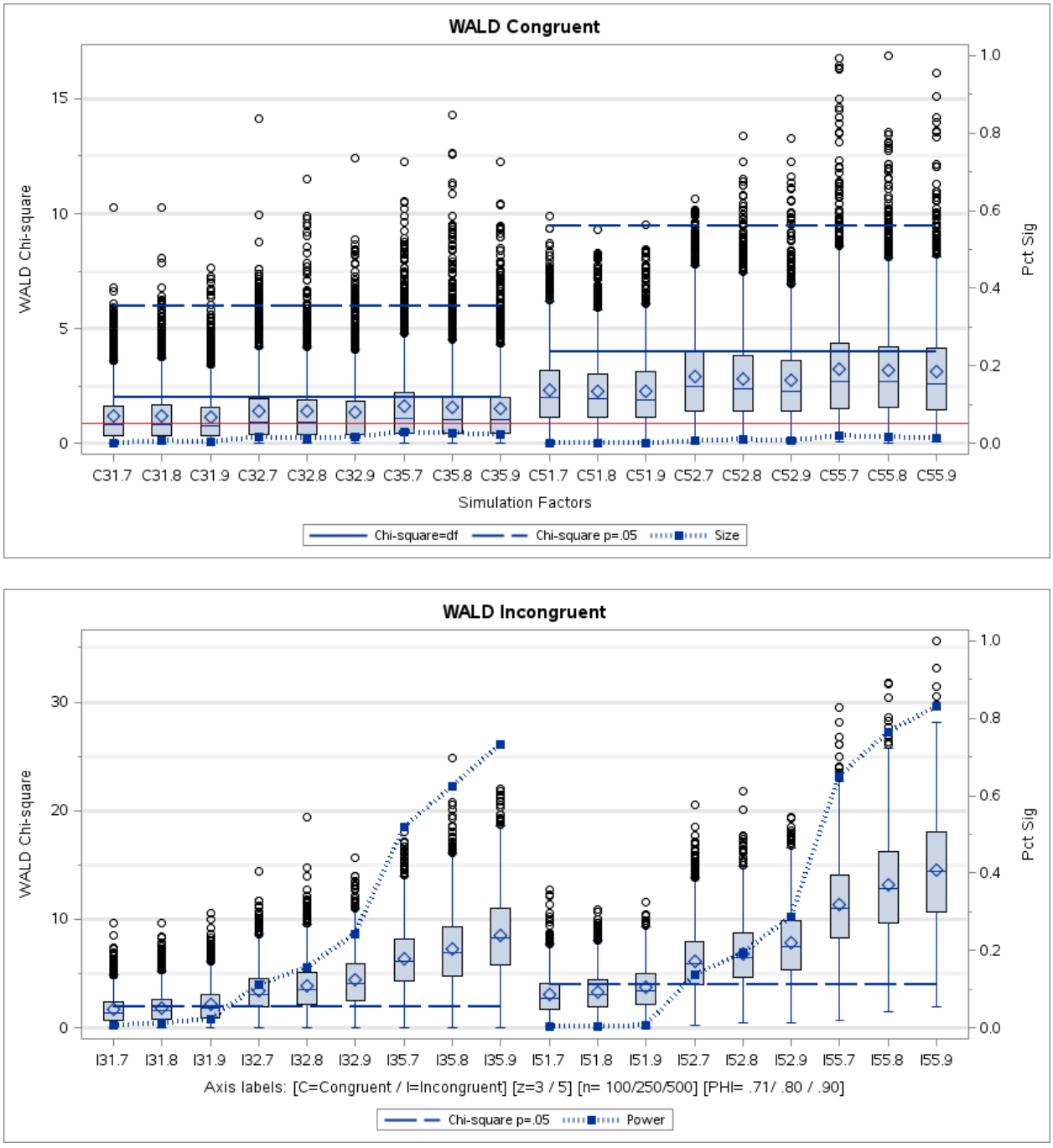


**Fig. 5**. WALD chi-square distributions for the congruent simulations are in the upper panel; incongruent simulations are in the lower panel. In both panels, k = 5 (three Z-variables) results are on the left, k = 7 (five Z-variables) results are on the right. The solid lines (left axis, upper panel) at 2 and 4 are the expected chi-squares under the null hypothesis. The dashed lines are the critical chi-square values with 2 and 4 d.f. The red solid line at p = .05 (upper panel, right axis) is the expected size of the test. Dotted lines (both panels, right axis) are the percent of significant findings at p = .05.

Power rather than size is relevant when the null hypothesis is false, as in the incongruent condition (lower panel, Fig. 5). The dotted lines with the filled square markers show the percent of null hypotheses that are correctly rejected (i.e., the power of the test). When n = 100, power is near 0. When n = 250, power is less than .30. Even when n = 500, the only simulation condition with power above the conventional .80 standard combines five $Z_i$ variables with a very high $\phi_{XY}$ = .9. The sample size has a much stronger effect when the null hypothesis is false than when it is true, so that the two panels in the figure illustrates the congruence x n interaction found in the GLM results (Table 1).

The DIFF test is more effective, as shown in Fig. 6. In the congruent condition (upper panel), the average chi-squares are consistently very near the expected values of 2 and 4, and the proportion of significant results is close to the expected p = .05. Therefore, DIFF has the appropriate size for rejecting the true null hypothesis. Power is substantially higher than in the WALD test for detecting measure incongruence (lower panel). When n = 100, power ranges from roughly .45 to .60 depending on the value of $\phi_{XY}$. Power is consistently above .90 when n = 250, and approaches 1.0 when n = 500. In Table 1, $\omega^2$ for the congruence x n interaction is .17 for DIFF versus .12 for WALD, as reflected in the disparate power findings in Figs. 5 and 6.

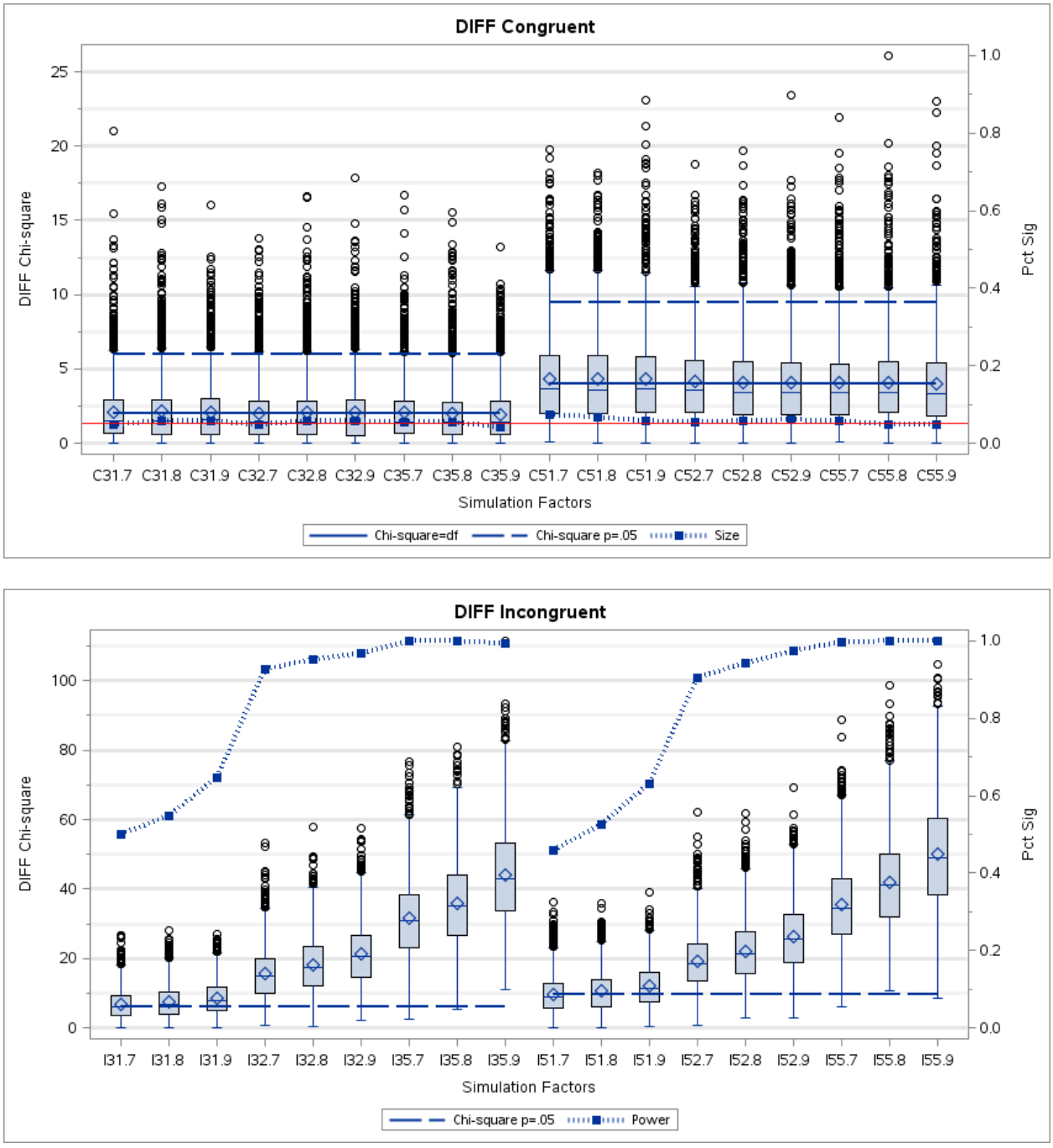


**Fig. 6**. DIFF test simulation results

**Notes:** DIFF chi-square distributions for the congruent simulations are in the upper panel; incongruent simulations are in the lower panel. In both panels, k = 5 (three Z-variables) results are on the left, k = 7 (five Z-variables) results are on the right. The solid lines (left axis, upper panel) at 2 and 4 are the expected chi-squares under the null hypothesis. The dashed lines are the critical chi-square values with 2 and 4 d.f. The red solid line at p = .05 (upper panel, right axis) is the expected size of the test. Dotted lines (both panels, right axis) are the percent of significant findings at p = .05.

Fig. 1 gives a baseline for interpreting $r_c$ and θ values for random correlations, and Figs. 3-6 show results in the context of a particular simulation design. In empirical applications measuring variables within nomological networks, correlations between constructs are expected to show more meaningful patterns as predicted by theory. To illustrate congruence assessment in practice, the next section presents three examples with real data: the first with observed data for multiple indicators of latent variables, the second with published correlations between measures of organizational market information-processing practices, and the third with published MTMM correlations from an analysis of personality dimensions assessed using multiple measures.

## 5. Empirical examples

We examine measure congruence in three examples. In the first two, we also apply three tests of discriminant validity: the Anderson-Gerbing criterion, the Fornell-Larcker criterion, and McDonald's (1999) test. The last test is not widely used, but it has the desirable feature of directly taking measure reliability into account (Franke & Sarstedt, 2019). AVE as considered in the Fornell-Larcker test can be calculated in terms of average item reliabilities, but unlike CR, the number of indicators is irrelevant to AVE. McDonald (1999, p. 212) does not propose an explicit test criterion, but gives an example that suggests a workable test incorporating the reliabilities of both constructs (Franke & Sarstedt, 2019):

$$CR_1^{.5} - CR_2^{.5} * \Phi_{21} > 0 \text{ and } CR_2^{.5} - CR_1^{.5} * \Phi_{21} > 0 \quad (4)$$

for constructs 1 and 2. That is, indicators should have a stronger relationship with their intended construct than with another construct. Paraphrasing McDonald's description, $CR_1^{.5} * \phi_{21} = .77$ versus $CR_2^{.5} = .78$ means that the indicators of construct 1 are almost as good a measure of construct 2 as are 2's own indicators.

### *5.1. Example 1: Data from Bove et al. (2009)*

Bove et al. (2009) present a model of how four measures of customer appraisals of service workers influence the customers' organizational citizenship behaviors (OCBs), such as positive word-of-mouth and participation in the organization's activities. Using data from 484 respondents, they apply the Fornell-Larcker (1981) criterion in terms of the *average* AVE between pairs of constructs, rather than for each construct individually. Farrell (2010) identifies three correlations between constructs that fail the actual Fornell-Larcker (1981) criterion. Analyzing data provided by one of Bove's coauthors (with OCB parcels rather than all 29 original items), we derive the results shown in Table 2.

**Table 2.** Empirical Example 1: Bove et al. (2009)

| Construct 1 | Construct 2 | $r_c$ | Theta (θ) | WALD (p<) (df = 2) | DIFF (p<) (df = 2) | PHI | FL 1 (p <) | FL 2 (p <) | McDonald 1 (p <) | McDonald 2 (p <) |
|---|---|---|---|---|---|---|---|---|---|---|
| Commitment | Credibility | 0.929 | 21.8 | 34.2 (.00) | 50.8 (.00) | 0.51 | .61 (.00) | .40 (.00) | .48 (.00) | .45 (.00) |
| Commitment | Benevolence | 0.964 | 15.3 | 48.9 (.00) | 53.9 (.00) | 0.64 | .47 (.00) | .22 (.00) | .38 (.00) | .30 (.00) |
| Commitment | Loyalty | 0.988 | 8.8 | 25.0 (.00) | 29.8 (.00) | 0.82 | .20 (.00) | -.03 (.42) | .21 (.00) | .12 (.00) |
| Commitment | OCB | 0.995 | 5.9 | 7.6 (.02) | 7.5 (.02) | 0.78 | .27 (.00) | -.09 (.06) | .29 (.00) | .12 (.00) |
| Credibility | Benevolence | 0.991 | 7.5 | 5.9 (.05) | 6.0 (.05) | 0.77 | .08 (.06) | .04 (.35) | .24 (.00) | .19 (.00) |
| Credibility | Loyalty | 0.973 | 13.3 | 41.1 (.00) | 55.6 (.00) | 0.72 | .15 (.00) | .13 (.00) | .28 (.00) | .24 (.00) |
| Credibility | OCB | 0.958 | 16.6 | 40.6 (.00) | 54.4 (.00) | 0.57 | .34 (.00) | .18 (.00) | .44 (.00) | .33 (.00) |
| Benevolence | Loyalty | 0.993 | 6.7 | 24.2 (.00) | 30.0 (.00) | 0.79 | -.00 (.95) | .02 (.74) | .18 (.00) | .19 (.00) |
| Benevolence | OCB | 0.986 | 9.7 | 9.1 (.00) | 46.8 (.00) | 0.69 | .15 (.00) | .03 (.57) | .31 (.00) | .24 (.00) |
| Loyalty | OCB | 0.998 | 3.5 | 5.5 (.06) | 6.7 (.05) | 0.80 | .01 (.90) | -.13 (.02) | .22 (.00) | .13 (.00) |

**Notes:** All PHI-values < 1, $p < .001$. FL 1 and FL 2 indicate the Fornell-Larcker criterion value when using the average variance extracted of Constructs 1 and 2, respectively. McDonald 1 and 2 indicate the criterion value applied to Constructs 1 and 2, respectively. Some results do not match Bove et al. (2009) and Farrell (2010) due to published rounding errors and the use of OCB parcels in the available data.

The results show that the $\phi$ values are generally large, ranging from .51 to .82, but are consistently significantly less than 1.0. The Fornell-Larcker values show the difference between $\phi^2$ and the AVE for each of the paired variables, corresponding to Constructs 1 and 2 (FL 1 and 2 in Table 2, respectively). Positive values imply discriminant validity according to the Fornell-Larcker criterion. Nonsignificant p-values suggest that confidence intervals are ambiguous about whether AVEs are larger or smaller than shared variances. For McDonald's (1999) test, using Construct 1 for $CR_1$ and Construct 2 for $CR_2$ in Eq. 4 (McDonald 1 and 2 in Table 2, respectively) shows that the measures are better indicators of their intended construct than the other construct in each pair.

Given the nature of the constructs, involving related worker perceptions and voluntary citizenship behaviors, the results in Table 2 unsurprisingly suggest congruence between variables. The least congruent pair involves customer commitment to the worker and perceptions of worker credibility, with $r_c$ =.929 and $\theta$ = 21.8. For random correlations with five variables, these values are close to the $90^{th}$ and $5^{th}$ percentiles respectively. All other pairs show even stronger congruence, with the greatest similarity between loyalty to the service worker and OCBs ($r_c$ =.998 and $\theta$ = 3.5). However, even with these extreme values, the DIFF tests reject congruence for all ten pairs of variables (p-values ranging from .000 to .049). The less powerful WALD test does not reject congruence for two pairs at a strict .05 level: for loyalty-OCBs, $p < .063$, and for credibility-beneficence, $p < .052$.

To summarize, while the pairwise tests give mixed evidence for discriminant validity, congruence assessment raises concerns about the distinctiveness of OCBs relative to three other variables: commitment, benevolence, and loyalty. Though the DIFF test rejects congruence for all three pairs at $p < .05$, the magnitudes of $r_c$ and $\theta$ suggests that each pair has similar

(proportionally equal) correlations. Loyalty also shows high congruence with commitment and benevolence, as would be expected in response to good or bad customer service. Overall, congruence assessment supports Bove et al.'s (2009) findings of an important relationship between service worker perceptions and customer OCBs. However, the results raise questions about whether the measured OCBs are better construed as behavioral outcomes of worker perceptions, or as affective indicators of attitudes toward the service provider.

### *5.2. Example 2: Data from Hult et al. (2005)*

Hult et al. (2005) test a model in which three cultural aspects and three information-processing aspects of market orientation influence organizational responsiveness to customers and competitors, which in turn influences performance. They also take six control variables into account. Except for firm size, age, and performance, all constructs are measured with multiple items. They note that for every pair of latent variables, a test of $\phi < 1$ is significant. However, AVE $< \phi^2$ for two pairs of information-processing variables, suggesting a lack of discriminant validity between information dissemination and generation, and between dissemination and shared interpretation. Note that this does not compromise their study's results, but might with other analysis approaches or nomological networks.

Hult et al. (2005) report correlations, AVEs, and shared variance ($\phi^2$) values for 217 firms. Table 3 presents results as available from analyzing the three pairs of information-processing variables. Congruence effects indicate that the variables have nearly proportional correlations with the other variables analyzed: $r_c \geq .986$ and $\theta \leq 9.6$. Nevertheless, the DIFF test significantly ($p < .036$) rejects congruence for all three pairs. The WALD test fails to reject congruence for the same pairs as indicated by the Fornell-Larcker criterion: information generation and dissemination ($p < .057$) and information dissemination and shared interpretation ($p < .261$).

McDonald's (1999) test directionally support discrimination for all three variable pairs. Note that the available information does not allow for assessing the Fornell-Larcker and McDonald's (1999) test for significance.

**Table 3.** Empirical Example 2: Hult et al. (2009)

| Variable 1 | Variable 2 | $r_c$ | Theta (θ) | WALD (p<) (df = 2) | DIFF (p<) (df = 2) | PHI | FL 1 (p<) | FL 2 (p<) | McDonald 1 (p<) | McDonald 2 (p<) |
|---|---|---|---|---|---|---|---|---|---|---|
| Information Generation | Information Dissemination | 0.990 | 8.0 | 21.9 (.06) | 40.3 (.00) | 0.72 | .11 (na) | -.08 (na) | .25 (na) | .23 (na) |
| Information Generation | Shared Interpretation | 0.986 | 9.6 | 26.0 (.02) | 58.1 (.00) | 0.68 | .17 (na) | .39 (na) | .28 (na) | .31 (na) |
| Information Dissemination | Shared Interpretation | 0.990 | 6.6 | 15.8 (.26) | 23.5 (.04) | 0.78 | -.17 (na) | .24 (na) | .18 (na) | .21 (na) |

**Notes:** All PHI-values < 1, $p < .001$, per Hult et al. (2009). na is not available. FL 1 and FL 2 indicate the Fornell-Larcker criterion value when using the average variance extracted of Constructs 1 and 2, respectively. McDonald 1 and 2 indicate the criterion value applied to Constructs 1 and 2, respectively.

In sum, the latent-variable correlations, McDonald's (1999) tests, and DIFF test results imply distinctions between the information-processing variables. However, the congruence effect sizes suggest that their correlations with the other variables in the data are generally very consistent (proportionally equal). Rather than raising concerns about the measures, this pattern supports Hult et al.'s (2005) treatment of market information processing as a single summated scale, rather than as three distinct dimensions.

### *5.3. Example 3: Data from Hong et al. (2008)*

The "Big Five" personality traits of openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism are a research focus in many settings (e.g., Park et al., 2020). Hong et al. (2008) report an MTMM investigation (n = 295) in which the methods are three different measures of the Big Five: the NEO Five-Factor Inventory (Costa & McCrae, 1992), the Five-Factor Nonverbal Personality Questionnaire (Paunonen et al., 2004), and an adjective rating form called B5-ADJ (Goldberg, 1992). For comparison with other MTMM analyses, and because of the large number of variable pairs, Table 4 shows effect sizes $r_c$ and $\theta$ without pairwise tests of significance or discriminant validity.

**Table 4.** Empirical Example 3: Hong, Paunonen and Slade (2008)

| | ***NEO-FFI*** | | | | | ***FF-NPQ*** | | | | | ***B5-ADJ*** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***NEO-FFI*** | **N** | **E** | **O** | **A** | **C** | **N** | **E** | **O** | **A** | **C** | **N** | **E** | **O** | **A** | **C** |
| Neuroticism | 0.86 | *37.75* | *55.97* | *47.87* | *45.45* | **12.47** | 32.77 | 55.27 | 75.85 | 53.77 | **30.59** | 60.39 | 78.77 | 65.44 | 45.18 |
| Extraversion | *0.79* | 0.78 | *61.62* | *36.13* | *58.40* | 40.80 | **24.38** | 48.38 | 82.15 | 68.12 | 42.49 | **41.46** | 70.42 | 55.81 | 60.11 |
| Openness | *-0.56* | *-0.48* | 0.67 | *82.07* | *85.70* | 61.89 | 63.42 | **33.31** | 84.71 | 88.32 | 57.74 | 65.07 | **62.53** | 78.51 | 64.86 |
| Agreeableness | *0.67* | *0.81* | *-0.14* | 0.76 | *45.12* | 44.54 | 32.42 | 73.66 | **60.84** | 49.00 | 52.19 | 65.68 | 76.30 | **60.64** | 63.42 |
| Conscientiousness | *0.70* | *0.52* | *-0.08* | *0.71* | 0.82 | 38.82 | 47.64 | 84.13 | 67.21 | **31.14** | 58.75 | 81.25 | 85.09 | 80.50 | **54.55** |
| ***FF-NPQ*** | | | | | | | | | | | | | | | |
| Neuroticism | **0.98** | 0.76 | -0.47 | 0.71 | 0.78 | 0.80 | *31.94* | *63.02* | *69.31* | *45.34* | **31.27** | 65.73 | 80.27 | 67.88 | 47.61 |
| Extraversion | 0.84 | **0.91** | -0.45 | 0.84 | 0.67 | *0.85* | 0.82 | *58.44* | *64.28* | *53.31* | 36.21 | **54.37** | 73.22 | 61.84 | 57.96 |
| Openness | -0.57 | -0.66 | **0.84** | -0.28 | -0.10 | *-0.45* | *-0.52* | 0.80 | *72.07* | *82.96* | 60.78 | 52.09 | **51.52** | 78.91 | 75.09 |
| Agreeableness | 0.24 | 0.14 | 0.09 | **0.49** | 0.39 | *0.35* | *0.43* | *0.31* | 0.76 | *49.22* | 70.76 | 83.93 | 80.77 | **72.53** | 73.97 |
| Conscientiousness | 0.59 | 0.37 | -0.03 | 0.66 | **0.86** | *0.70* | *0.60* | *0.12* | *0.65* | 0.79 | 56.61 | 85.85 | 84.51 | 78.37 | **49.41** |
| ***B5-ADJ*** | | | | | | | | | | | | | | | |
| Neuroticism | **0.86** | 0.74 | -0.53 | 0.61 | 0.52 | **0.85** | 0.81 | -0.49 | 0.33 | 0.55 | 0.84 | *57.52* | *79.19* | *63.61* | *53.83* |
| Extraversion | 0.49 | **0.75** | -0.42 | 0.41 | 0.15 | 0.41 | **0.58** | -0.61 | -0.11 | 0.07 | *0.54* | 0.87 | *72.88* | *57.56* | *67.60* |
| Openness | -0.19 | -0.34 | **0.46** | -0.24 | -0.09 | -0.17 | -0.29 | **0.62** | 0.16 | 0.10 | *-0.19* | *-0.29* | 0.74 | *83.68* | *81.48* |
| Agreeableness | 0.42 | 0.56 | -0.20 | **0.49** | 0.17 | 0.38 | 0.47 | -0.19 | **0.30** | 0.20 | *0.44* | *0.54* | *0.11* | 0.85 | *56.08* |
| Conscientiousness | 0.70 | 0.50 | -0.42 | 0.45 | **0.58** | 0.67 | 0.53 | -0.26 | 0.28 | **0.65** | *0.59* | *0.38* | *0.15* | *0.56* | 0.81 |

**Notes:** $r_c$ is below the diagonal, reliabilities (from Hong et al. 2008) are on the diagonal, and theta (θ) is above the diagonal. Monotrait-heteromethod (validity) correlations are in bold. Heterotrait-monomethod correlations are in italics.

Following Hong et al. (2008), the validity diagonals (same traits, different methods) are shown in bold in Table 4. Heterotrait-monomethod correlations are in italics. According to the seminal Campbell and Fiske (1959) criteria, values in the validity diagonals should be large enough to merit further consideration of convergent validity. At the other extreme (different variables and different methods), for each variable, its validity-diagonal value should be larger than the heterotrait-heteromethod correlations (plain text in Table 4). An intermediate criterion is that validity-diagonal correlations should be larger than heterotrait-monomethod correlations; that is, the correlations should reflect common-trait effects more strongly than common-method effects. Finally, heterotrait triangles should show consistent patterns regardless of method. Given the correlations reproduced in Table 4, Hong et al. (2008, p. 162) see "evidence of generally good convergent and discriminant validity according to the Campbell and Fiske (1959) criteria." However, alternative CFA analysis procedures "indicated that discriminant validity of the Big Five factors is at best moderate" (p. 163).

For comparison with the values in Table 4, extending Fig. 1 to include k = 15 (for three methods and five traits) gives 90$^{th}$, 95$^{th}$, and 99$^{th}$ percentiles of $r_c$ = .629, .704, and .754 respectively. For θ, the 10$^{th}$, 5$^{th}$, and 1$^{st}$ percentiles (angles closer to 0$^{o}$) are 51.0, 45.2, and 35.1 degrees respectively. Therefore, several entries in the validity diagonals do not reflect congruence levels much better than chance. Most but not all of the validity-diagonal values are larger than the heterotrait-monomethod congruences. Visual examination suggests multiple discrepancies between patterns of correlations in the heterotrait triangles. Overall, the results in Table 4 suggest that measures of neuroticism and extraversion are reasonably congruent across methods, though less so for the FF-NPQ and B5-ADJ measures. Openness and conscientiousness are relatively congruent for the NEO-FFI and FF-NPQ measures. Agreeableness correlations are

not especially congruent between any pairs of measures.

In short, congruence assessment, which considers all measures simultaneously without imposing an a priori measurement model, does not strongly support Hong, Paunonen and Slade's (2008) conclusions based on the Campbell and Fiske (1959) criteria or CFA results. Except for neuroticism, the B5-ADJ dimensions are not particularly congruent with each other or across measures. Neuroticism and extraversion are substantially more congruent than random correlations, but the $r_c$ and $\theta$ values clearly allow for inconsistent correlations with other variables. Thus, the congruence results indicate that the Big Five dimensions are empirically distinct, as expected, but raise questions about exactly what aspects of personality are captured by the alternative measurement scales.

## 6. Discussion

### *6.1. Theoretical implications of congruence*

Campbell and Fiske's (1959) introduction of convergent and discriminant validity has influenced measurement methodology for more than half a century. Convergent validity seems straightforward: Do measures of items and constructs correlate significantly as expected, and to what extent? Discriminant validity has received considerably more research attention in recent years (including among others Franke & Sarstedt, 2019; Henseler et al., 2015; Matthes & Ball, 2018; Rönkkö & Cho, 2021; Schaffer et al., 2016; Voorhees et al., 2016). As stated by Campbell and Fiske (1959, p. 84), "One cannot define without implying distinctions, and the verification of these distinctions is an important part of the validational process." Or put more succinctly, "knowledge is knowledge of differences" (Fiske, 1982, p. 89). However, when measures of the same construct have different labels, theory development and testing, literature searches, survey length, and empirical analyses can all be compromised.

A construct has meaning only in the context of a nomological network: "Constructs are of interest only if they are connected to other constructs… Failure of the measure to relate to measures of other constructs (assuming the latter are accepted as valid) leads to the modification of the measure (and hence the construct), modification of the theory connecting the construct to the other constructs, or the abandonment of both" (Schwab, 1980, pp. 6-7). Congruence assessment considers similarities and differences between pairs of variables relative to a network of other variables, not in isolation. High congruence, like high correlations, "challenge[s] the proponent of distinct constructs to locate or create the circumstances under which the variables in question are distinguishable" (Messick, 1989, p. 51). An example can be found in Bove et al. (2009): Though the commitment and loyalty measures (directionally) fail the Fornell-Larcker criterion, their respective correlations of .51 and .72 with worker credibility differ significantly ($p < .01$). Whether this difference is substantively important within the nomological network is a different question than a yes-no conclusion about discriminant validity based on shared variance and AVEs.

Even when "the true correlation between two traits is *meaningfully* less than unity" (Werts et al., 1974, p. 281, italics added), high congruence calls for further examination of the measures and data. Consistent correlations could result from common method variance, which can be addressed in data collection and analysis (Podsakoff et al., 2012). Alternatively, and more important theoretically, the variables in the network could have other common causes. Identifying, measuring, and controlling such influences would enhance understanding of relationships between other variables in the network.

A test or measure is neither valid nor invalid in general: "What is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators"

(Messick, 1989, p. 13). Therefore, conclusions about two variables' congruence, and more broadly about construct validity, logically depend on which other variables are considered within a nomological network. Evidence for or against congruence between the same two variables may meet theoretical expectations in different settings. However, as research in an area evolves, understanding "a theoretical construct is a matter of elaborating the nomological network in which it occurs, or of increasing the definiteness of the components" (Cronbach & Meehl, 1955, p. 290).

Whitely (1983, p. 180) discusses the scope of a network in terms of its nomothetic span, which "indicates the importance of a test as a measure of individual differences." Implication analysis calls for making theories elaborate. As a first step in theory testing, "the implications of a theory are developed for all sorts of conditions and circumstances … and these implications are then empirically examined… A useful theory should suggest a wide array of expansions" (Lieberson & Horwich, 2008, pp. 22-23). In a mature research area, meta-analytic investigations may examine congruence between variables across much broader networks than any single study reports. For example, King and King (1990, p. 59) use results from two previous meta-analyses to conclude, "the meta-analytic findings suggest that there is little difference between the two constructs [i.e., role conflict and role ambiguity], at least concerning patterns of relationships with other variables." Thus, congruence assessment can contribute to theory and measure validation as a research area develops.

### *6.2. Practical implications and future research*

Congruence assessment is well established in the measurement literature, including a discussion and version of Eq. 2 in the enormously influential "Two-Step Approach" paper by Anderson and Gerbing (1988). However, it has played only a very minor role in validation

practices over the years. One contribution of this paper is to show how and why congruence should play an important role in measure validation and theory testing. Another contribution is to show how congruence can be summarized in two related effect size indicators, $r_c$ and $\theta$, and interpreted in terms of statistical significance using contemporary software in ways that were not possible in the past. Therefore, congruence assessment is a feasible tool for modern empirical research.

Congruence of two measures relative to networks of other variables, especially diverse and extensive networks, suggest they are measures of the same thing. Why, then, should these variables not correlate perfectly? Several factors may produce weak equivalence ($\rho \neq 1$ in Eq. 1) between otherwise equivalent measures. Some possible sources for weak equivalence and starting points for identifying congruence issues include:

- *The number of items used in multi-item measures.* As is well known, for a given average correlation between indicators, using more items increases reliability and reduces measurement error (e.g., Nunnally, 1978, p. 211).
- *Random sampling of measurement items.* The domain-sampling model provides a framework for examining measurement error (e.g., Nunnally, 1978, Chapter 6). It posits an infinitely large hypothetical correlation matrix of items that all correlate the same *on average*. By chance, two sets of items from the same domain may correlate more or less than average, increasing or decreasing reliability for a given number of items.
- *Measurement artifacts*, which can influence correlations between observed and true scores (e.g., Schmidt & Hunter, 2014). One is range restriction, which could be reduced in measures or samples that differ in how much of the overall possible range of responses is captured. Others are differing patterns of skewness or variable dichotomization, which can limit the

range of possible correlations (i.e., to greater than -1 or less than +1).

- *Test bias*, which can influence response means, variances, reliabilities, and correlations (e.g., Drasgow, 1982). Though bias is often considered in how the same measure affects different groups, it could also affect how different measures affect a single group.
- *Bifactor models*, in which measures are influenced both by a general underlying trait or "global" factor and another "local" factor that affects a subset of the variables. Variables X and Y could largely share the same global factor, but have distinct local factors (e.g., Raykov & Bluemke, 2020). Depending on whether and how the local factors relate to $Z_i$, the $XZ_i$ and $YZ_i$ correlations could be congruent even though X and Y do not *exclusively* measure the same thing.

These factors could have different implications for conventional tests of discriminant validity. Artifacts that reduce values of $\phi_{XY}$ in testing $\phi_{XY} = 1$ could imply discriminant validity according to Anderson and Gerbing (1988), but reduced shared variance ($\phi_{XY}^2$) relative to AVE could imply a failure of discriminant validity according to Fornell and Larcker (1981). As shown in the simulations, neither of these criteria has direct implications for congruence assessment. That is, $\rho \neq 1$ is problematic for two popular tests of discriminant validity, but is explicitly accommodated in assessing congruence across nomological networks.

In addition to further consideration of these possible influences on proportionality, several areas for future research on congruence seem promising. One is simply to apply the procedures described in this study to additional extant or future examples, to provide further grounds for interpreting congruence effect sizes and tests, and to see whether they offer new insights relative to conventional practices. Congruence is grounded in patterns of correlations, not their magnitudes. Therefore, like MTMM analyses and HTMT, it can be used for descriptive,

atheoretical purposes, such as comparing variable interrelationships over time with early versus late survey respondents or with test-retest and panel data. Congruence assessment across traits and methods in MTMM matrices could also provide new measures of effect sizes and significance relative to more conventional approaches (e.g., Hong et al., 2008).

Hunter (1973, p. 61) makes the point that, "since similarity coefficients are based on more data than are direct correlation coefficients, they will be more stable in the face of sampling error." Therefore, it may be useful to explore ways of visualizing and interpreting congruence coefficients. If they are all positive, or can be recoded as positive because of arbitrary underlying metrics (e.g., item reversals in data collection), $r_c$ and $\theta$ can be treated as distance measures. Steenbergen (2000) suggests that multidimensional scaling can reveal patterns of relationships that would be harder to detect in the coefficients themselves. This approach could be useful in examining MTMM results (as in Table 4). Hunter (1973), Tryon (1958), and others discuss "cluster analysis" of congruence coefficients as a superior alternative to factor analysis in identifying dimensions of individual differences, which could be used in item screening prior to CFA or structural analysis of the nomological network.

An interesting research question is whether and how congruence may be related to multicollinearity in structural models (e.g., Grewal et al., 2004). Strong linear relationships between predictors increase the variances of their structural coefficients, potentially leading to estimates that are nonsignificant, have the wrong sign, or are affected by small changes in sample composition. Congruent predictors are not necessarily highly correlated, but they are redundant in the sense of having the same (proportional) relationships with other variables. Monte Carlo simulations under controlled conditions should clarify when congruent predictors may be effectively treated as distinct predictors, or when they are better treated as indicators of

the same underlying construct (as in Fig. 2, lower panel). Collinearity effects may also be relevant in the context of composite-based SEM, discussed in Appendix 2.

The simulation results summarized in Table 1 and Figs. 3-6 are based on 72 experimental conditions, so the design is already fairly elaborate. However, it might be useful to determine whether analysis of nonnormal data affects the WALD and DIFF tests. Another possible extension is to consider additional patterns of congruence and incongruence. This study focuses on external consistency (relationships between constructs), but internal consistency (relationships within constructs) may provide new insights relative to conventional consideration of exploratory or confirmatory factor analyses. For both applications, the WALD test is convenient, since it requires a single analysis versus the two required for the DIFF test. However, accuracy is more important than convenience in measure validation.

### *6.3. Conclusion*

Measures of different constructs can have high pairwise correlations, and alternative measures of the same construct do not necessarily correlate highly (Cronbach, 1989; Messick, 1989). Our thesis is that constructs can only be validated in the context of nomological networks, rather than by comparing pairs of variables in isolation. Congruence assessment, unlike conventional criteria for convergent and discriminant validity, is fundamental for establishing whether and to what extent measures of focal constructs are consistent or distinct. Therefore, the procedures for congruence assessment presented in this paper can play a key role in validating focal variables relative to their theoretically implied antecedents, consequences, and correlates.

## References


APA Committee on Test Standards (1952). Technical recommendations for psychological tests and diagnostic techniques: Preliminary proposal. *American Psychologist*, *7*(8), 461-476. doi: 10.1037/h0056631.

Anderson, J. C., & Gerbing, D. W. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research*, *19*(4), 453-460. doi: 10.1177/002224378201900407.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411-423. doi: 10.1037/0033-2909.103.3.411.

Bove, L. L., Pervan, S. J., Beatty, S. E., & Shiu, E. (2009). Service worker role in encouraging customer organizational citizenship behaviors. *Journal of Business Research*, *62*(7), 698-705. doi: 10.1016/j.jbusres.2008.07.003.

Brooke, P. P., Russell, D. W., & Price, J. L. (1988). Discriminant validation of measures of job satisfaction, job involvement, and organizational commitment. *Journal of Applied Psychology*, *73*(2), 139-145. doi: 10.1037/0021-9010.73.2.139.

Burt, C. (1937). Methods of factor-analysis with and without successive approximation. *British Journal of Educational Psychology*, *7*(2), 172-195. doi: 10.1111/j.2044-8279.1937.tb03166.x.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105. doi: 10.1037/h0046016.

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods*, *15*(1), 17-32, doi: 0.1177/1094428110392383.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. doi: 10.1007/BF02310555.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.). *Intelligence: Measurement, theory, and public policy* (pp. 147-171) Chicago, IL: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302. doi: 10.1037/h0040957.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391-418. doi: 10.1177/0013164404266386.

Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin*, *92*(2), 526-531. doi: 0033-2909.92.2.526.

Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009).

*Journal of Business Research*, *63*(3), 324-327. doi: 10.1016/j.jbusres.2009.05.003.

Fiske, D. (1982). Convergent-discriminant validation in measurements and research strategies. In D. Brinberg, & L. Kidder (Eds.). *Forms of validity in research* (pp. 77-92). San Francisco, CA: Jossey-Bass.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(February), 39-50. doi: 10.1177/002224378101800104.

Franke, G., & Sarstedt, M. (2019). Heuristics versus statistics in discriminant validity testing: a comparison of four procedures. *Internet Research*, *29*(3), 430-447. doi: 10.1108/IntR-12-2017-0515.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26-42. doi: 10.1037/1040-3590.4.1.26.

Grewal, R., Cote, J. A., & Baumgartner, H. (2004). Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Marketing Science*, *23*(4), 519-529. doi: 10.1287/mksc.1040.0070.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115-135. doi: 10.1007/s11747-014-0403-8.

Hong, R. Y., Paunonen, S. V., & Slade, H. P. (2008). Big Five personality factors and the prediction of behavior: A multitrait–multimethod approach. *Personality and Individual Differences*, *45*(2), 160-166. doi: 10.1016/j.paid.2008.03.015.

Hult, G. T. M., Ketchen Jr, D. J., & Slater, S. F. (2005). Market orientation and performance: an integration of disparate approaches. *Strategic Management Journal*, *26*(12), 1173-1181. doi: 10.1002/smj.494.

Hunter, A. A. (1973). Methods of reordering the correlation matrix to facilitate visual inspection and preliminary cluster analysis. *Journal of Educational Measurement*, *10*(1), 51-61. doi: 10.1111/j.1745-3984.1973.tb00782.x.

Hwang, H., & Y. Takane (2004). Generalized structured component analysis. *Psychometrika*, *69*(1), 81-99. doi: 10.1007/BF02295841.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*(2), 239-251. doi: 10.2307/2334833

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109-133. doi: 10.1007/BF02291393.

King, L. A., & King, D. W. (1990). Role conflict and role ambiguity: A critical assessment of construct validity. *Psychological Bulletin*, *107*(1), 48-64. doi: 10.1037/0033-2909.107.1.48.

Kroger, R.G. (1968). Conceptual and empirical independence in test validation: A note on Campbell and Fiske's "discriminant validity". *Educational and Psychological Measurement*, *28*(2), 383-387. doi: 10.1177/001316446802800217.

Lieberson, S., & Horwich, J. (2008). Implication analysis: a pragmatic proposal for linking theory and data in

the social sciences. *Sociological Methodology*, *38*(1), 1-50. doi: 10.1111/j.1467-9531.2008.00199.x.

Lord, F. M. (1957). A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. *Psychometrika*, *22*(3), 207-220. doi: 10.1007/BF02289122.

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*(2), 57-64. doi: 10.1027/1614-1881.2.2.57.

Matthes, J.M., & Ball, A.D. (2018). Discriminant validity assessment in marketing research. International *Journal of Market Research*, *61*(2), 210-222. doi: 10.1177/1470785318793263.

McDonald, R.P. (1999), *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. Linn (Ed.) *Educational measurement* (pp. 13-103). New York, NJ: American Council on Education.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NJ: McGraw-Hill.

Park, H. H., Wiernik, B. M., Oh, I.-S., Gonzalez-Mulé, E., Ones, D. S., & Lee, Y. (2020). Meta-analytic five-factor model personality intercorrelations: Eeny, meeny, miney, moe, how, which, why, and where to go. *Journal of Applied Psychology*, *105*(12), 1490-1529. doi: 10.1037/apl0000476.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287-312. doi: 10.1207/S15328007SEM0802_7.

Paunonen, S. V., Jackson, D. N., & Ashton, M. C. (2004). *Nonverbal personality questionnaire (NPQ) and five-factor nonverbal personality questionnaire (FF-NPQ)*. Port Huron, MI: Sigma Assessment Systems, Incorporated.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539-569. doi: 10.1146/annurev-psych-120710-100452.

Raykov, T., West, B. T., & Traynor, A. (2015). Evaluation of coefficient alpha for multiple-component measuring instruments in complex sample designs. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 429-438. doi: 10.1080/10705511.2014.936081.

Raykov, T., & Bluemke, M. (2020). Examining multidimensional measuring instruments for proximity to unidimensional structure using latent variable modeling. *Educational and Psychological Measurement*, 0013164420940764. doi: 10.1080/10705511.2014.936081.

Rönkkö, M., & Cho, E. (2021). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, Advance online publication. doi: 10.1177/1094428120968614

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Schwab, D. P. (1980). Construct validity in organizational behavior. *Research in Organizational Behavior*, *2*, 3-43.

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods*, *19*(1), 80-110. doi: 10.1177/1094428115598239.

Steenbergen, M. R. (2000). Item similarity in scale analysis. *Political Analysis*, *8*(3), 261-283. doi: 10.1093/oxfordjournals.pan.a029816.

Tryon, R. C. (1958). Cumulative communality cluster analysis. *Educational and Psychological Measurement, 18*(1), 3-35. doi: 10.1177/001316445801800102.

van der Sluis, S., Dolan, C. V., & Stoel, R. D. (2005). A note on testing perfect correlations in SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(4), 551-577. doi: 10.1207/s15328007sem1204_3.

Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant validity testing in marketing: an analysis, causes for concern, and proposed remedies. *Journal of the Academy of Marketing Science*, *44*(1), 119-134. doi: 10.1007/s11747-015-0455-4.

Werts, C. E., Linn, R. L., & Jöreskog, K. G. (1974). Quantifying unmeasured variables. In H. M. Blalock (Ed.). *Measurement in the social sciences* (pp. 270-292). Chicago, IL: Aldine Publishing, 270-292.

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179-197. doi: 10.1037/0033-2909.93.1.179.

Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction, part II* (pp. 1-54). Amsterdam, Netherlands: North Holland.

Zeller, R. A., Zeller, R. A., & Carmines, E. G. (1980). *Measurement in the social sciences: The link between theory and data*. London: Cambridge University Press.

**Appendix 1: Congruence calculations and significance tests**

For the first two examples, data are from Bove et al. (2009) and may not exactly match Table 2 due to rounding. Composite reliabilities (CR) are in bold.

*Example Spreadsheet Using Microsoft Excel*

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Formula for column: | | | =Bi*Ci | =Bi^2 | =Ci^2 |
| 2 | | X | Y | Numerator | Denominator X | Denominator Y |
| 3 | X | **0.93** | 0.51 | 0.47 | 0.86 | 0.26 |
| 4 | Y | 0.51 | **0.88** | 0.45 | 0.26 | 0.77 |
| 5 | Z1 | 0.64 | 0.77 | 0.49 | 0.40 | 0.59 |
| 6 | Z2 | 0.82 | 0.72 | 0.59 | 0.68 | 0.51 |
| 7 | Z3 | 0.76 | 0.54 | 0.41 | 0.57 | 0.29 |
| 8 | Sum | | | 2.40 | 2.77 | 2.42 |
| 9 | rc | 0.929 | Formula =D9/(E9^0.5*F9^0.5) | | | |
| 10 | theta | 21.8 | Formula =DEGREES(ACOS(ABS(B10))) | | | |

An alternative formula for $r_c$ that uses only columns B and C is
=SUMPRODUCT(B3:B7,C3:C7)/(SUMSQ(B3:B7)^0.5*SUMSQ(C3:C7)^0.5)

*Example Code Using SAS PROC DISTANCE*

Capitalization is optional. Output is not shown. For more discussion and code for matrix algebra calculations, see https://blogs.sas.com/content/iml/2019/09/03/cosine-similarity.html.

```
title ‘rc (“cosine coefficient”) for Bove et al. (2009)’;
data bove; input comm cred bene loy ocb;
cards;
.93 .51 .64 .82 .76
.51 .89 .77 .72 .54
.64 .77 .83 .79 .67
.82 .72 .79 .85 .78
.76 .54 .67 .78 .76
;
proc distance data=bove out=cos method=cosine;
var ratio(_numeric_); run;

proc print data=cos; run; quit;
```

*Example WALD and DIFF Tests Using Mplus*

General setup (not shown, with I = 3). Specify variables, input data, analysis method, etc. Create standardized measurement models for X, Y, and $Z_1$ – $Z_I$. Label X-Z and Y-Z correlations:

```
x with y*;
x with z1* (xz1);
x with z2* (xz2);
x with z3* (xz3);
y with z1* (yz1);
y with z2* (yz2);
y with z3* (yz3);
z1 with z2*;
z1 with z3*;
z2 with z3*;
```

*WALD Test:*

Create new variables for ratios of correlations, and test them for equality:

```
model constraint:
  new (ratio1* ratio2* ratio3*);
  ratio1 = xz1 / yz1;
  ratio2 = xz2 / yz2;
  ratio3 = xz3 / yz3;

model test:
  ratio1 = ratio2;
  ratio1 = ratio3;
```

*DIFF Test:*

Create a new variable RHO, and constrain the XZ correlations to equal RHO times the YZ correlations. Compare the resulting model chi-square with an unconstrained confirmatory factor analysis result (e.g., from the WALD test analysis):

```
model constraint:
  new (rho*);
  xz1=rho*yz1;
  xz2=rho*yz2;
  xz3=rho*yz3;
```

## Appendix 2
## Congruence testing in composite-based SEM

Factor-based structural equation modeling employs the maximum likelihood method to estimate the model parameters. Composite-based SEM such as partial least squares (PLS-SEM; Wold, 1982) and generalized structured component analysis (GSCA; Hwang & Takane, 2014) minimizes a least squares criterion in parameter estimation. This fundamental difference has an important consequence for testing model fit. Factor-based SEM methods produce a model-implied correlation matrix that can be assumed to follow a chi-square distribution, while composite-based SEM methods cannot make this claim (Dijkstra & Henseler, 2015; Hwang & Takane, 2014). Therefore, chi-square difference tests are not available in composite-based SEM.

More generally, eminent SEM methodologists have noted several technical challenges with model fit statistics in PLS-SEM, and suggest that goodness-of-fit indices are questionable in a PLS-SEM context (Hair et al., 2019). Furthermore, PLS-SEM does not allow for imposing model constraints, which makes the method unsuitable for tests of model fit such as the WALD and DIFF tests. However, $r_c$ and $\theta$ can still be calculated using the same approaches as described in the first two examples of Appendix 1.

Conversely, GSCA minimizes a single least-squares criterion and thus alternative goodness-of-fit indices are available, such as standardized root mean square residual (SRMR) and the goodness of fit index (GFI; Ryoo & Hwang, 2017), whose efficacy has already been established in this methodological context (Cho et al., 2020). While GSCA currently allows for fixing of parameters, it does not allow for testing of model constraints, and thus the WALD and DIFF tests as described in the body of the paper and Appendix 1 cannot be implemented.

However, researchers using GSCA can apply nonparametric bootstrapping to construct the

sampling distribution of a test statistic of the deviance between two models such as an SRMR difference (Hwang & Takane, 2014, Chapter 3.1.2). Such a test could compare the goodness-of-fit of a standard model (such as the model in the upper panel in Fig. 2) with a nested model that includes a higher-order construct (such as the model in the lower panel in Fig. 2). A significantly different model fit would indicate that the lower-order constructs are not congruent. However, testing such an approach's performance and drawing conclusions on its suitability are beyond the scope of this article. We invite further research into the topic.

## References


Cho, G., Hwang, H., Sarstedt, M., & Ringle, C. M. (2020). Cutoff criteria for overall model fit indexes in generalized structured component analysis. *Journal of Marketing Analytics*, *8*(4), 189-202.

Dijkstra, T. K., & Henseler, J. (2015). Consistent and asymptotically normal PLS estimators for linear structural equations. *Computational Statistics & Data Analysis*, *81*(January), 10-23.

Hair, J. F., Sarstedt, M., & Ringle, C. M. (2019). Rethinking some of the rethinking of partial least squares. *European Journal of Marketing*, *53*(4), 566-584.

Hwang, H., & Takane, Y. (2014). *Generalized structured component analysis: A component-based approach to structural equation modeling*. Boca Raton, FL: CRC Press.

Ryoo, J. H., & Hwang, H. (2017). Model evaluation in generalized structured component analysis using confirmatory tetrad analysis. *Frontiers in Psychology*, *8*(916), 1-10.

Wold, H. O. A., 1982. Soft modelling: the basic design and some extensions. In: Jöreskog, K.G., Wold, H.O.A. (Eds.), *Systems under indirect observation. Causality, structure, prediction. Vol. II* (pp. 1-54). Amsterdam, New York, Oxford: North-Holland.


## Author biography

George R. Franke retired in 2019 as a Reese Phifer Professor of Marketing in the Culverhouse College of Business at the University of Alabama. In 25 years at Alabama and 12 at other universities, he received both best-paper and outstanding-reviewer awards from the *Journal of Advertising* and *Journal of Public Policy & Marketing*, and best-paper awards from the *Journal*

*of Marketing Research* and *Journal of Marketing Theory and Practice*. He also received three William R. Darden Best Marketing Research Paper Awards at conferences of the Academy of Marketing Science, where he and Marko Sarstedt presented an earlier version of this paper.

Marko Sarstedt is a chaired professor of marketing at the Otto-von-Guericke-University Magdeburg (Germany) and Adjunct Research Professor at Babeș-Bolyai-Universität Cluj. His main research interests are the advancement of research methods to further the understanding of consumer behavior. His research has been published in, for example, *Nature Human Behaviour*, *Journal of Marketing Research*, *Journal of the Academy of Marketing Science*, *International Journal of Research in Marketing*, *Organizational Research Methods*, *MIS Quarterly*, *Psychometrika*, *Journal of the Association for Information Systems*, *Decision Sciences*, *Journal of Business Research* and *Multivariate Behavioral Research*. Marko has coedited several special issues of leading journals and co-authored several widely adopted textbooks, including "A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)." He has been named member at Clarivate Analytic's 2019 Highly Cited Researcher List.

Nicholas P. Danks is an assistant professor of business analytics at Trinity College, Dublin. His research focuses on structural equation modeling, partial least squares, predictive methodology, and programming. Nicholas is a co-author and the primary maintainer of SEMinR, an open-source package for the R Statistical Environment for the estimation and evaluation of structural equation models. He received a William R. Darden Best Marketing Research Paper Award at the

conference of the Academy of Marketing Science. Nicholas has published in *Decision Sciences* and *Journal of Business Research.*