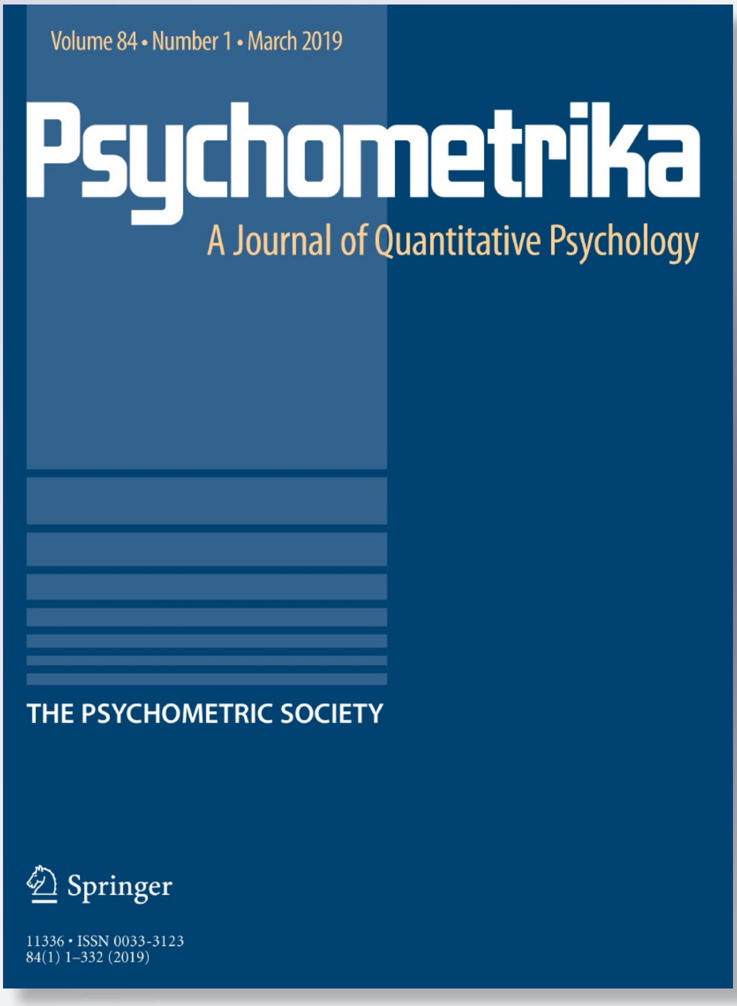
Tests of Matrix Structure for Construct Validation

Brian D. Segal, Thomas Braun, Richard Gonzalez & Michael R. Elliott

Psychometrika

ISSN 0033-3123
Volume 84
Number 1

Psychometrika (2019) 84:65-83
DOI 10.1007/s11336-018-9647-4



Volume 84 • Number 1 • March 2019

Psychometrika

A Journal of Quantitative Psychology

THE PSYCHOMETRIC SOCIETY

Springer

11336 • ISSN 0033-3123
84(1) 1–332 (2019)

PSYCHOMETRIKA—VOL. 84, NO. 1, 65–83
MARCH 2019
https://doi.org/10.1007/s11336-018-9647-4

TESTS OF MATRIX STRUCTURE FOR CONSTRUCT VALIDATION

BRIAN D. SEGAL, THOMAS BRAUN, RICHARD GONZALEZ AND MICHAEL R. ELLIOTT
UNIVERSITY OF MICHIGAN

Psychologists and other behavioral scientists are frequently interested in whether a questionnaire measures a latent construct. Attempts to address this issue are referred to as construct validation. We describe and extend nonparametric hypothesis testing procedures to assess matrix structures, which can be used for construct validation. These methods are based on a quadratic assignment framework and can be used either by themselves or to check the robustness of other methods. We investigate the performance of these matrix structure tests through simulations and demonstrate their use by analyzing a big five personality traits questionnaire administered as part of the Health and Retirement Study. We also derive rates of convergence for our overall test to better understand its behavior.

Key words: permutation testing, hubert's gamma, quadratic assignment.

1. Introduction

Psychologists and other behavioral scientists are frequently interested in whether a survey or questionnaire measures the concepts it purports to measure. Attempts to address this issue are referred to as construct validation. Since the construct cannot be directly observed, it is impossible to assess its validity directly. Instead, researchers divide construct validity into different aspects that can be addressed separately. These different aspects are called criterion-related validity, convergent validity, discriminant validity, and content validity. As Kline (2011) describes, criterion-related validity concerns the consistency of the test with external measures, convergent and discriminant validity refer to the magnitudes of correlations between test questions, and content validity is the degree to which the questions can be interpreted to represent the underlying scientific construct. By considering these different aspects of validity together, researchers can produce an overall body of evidence either in favor of or against validating a construct.

The statistical aspects of construct validation are covered by convergent and discriminant validity. Convergent validity occurs when the magnitudes of the correlations are high between items that are hypothesized to measure the same construct, and discriminant validity occurs when the magnitudes of the correlations are low between items hypothesized to measure different constructs (Kline 2011). In this paper, we describe and extend tests for matrix structure that can be used to assess convergent and discriminant validity, and derive rates of convergence for the overall test. These matrix structure tests can be used either by themselves or to check the robustness of other methods, such as confirmatory factor analysis (CFA).

In Sect. 2, we provide a motivating example. In Sect. 3, we describe and extend methods for testing matrix structure based on the quadratic assignment framework of Hubert and Schultz (1976), and derive rates of convergence for the overall test. In Sect. 4, we discuss related methods, including linear models, pattern hypothesis tests of correlation coefficients, and CFA. In Sect. 5,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11336-018-9647-4) contains supplementary material, which is available to authorized users.

Correspondence should be made to Brian D. Segal, University of Michigan, Ann Arbor, MI, USA. Email: bdsegal@umich.edu we investigate the behavior of these methods through simulations, and in Sect. 6, we demonstrate these methods by analyzing the big five personality traits questionnaire conducted as part of the 2010 Health and Retirement Survey (HRS 2016). In Sect. 7, we discuss the benefits and limitations of using tests of matrix structure for construct validation, as well as potential extensions. As noted in Sect. 8, we have implemented the methods described in this paper in the R package (`matrixTest`).

2. Motivating Example

As a motivating example, we analyze the big five personality traits questionnaire that was given as part of the 2010 Health and Retirement Study (HRS 2016). HRS is a "longitudinal panel study that surveys a representative sample of approximately 20,000 Americans over the age of 50 every two years" (HRS 2016). The big five personality traits questionnaire is given as part of the HRS Psychosocial and Lifestyle Questionnaire, which is administered to a rotating, random selection of 50% of the HRS respondents. The HRS data are publicly available at http://hrsonline.isr.umich.edu. The Psychosocial and Lifestyle Questionnaire is part of the core data release, in the file labeled `LB_R` (leave-behind, respondent).

In 2010, 7215 respondents provided complete responses to the big five personality trait questionnaire, and an additional 1050 subjects provided partial responses. The big five personality traits questionnaire contains 31 items, each of which was recorded on a four-point Likert scale. In what follows, we did a complete case analysis and did not incorporate sampling weights into the estimation of correlation coefficients, though this could be done in future analyses.

To assess convergent and divergent validity, we were interested in the magnitude of the correlations, but not the direction. Figure 1 shows the absolute values of Spearman's rank correlation matrix for the 31 items in the questionnaire, ordered by the hypothesized groups, which are outlined. From upper left to lower right, the outlined groups are: (1) neuroticism, (2) extroversion, (3) agreeableness, (4) openness to experience, and (5) conscientiousness. The questionnaire items are described in "Web Appendix C in Supplementary material".

From a visual inspection of Fig. 1, the first block (neuroticism) appears to exhibit both convergent validity (high within-block correlation) and divergent validity (low between-block correlation). The second, third and fourth blocks (extroversion, agreeableness, and openness to experience) appear to exhibit convergent validity, though the relatively high correlations between these blocks makes it unclear whether they also exhibit divergent validity. The fifth block (conscientiousness) does not appear to exhibit either convergent or divergent validity. We next develop methods to formally test convergent and divergent validity using nonparametric tests of matrix structure.

3. Tests of Matrix Structure

Several authors have developed methods for testing matrix structure, including Bock and Bargmann (1966), Srivastava (1966), McDonald (1974) and Jöoreskog (1978). The approach we describe has a similar goal to these methods, but differs in the way hypothesized matrix structures are assessed. Most notably, our approach sets up a traditional null hypothesis that researchers seek to reject, and does not use a goodness of fit (GOF) test or index to evaluate model fit.

3.1. Block Diagonal Structure

Let A be a $p \times p$ symmetric matrix. In our applications, A is typically the covariance or correlation matrix, or the absolute values of the covariance or correlation matrix. We are interested in whether A is approximately block diagonal:

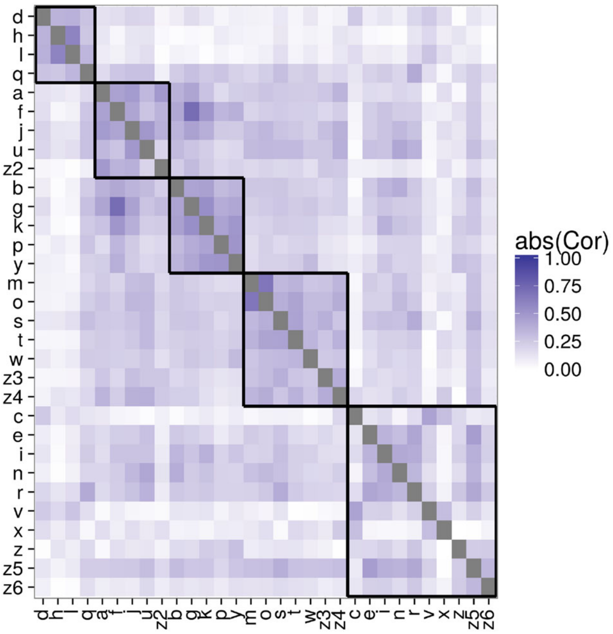


FIGURE 1.
Absolute values of the Spearman rank correlation matrix for the HRS big five personality traits questionnaire ordered by hypothesized groups. From upper left to lower right, the groups are: (1) neuroticism, (2) extroversion, (3) agreeableness, (4) openness to experience, and (5) conscientiousness. Diagonal elements are all equal to 1, and are not included in the color gradient. Item labels (d, h, l, …) are taken from the HRS questionnaire. The items are described in "Web Appendix C in Supplementary material".

$$A = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_K \end{pmatrix}$$

where blocks A_1 through A_K have respective dimensions $p_1 \times p_1, \ldots, p_K \times p_K$, and $\sum_{k=1}^{K} p_k \leq p$. When A is the covariance matrix, this is the structure implied by a CFA model in which each item loads onto no more than one latent variable. Throughout this paper, we use the terms *group* and *block* interchangeably.

By approximately block diagonal, we mean that the elements in blocks $A_1, \ldots, A_K$ are larger in absolute value than elements in the non-blocks. If A were perfectly block diagonal, all elements in blocks $A_1, \ldots, A_K$ would be nonzero, and all other elements would be zero. Figure 1 is an example where A is the element-wise absolute values of the correlation matrix, with $p = 31$ variables and a hypothesized $K = 5$ blocks of sizes $p_1 = 4$, $p_2 = 5$, $p_3 = 5$, $p_4 = 7$, and $p_5 = 10$. If we exclude the fifth block from Fig. 1, then $\sum_{k=1}^{4} p_k < p$ and the hypothesized block diagonal structure would not extend all the way to the bottom right corner of the correlation matrix.

3.2. *Hubert's* Γ

Hubert's Γ (Hubert and Schultz 1976) was originally proposed by Mantel (1967). Consequently, some authors, including Good (2000), refer to the statistic as Mantel's U. However, we follow most authors, including Jain and Dubes (1988), Halkidi et al. (2001) and Zaki and Meira (2014) and refer to the statistic as Hubert's Γ, especially since our methods are based on the quadratic assignment framework of Hubert and Schultz (1976).

To define Hubert's Γ, let v_i be the label for the variable in row and column i of matrix A and let Δ be a $p \times p$ matrix with element δ_{ij} in row i and column j, where

$$\delta_{ij} = \begin{cases} 1 & \text{if variables } v_i \text{ and } v_j \text{ are hypothesized to belong to the same block} \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, we denote the element in row i and column j of A as a_{ij}. Let $N = p(p-1)/2$ be the number of upper triangular elements in A, where the upper triangular elements form the set $\{a_{ij} : i < j\}$. Let $\boldsymbol{a} = (a_{12}, a_{13}, a_{23}, a_{14}, a_{24}, \ldots, a_{N-1,N})^T$ be the $N \times 1$ vector of the upper triangular elements of A, and let $\boldsymbol{\delta} = (\delta_{12}, \delta_{13}, \delta_{23}, \delta_{14}, \delta_{24}, \ldots, \delta_{N-1,N})^T$ be the $N \times 1$ vector of the upper triangular elements of Δ. Since the A and Δ matrices are symmetric, we do not need to consider the lower triangular elements. Hubert's Γ is defined as the mean element-wise product between the upper triangular elements of A and Δ, given by $\Gamma = N^{-1} \sum_{i<j} a_{ij}\delta_{ij} = N^{-1}\boldsymbol{a}^T\boldsymbol{\delta}$.

We use the normalized Γ, which is more interpretable. Let $\bar{a} = N^{-1} \sum_{i<j} a_{ij}$ and $\hat{\sigma}_a^2 = (N-1)^{-1} \sum_{i<j} (a_{ij} - \bar{a})^2$ be the sample mean and variance of the elements in $\boldsymbol{a}$, let $\bar{\delta} = N^{-1} \sum_{i<j} \delta_{ij}$ and $\hat{\sigma}_\delta^2 = (N-1)^{-1} \sum_{i<j} (\delta_{ij} - \bar{\delta})^2$ be the sample mean and variance of the elements in $\boldsymbol{\delta}$, and let $\hat{\sigma}_{a\delta}^2 = (N-1)^{-1} \sum_{i<j} (a_{ij} - \bar{a})(\delta_{ij} - \bar{\delta})$ be the sample covariance between $\boldsymbol{a}$ and $\boldsymbol{\delta}$. Then the normalized Γ, which we denote as Γ_{norm}, is defined as the Pearson correlation between $\boldsymbol{a}$ and $\boldsymbol{\delta}$, given by

$$\Gamma_{\text{norm}} = \frac{\sum_{i<j} (a_{ij} - \bar{a})(\delta_{ij} - \bar{\delta})}{\sqrt{\sum_{i<j} (a_{ij} - \bar{a})^2 \sum_{i<j} (\delta_{ij} - \bar{\delta})^2}} = \frac{\hat{\sigma}_{a\delta}^2}{\hat{\sigma}_a \hat{\sigma}_\delta}. \tag{1}$$

Since Γ_{norm} is a correlation, $-1 \leq \Gamma_{\text{norm}} \leq 1$.

In general, the Δ matrix can be replaced by any conformable matrix in calculating Γ and Γ_{norm} depending on the hypothesis a researcher wants to test. As we show in Sect. 4.1, Γ_{norm} with Δ as defined above is related to the slope from a linear model that contrasts the within-block elements with the between-block elements.

Large positive values of Γ_{norm} (values near 1) indicate that overall, the clustering has a high degree of convergent and discriminant validity. If Γ_{norm} is near zero, then either the clusters have low levels of convergent validity, discriminant validity, or both. If Γ_{norm} is large and negative, then we have likely flipped blocks with non-blocks, and would have reason to revisit the exploratory analysis.

3.3. *Permutation Test*

The null hypothesis in our permutation test is that off-diagonal elements of A are exchangeable. Rejecting the null is evidence in favor of the hypothesized latent structure.

Other authors, including Jain and Dubes (1988), have proposed permutation tests with Hubert's Γ to test overall null hypotheses. However, in most existing applications A is an $n \times n$ matrix that measures proximity between subjects as opposed to a $p \times p$ matrix that measures correlation between items on a questionnaire. Application of the permutation test to correlation matrices requires additional considerations when defining the null hypothesis, as described in Sect. 3.3.1, and also when defining and deriving rates of convergence, as described in Sect. 3.5 and "Web Appendix A in Supplementary material".

In addition to the overall test, we propose a block-specific test in Sect. 3.3.2. To the best of our knowledge, the block-specific test has not been proposed previously, and we find it to be highly informative in our simulations and application.

3.3.1. Overall Test As before, let v_i be the label of the ith column and row of A, and let $v = (v_1, \ldots, v_p)$ be the ordered sequence of labels. For example, if A is the matrix of correlations among items on a questionnaire, then v_i would be the ith item on the questionnaire. Also, let π be a permutation of the indices of v, such that $\pi(i)$ is the index to which i is mapped after the permutation. For example, if the indices $(1, 2, 3)$ are permuted to $(3, 1, 2)$, then $\pi(1) = 2$. Let $v^* = (v_1^*, \ldots, v_p^*)$ be a permuted sequence of labels, where $v^*_{\pi(i)} = v_i, i = 1, \ldots, p$. For example, in Sect. 2, the items in the Big Five questionnaire are labeled as $v = (v_1 = a, v_2 = b, \ldots, v_{31} = z6)$, and under the hypothesized ordering shown along the rows and columns of Fig. 1, $v^* = (v_1^* = d, v_2^* = h, v_3^* = l, \ldots, v_{31}^* = z6)$.

In the permutation test, we keep the Δ matrix constant, permute the order of the labels in A, and recompute the test statistic Γ_{norm}. In keeping Δ constant, we are conditioning on the hypothesized number of blocks K and block sizes $p_k, k = 1, \ldots, K$. This conditioning is an important constraint needed in the permutation test.

If we randomly sample B permutations $\pi_1, \ldots \pi_B$ with replacement, then the Monte Carlo (MC) approximation to the two-sided permutation p-value is (Lehmann and Romano 2005)

$$\tilde{p} = \frac{1}{B+1}\left[\sum_{b=1}^{B} \mathbb{1}\left(\left|\Gamma^b_{\text{norm}}\right| \geq \left|\Gamma^0_{\text{norm}}\right|\right) + 1\right],$$

where $\mathbb{1}$ is an indicator function, Γ^0_{norm} is the test statistic under the hypothesized clustering, and Γ^b_{norm} is the test statistic from the bth randomly sampled permutation π_b. That is, $\tilde{p}$ represents the proportion of MC resamples with test statistics that exceed the observed test statistic under the hypothesized clustering.

Exchangeable off-diagonal elements implies a variety of matrix structures, including constant off-diagonal elements (referred to by Steiger (1980a) as equicorrelation in the case where A is the correlation matrix) and white noise. Under constant off-diagonal elements, A is of the form $A = a\mathbf{1}\mathbf{1}' + (\boldsymbol{b} - a\mathbf{1})'I$ for some $a \in \mathbb{R}$ and $\boldsymbol{b} \in \mathbb{R}^p$, where $\mathbf{1}$ is a column vector of 1's and I is the identity matrix (for correlation matrices, $\boldsymbol{b} = \mathbf{1}$ and $a \in [-1, 1]$):

$$A = \begin{pmatrix} b_1 & & & \Large a \\ & b_2 & & \\ & & \ddots & \\ \Large a & & & b_p \end{pmatrix}.$$

More generally, under white noise we assume the off-diagonal elements $a_{ij} \sim P, i < j$ for some common distribution P. If A is a covariance or correlation matrix, then we have the additional constraint that A is symmetric and positive semi-definite. If P has zero variance, we obtain constant off-diagonals.

3.3.2. Block-Specific Test In addition to the overall test, we can test each block individually to see if the magnitude of the within-block elements are larger than the magnitude of the corresponding between-block elements. To this end, let $\Gamma_{\text{norm},k}$ be the same as above, except that the sum is restricted to (i, j) such that at least one of v_i, v_j is in block k. As before, we remove variance terms from the sum. To be precise, let $\mathcal{V}_k$ be the set of labels assigned to block k, and let $\mathcal{I}_k = \{(i, j) : v_i \in \mathcal{V}_k \text{ or } v_j \in \mathcal{V}_k, i < j\}$ be the set of ordered index pairs with at least one index in block k. Let $N_k = |\mathcal{V}_k|$ be the number of elements in $\mathcal{V}_k$, and let $\bar{a}_k = N_k^{-1} \sum_{(i,j)\in\mathcal{I}_k} a_{ij}$ and $\hat{\sigma}^2_{a,k} = (N_k - 1)^{-1} \sum_{(i,j)\in\mathcal{I}_k} (a_{ij} - \bar{a}_k)^2$ be the sample mean and variance of elements in the set $\{a_{ij} : (i, j) \in \mathcal{I}_k\}$, and $\bar{\delta}_k = N_k^{-1} \sum_{(i,j)\in\mathcal{I}_k} \delta_{ij}$ and $\hat{\sigma}^2_{\delta,k} = (N_k - 1)^{-1} \sum_{(i,j)\in\mathcal{I}_k} (\delta_{ij} - \bar{\delta}_k)^2$ be the sample mean and variance of elements in the set $\{\delta_{ij} : (i, j) \in \mathcal{I}_k\}$. Also, let $\hat{\sigma}^2_{a\delta,k} = (N_k - 1)^{-1} \sum_{(i,j)\in\mathcal{I}_k} (a_{ij} - \bar{a}_k)(\delta_{ij} - \bar{\delta}_k)$ be the sample covariance. Then we define

$$\Gamma_{\text{norm},k} = \frac{\sum_{(i,j)\in\mathcal{I}_k} (a_{ij} - \bar{a}_k)(\delta_{ij} - \bar{\delta}_k)}{\sqrt{\sum_{(i,j)\in\mathcal{I}_k} (a_{ij} - \bar{a}_k)^2 \sum_{(i,j)\in\mathcal{I}_k} (\delta_{ij} - \bar{\delta}_k)^2}} = \frac{\hat{\sigma}^2_{a\delta,k}}{\hat{\sigma}_{a,k}\hat{\sigma}_{\delta,k}}.$$

When testing multiple blocks, to control the family-wise error rate we follow Westfall and Young (1993) and for each permutation π_b set $\Gamma^{\max,b}_{\text{norm}} = \max_{k\in\{1,\ldots,K\}} |\Gamma^b_{\text{norm},k}|$, where $\Gamma^b_{\text{norm},k}$ is the computed statistic for block k under permutation π_b. We then compute the MC estimate of the two-sided permutation p-value for block k as

$$\tilde{p}_k = \frac{1}{B+1}\left[\sum_{b=1}^{B} \mathbb{1}\left(\Gamma^{\max,b}_{\text{norm}} \geq \left|\Gamma^0_{\text{norm},k}\right|\right) + 1\right].$$

3.4. *Recommendations for Choosing Matrix A*

In construct validation, the primary question concerns the magnitude of association, as opposed to the direction. Furthermore, in most questionnaires, the direction of correlation is arbitrary. For example, in the HRS big five personality questionnaire, some items are reverse coded to preserve positive correlations among items hypothesized to measure the same latent construct. Consequently, in some applications A could be set to the element-wise absolute correlations, as in the motivating example in Sect. 2. By using the absolute values of the correlations, we avoid potentially overlooking associations between items that are coded in such a way that their correlations are negative.

We use Spearman's rho so that our test is robust to non-normal data and nonlinear associations. However, we speculate that other nonparametric correlation coefficients would also be reasonable, such as Kendall's tau and Goodman and Kruskal's gamma. Ultimately, we recommend that researchers use a matrix A that best measures the phenomenon of interest, which may differ across applications.

3.5. *Convergence Rates*

In this section, we denote the estimated quantities obtained with n observations as $\boldsymbol{a}^n = (a^n_1, \ldots, a^n_N)^T$, and describe the convergence rate as $n \to \infty$. In data analyses, we use Monte Carlo methods to approximate the permutation p value obtained with the estimated quantities $\boldsymbol{a}^n$. We denote the permutation p value with the estimated quantities as $\hat{p}(\boldsymbol{a}^n)$. However, we would ideally approximate the permutation p value obtained with the true population values, which we denote as $\hat{p}(\boldsymbol{\rho})$, where $\boldsymbol{\rho}$ are the true population values. Assuming $\boldsymbol{a}^n$ is a consistent estimator of $\boldsymbol{\rho}$, $\boldsymbol{a}^n \to \boldsymbol{\rho}$ as $n \to \infty$. In this section, we address the rate at which the overall permutation p value computed with the estimated values $\hat{p}(\boldsymbol{a}^n)$ converges to the overall permutation p value computed with the true values $\hat{p}(\boldsymbol{\rho})$. These results hold for the overall test.

As stated in Theorem 1, under fairly general conditions, the permutation p value for the overall test has the same rate of convergence as the elements of $\boldsymbol{a}^n$.

Theorem 1. *Let a_j^n be the sample estimates of ρ_j, $j = 1, \ldots, N$, and suppose that for all j, $|a_j^n - \rho_j| = O_p(g(n))$ for some strictly decreasing function g, such that $g(n) \to 0$ as $n \to \infty$. Also suppose that the permutation distribution $\hat{R}_N(t)$ has limiting distribution $R(t)$ such that the density of $R(t)$, denoted as $f(t)$, exists and $\sup_t f(t) < \infty$. Then for N sufficiently large, $|\hat{p}(\boldsymbol{a}^n) - \hat{p}(\boldsymbol{\rho})| = O_p(g(n))$.*

Furthermore, as described in Corollary 1, when $\boldsymbol{a}^n$ are Pearson's or Spearman's correlations, $|\hat{p}(\boldsymbol{a}^n) - \hat{p}(\boldsymbol{\rho})| = O_p(1/\sqrt{n})$. As described in Corollary 2, the same rate holds when using the absolute values of Pearson's or Spearman's correlations.

Corollary 1. *Let $\boldsymbol{a}^n$ be Pearson's or Spearman's correlation coefficients estimated from n independent and identically distributed (i.i.d.) observations. Let $\tau_j^2 = Var(a_j^n)$ and assume $\tau_j^2 < \infty$ for $j = 1, \ldots, N$. Also suppose that the permutation distribution $\hat{R}_N(t)$ has limiting distribution $R(t)$ such that the density of $R(t)$, denoted as $f(t)$, exists and $\sup_t f(t) < \infty$. Then for N sufficiently large, $|\hat{p}(\boldsymbol{a}^n) - \hat{p}(\boldsymbol{\rho})| = O_p(1/\sqrt{n})$.*

Corollary 2. *Under the same conditions as Corollary 1, but with $\boldsymbol{a}^n$ and $\boldsymbol{\rho}$ replaced with absolute values of Pearson's or Spearman's correlations, we also have $|\hat{p}(\boldsymbol{a}^n) - \hat{p}(\boldsymbol{\rho})| = O_p(1/\sqrt{n})$.*

For details and proofs, please see "Web Appendix A in Supplementary material".

4. Comparison to Related Methods

4.1. Linear Model and t test

To better understand and interpret Γ_{norm}, we note that because Γ_{norm} is a correlation, it is permutationally equivalent to the ordinary least squares coefficient from a simple linear regression model where the outcomes are the absolute values of the correlation coefficients $\boldsymbol{a}$ and the covariates are the indicators $\boldsymbol{\delta}$.

To see this, we write the linear model as

$$\mathbb{E}[\boldsymbol{a}] = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{\delta} \tag{2}$$

where $\mathbf{1}$ is an $N \times 1$ vector. The ordinary least squares estimate for (2) is $\hat{\beta}_1 = (\hat{\sigma}_a/\hat{\sigma}_\delta)\Gamma_{\text{norm}}$.

Let $\mathcal{W}_k = \{(i, j) : v_i \in \mathcal{V}_k, v_j \in \mathcal{V}_k, i < j\}$ be the set of ordered index pairs for upper triangular elements such that both indices are in block k, let $N_{\text{in},k} = |\mathcal{W}_k|$ be the number of elements in $\mathcal{W}_k$, and $N_{\text{in}} = \sum_k N_{\text{in},k}$ be the total number of upper triangular within-block elements. Also, let $\mathcal{W}_{\text{out}} = \{(i, j) : (i, j) \notin \mathcal{W}_k, k = 1, \ldots, K, i < j\}$ be the set of ordered index pairs for upper triangular elements not in blocks, and $N_{\text{out}} = |\mathcal{W}_{\text{out}}|$ be the number of non-block elements. Then, because Δ is a matrix of zeros and ones, we have $\hat{\beta}_1 = \bar{a}_{\text{in}} - \bar{a}_{\text{out}}$, where $\bar{a}_{\text{in}} = N_{\text{in}}^{-1} \sum_k \sum_{(i,j)\in\mathcal{W}_k} a_{ij}$ and $\bar{a}_{\text{out}} = N_{\text{out}}^{-}1 \sum_{(i,j)\in\mathcal{W}_{\text{out}}} a_{ij}$ are the mean within-block and between-block elements, respectively. In the overall test, $\hat{\sigma}_a^2$ and $\hat{\sigma}_\delta^2$ are constant across permutations. Therefore, there is a one-to-one relationship between Γ_{norm} and $\hat{\beta}_1$, and they are permutationaly equivalent. In other words, $\hat{\beta}_1$ could be substituted for Γ_{norm} in the permutation test to obtain the same permutation p-value. When restricting to subsets of the matrix to evaluate $\Gamma_{\text{norm},k}$, $\hat{\sigma}_{a,k}^2$ is no longer constant across permutations, so $\Gamma_{\text{norm},k}$ and $\hat{\beta}_{1,k}$ are no longer permutationaly equivalent.

We also note that the t-statistic with unequal variance has potential advantages over the statistics Γ_{norm} and $\hat{\beta}_1$. In particular, the t-statistic with unequal variance controls the type I error rate in permutation tests under the null $H_0 : \bar{a}_{\text{in}} = \bar{a}_{\text{out}}$ versus $H_1 : \bar{a}_{\text{in}} \neq \bar{a}_{\text{out}}$ even if the variance of the within-block and between-block correlations are different (Chung and Romano, 2013). The t-statistic with unequal variance is given by

$$t = \frac{\bar{a}_{\text{in}} - \bar{a}_{\text{out}}}{\sqrt{\hat{\sigma}^2_{\text{in}}/N_{\text{in}} + \hat{\sigma}^2_{\text{out}}/N_{\text{out}}}} \tag{3}$$

where $\hat{\sigma}^2_{\text{in}} = (N_{\text{in}} - 1)^{-1} \sum_k \sum_{(i,j)\in\mathcal{W}_k} (a_{ij} - \bar{a}_{\text{in}})^2$ and $\hat{\sigma}^2_{\text{out}} = (N_{\text{out}} - 1)^{-1} \sum_{(i,j)\in\mathcal{W}_{\text{out}}} (a_{ij} - \bar{a}_{\text{out}})^2$ are the sample variances of the within-block and between-block upper triangular elements of A, respectively.

Due to the results of Chung and Romano (2013), it may be beneficial to use the studentized statistic t given by (3) in future work in place of Hubert's Γ, as it leads to permutation tests that are valid under a wider range of scenarios than those we examined in our simulations. However, in our simulations, the use of (3) in the permutation test gave nearly identical results to those obtained with Γ_{norm}.

4.2. Goodness of Fit (GOF) Tests

Several statistical methods used in construct validation rely on a goodness of fit (GOF) test, including CFA and pattern hypothesis tests (Steiger 2007). Frequently, GOF tests are based on χ^2 statistics. In general terms, the null hypothesis in GOF tests is H_0: "the model fits" and the alternative is H_1: "the model does not fit." Under this framework, failure to reject the null is evidence in favor of the scientific theory. This is in contrast to the tests of matrix structure described in Sect. 3.3, for which rejection of the null is evidence in favor of the scientific theory.

Since GOF tests reverse the usual role of the null and alternative hypotheses, the interpretation of type I and II errors is also reversed. To guard against making false scientific claims, one needs to avoid accepting the null when the alternative is true—a type II error. Similarly, to increase the chances of finding evidence in favor of a scientific theory, one needs to avoid rejecting the null when the null is true. Given the analogy with statistical power, we refer to this as type I power. Since this is not a standard term, we define it in Definition 1.

Definition 1. (Type I power) Type I power is the probability of failing to reject the null hypothesis when the null hypothesis is true: Pr(fail to reject $H_0|H_0$ true).

The reversal in GOF tests of the standard scientific interpretation of Type I and II errors may have several implications for the reliability of GOF tests in evaluating scientific hypotheses. In particular, failure to control type II errors in GOF tests could lead to higher than expected rates of false scientific claims, and low type I power would make it difficult to find evidence in favor of a scientific claim. Table 1 shows these differing interpretations, and proposes the terms "GOF false alarms" and "GOF missed opportunities" to describe the potential errors when conducting a GOF test. We are unaware of work aimed at controlling type II error rates in GOF tests, but several researchers have suggested ways to address low type I power. Contrary to standard statistical power, type I power decreases as sample size increases, making low Type I power a pervasive problem.

4.2.1. GOF Tests in Structural Equation Models (SEMs) To address low type I power in structural equation models (SEMs), including CFA, researchers have developed alternative fit indices, many of which adjust the χ^2 GOF statistic based on the degrees of freedom, such as the comparative fit index (CFI) (Bentler 1990) and Tucker–Lewis Index (TLI) (Tucker and Lewis 1973). The root mean squared error of approximation (RMSEA) (Steiger and Lind 1980, Steiger 1990) is another commonly used fit index. However, as shown in Sect. 5.1, the type I power of

TABLE 1.
Comparison of interpretation of errors under traditional and GOF frameworks.

	Truth	
Decision	H_0	H_1
H_0 $(p - \text{val} \geq c)$	Correct failure to reject H_0 **Type I power**	Type II error (missed opportunity) **GOF false alarm**
H_1 $(p - \text{val} < c)$	Type I error (false alarm) **GOF missed opportunity**	Power **Correct rejection of H_0**

Within each cell, the traditional interpretation is on the first line, and the GOF interpretation is on the second line in bold. H_0 is the null hypothesis and H_1 is the alternative hypothesis (H_0 rejected if p value $< c$ for cutoff value c).

CLI, TLI, and RMSEA decreases as sample size increases, though not as dramatically as for unadjusted χ^2 GOF statistics.

Many of the rules of thumb for interpreting fit indices have roots in the work of Hu and Bentler (1999). For CFI and TLI, values above 0.95 are commonly considered to indicate acceptable fit (Hu and Bentler 1999). However, Hooper et al. (2008) notes that some researchers have suggested a cutoff value of 0.9 for CFI and 0.8 for TLI. We show simulation results with all three cutoffs in Sect. 5.

For RMSEA, values less than 0.06 (Hu and Bentler 1999) or 0.07 (Steiger 2007) are commonly considered to indicate acceptable fit, though recommendations vary. For example, Browne and Cudeck (1992) suggest that a value less than 0.05 indicates close fit, that values as large as 0.08 may show reasonable fit, and that values greater than 0.1 indicate a lack of fit. We show simulation results with cutoffs of 0.05, 0.07, and 0.1 in Sect. 5.

As Barrett (2007) notes, some simulation studies, including Marsh et al. (2004), Beauducel and Wittmann (2005), Yuan (2005), and Fan and Sivo (2005), have cast doubt on the reliability of these rules of thumb for CFI, TLI, and RMSEA. We note that the criticism of Barrett (2007) is controversial, and Steiger (2007) offers a rebuttal. Kline (2011) and Hu and Bentler (1999) offer discussions on fit statistics and indices for SEMs, and we refer the reader to these sources for details.

4.2.2. Pattern Hypothesis GOF Tests As Steiger (1980b) describes, a pattern hypothesis is “any hypothesis that states that some of its elements are equal to each other and/or to specified numerical values." Using the same notation as before, let $\boldsymbol{a}$ be the $N \times 1$ vector of upper triangular elements of A. Pattern hypotheses are of the form (Steiger 1980b)

$$H_0: \ \boldsymbol{a} = L\boldsymbol{\beta} + \boldsymbol{a}^*, \tag{4}$$

where $\boldsymbol{\beta}$ is a $q \times 1$ vector of parameters to be estimated, $\boldsymbol{a}^*$ is $q \times 1$ vector of constants, and L is an $N \times q$ matrix of zeros and ones, with

$$L_{ij} = \begin{cases} 1 & \text{if the } ith \text{ element of } \boldsymbol{a} \text{ is hypothesized to equal } \beta_j \\ 0 & \text{otherwise,} \end{cases}$$

In the case where A is a covariance matrix, pattern hypothesis tests are related to the analysis of covariance structures (Bock and Bargmann 1966).

If we set $q = 2$, then (4) would be a re-parameterization of (2). In this case, to recover (2) from (4), we would set $\boldsymbol{a}^*$ to zero and reparameterize L as $L = [\mathbf{1}, \boldsymbol{\delta}]$. This changes L from being a cell-means coding to a reference cell coding.

For the rest of this section, we assume A is the Pearson correlation matrix for underlying data $\boldsymbol{y}_l = (y_{l1}, \ldots, y_{lp})^T, l = 1, \ldots, n$, in which we have n observations of p variables. In particular, let $\bar{y}_i = n^{-1}\sum_{l=1}^{n} y_{li}$, $\hat{\sigma}_i^2 = (n-1)^{-1}\sum_{l=1}^{n}(y_{li} - \bar{y}_i)^2$ and $\hat{\sigma}_{ij}^2 = (n-1)^{-1}\sum_{l=1}^{n}(y_{li} - \bar{y}_i)(y_{lj} - \bar{y}_j)$. Then $a_{ij} = \hat{\sigma}_{ij}^2/(\hat{\sigma}_i\hat{\sigma}_j)$. In this case, we set $\boldsymbol{r} = \boldsymbol{a}$ to use more familiar notation. If the underlying data are i.i.d. multivariate normal, then we can induce normality on the correlation coefficients by taking the Fisher r-to-z variance stabilizing transformation, denoted as $z(r)$, where (Fisher 1921)

$$z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) = \operatorname{arctanh}(r).$$

The Fisher transformation improves the normal approximation to the distribution of the correlation coefficients, even if the underlying data are not normal, though the form of the $N \times N$ covariance matrix $\mathrm{Var}(z(\boldsymbol{r}))$ may not be the same as for normal data (Hawkins, 1989).

Following Steiger (1980b), we test the null hypothesis (4) with the GOF χ^2 statistic

$$X_2 = (n-3)\left[z(\boldsymbol{r}) - z(\hat{\boldsymbol{r}}_{\mathrm{GLS}})\right]^T S_{LS}^{-1}\left[z(\boldsymbol{r}) - z(\hat{\boldsymbol{r}}_{\mathrm{GLS}})\right], \tag{5}$$

where $\hat{\boldsymbol{r}}_{\mathrm{GLS}} = L(L^T\hat{\Sigma}_{\mathrm{LS}}^{-1}L)^{-1}L^T\hat{\Sigma}_{\mathrm{LS}}^{-1}\boldsymbol{r}$, $\hat{\Sigma}_{\mathrm{LS}}$ is the covariance matrix with elements given by Steiger (1980b) with $\hat{\boldsymbol{r}}_{\mathrm{LS}} = L(L^TL)^{-1}L^T\boldsymbol{r}$ substituted for $\boldsymbol{r}$, and S_{LS} is the covariance matrix with elements also given by Steiger (1980b). Asymptotically, X_2 follows a χ^2 distribution with $N-2$ degrees of freedom (Steiger 1980b).

The permutation test with Γ_{norm} and the GOF χ^2 test with (5) are similar, but with important differences. In (4), and assuming $q = 2$, let $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. Then the permutation test is similar to obtaining a p value for the null hypothesis $H_0 : \beta_1 = 0$, whereas (5) gives a p value for the GOF null hypothesis H_0 : "the model fits." In addition, the permutation test is nonparametric and relies only on the exchangeability of off-diagonal elements, as opposed to the GOF test with (5), which relies on asymptotic approximations to obtain the reference distribution. The permutation test is also applicable for a variety of matrices A, whereas the asymptotic reference distribution for (5) is valid only for certain types of matrices.

5. Simulations

In this section, we simulated data under two scenarios: (1) block diagonal structure, and (2) random off-diagonal values (white noise). For each scenario, we generated 1000 datasets for each sample size. For simulations under the permutation null hypothesis, we used sample sizes of $n = 10$, 100, and 1000 with $B = 1000$ resamples. For simulations under the permutation alternative hypothesis, we used samples sizes of $n = 10$, 50, 100, and 1000 with $B = 10{,}000$ resamples to better approximate small p values and statistical power. For all simulations, we used $K = 4$ blocks of sizes $p_1 = 5$, $p_2 = 7$, $p_3 = 9$, $p_4 = 11$, so that the total number of variables was $p = \sum_k p_k = 32$. In all figures, the block numbers begin in the upper left and end in the lower right, i.e., block $k = 1$ is in the top left corner, and block $k = 4$ is in the bottom right corner.

In the matrix structure testing framework, Sect. 5.1 is under the alternative hypothesis (H_1 is true) and Sect. 5.2 is under the null hypothesis (H_0 is true). In the GOF framework, the model is correctly specified in Sect. 5.1 (H_0 is true) and misspecified in Sect. 5.2 (H_1 is true).

We followed our recommendations in Sect. 3.4 and used absolute Spearman correlation coefficients when computing Γ_{norm}, though we acknowledge that other choices for the matrix A are possible. For the pattern hypothesis test, we used Pearson's correlation and Fisher's r-to-z transform to compute X_2, as described in Sect. 4.2.2. To obtain CFI, TLI, and RMSEA, we fit CFA models with $K = 4$ latent factors and p_k items loading onto the kth factor, with p_k given above. In the CFA models, each item loaded onto exactly one factor. The tables in this section do not directly compare the results with the permutation test against those from CFI, TLI, and RMSEA, because different types of errors are relevant for the two approaches.

In "Web Appendix B in Supplementary material", we also show simulations under four additional scenarios: (1) constant off-diagonal values, (2) block diagonal structure on a subset of the matrix and white noise on the rest of the matrix (partial block diagonal structure), (3) a true CFA generating process, and (4) a true CFA generating process followed by the discretization of the outcome. For the GOF tests, the first two scenarios are under the alternative and the third and fourth are under the null. CFI, TLI, and RMSEA gave high GOF false alarm rates in the white noise scenario, low GOF false alarm rates in the partial block diagonal scenario, and moderate to high type I power for sample sizes of $n = 100$ and 1000 under the true CFA generating process with both continuous and discretized outcomes. In neither of the first two scenarios did the GOF indices identify the source of poor fit.

For the permutation test, the constant off-diagonal scenario is under the null, and simulation results show that the permutation test controls the type I error rate at the nominal level, as expected. The partial block diagonal and CFA generating scenarios are under the alternative for the permutation test, and simulation results show that the permutation test has high power.

5.1. Block Diagonal Structure

To simulate data under the scenario of a block diagonal correlation matrix, we began by generating the square root of the variance matrix $\Sigma^{1/2}$ such that variables within groups would be correlated with each other, and variables across groups would have minimal but nonzero correlations. In particular, we set $\Sigma_{ij}^{1/2} = \sum_k \mathbb{1}[v_i \in \mathcal{V}_k, v_j \in \mathcal{V}_k]r_k + u_{ij}$, where $r_1 = 0.25$, $r_2 = 0.2$, $r_3 = 0.23$, $r_4 = 0.15$, and $u_{ij} \sim N(0, 0.01)$.

For each sample size, we simulated 1000 $n \times p$ datasets, Y_t, $t = 1, \dots, 1000$, where

$$Y_t = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix},$$

and $\boldsymbol{y}_l = (y_{l1}, \dots, y_{lp})^T$, $l = 1, \dots, n$, were generated independently as $N(\boldsymbol{0}, \Sigma_t)$ random vectors with Σ_t generated as described above. We then created corresponding $n \times p$ datasets Z_t, $t = 1, \dots$, 1000, of ordinal variables where for each dataset, $z_{li} = 1$ if $y_{li} < -2$, $z_{li} = 2$ if $-2 \le y_{li} < -1$, $z_{li} = 3$ if $-1 \le y_{li} < 0$, $z_{li} = 4$ if $0 \le y_{li} < 1$, $z_{li} = 5$ if $1 \le y_{li} < 2$, and $z_{li} = 6$ if $2 \le y_{li}$.

For each dataset, we estimated Spearman's correlation matrix, which we denote as $C = C(Z)$, and conducted a permutation test with Hubert's Γ on $A = \text{abs}(C)$ where the absolute values are taken element-wise. We used $B = 10,000$ MC resamples for the permutation tests. We also computed X_2 with the Pearson correlation matrix of Z (treating the ordinal data as numeric), and fit a CFA model with the data Z (treating the data as ordinal) using the `lavaan` package (Rosseel 2012) for R (R Core Team 2017).

Figure 2 shows the estimated Spearman's absolute correlation matrices A from a single simulation for sample sizes of $n = 10$, 100, and 1000.

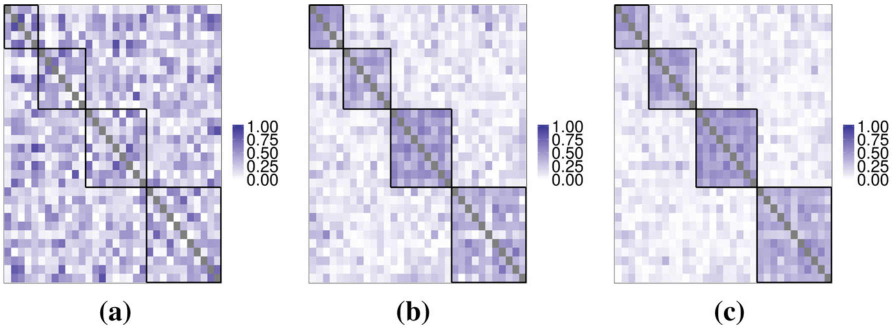


FIGURE 2.
Block diagonal: estimated Spearman's correlation coefficient (absolute values) from a single simulation at sample sizes of $n = 10$, 100, and 1000. **a** $n = 10$. **b** $n = 100$. **c** $n = 1000$.

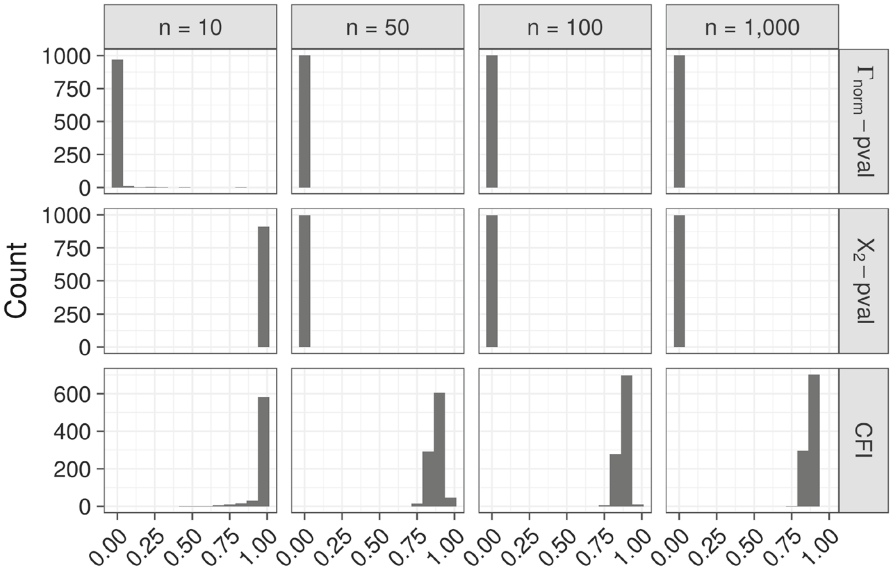


FIGURE 3.
Overall test for block diagonal scenario: permutation p-values with Γ_{norm} and $B = 10{,}000$ MC resamples, p values from the X_2 pattern hypothesis test, and CFI values from a CFA model. For each sample size, we did 1000 simulations. Results with TLI are similar to those for CFI and are not shown.

Figure 3 shows the distribution of p values from a permutation test with Γ_{norm} and $B = 10{,}000$ resamples, p values from the X_2 pattern hypothesis test, and CFI values from a CFA model. Figure 4 shows the distribution of RMSEA values. As seen in Figs. 3 and 4, the distribution of p values from Γ_{norm} is heavily left-skewed, which is as expected under the alternative hypothesis. The p values from the X_2 statistic quickly move from close to 1 to close to 0 as the sample size increases, and the CFI values cluster around 0.8 to 0.9 for all sample sizes. However, as shown in Table 3, the distribution of CFI values shifts downward as sample size increases, though not as dramatically as for p values from X_2. The RMSEA values tend to cluster around 0 for the sample size of $n = 10$, but have a central tendency around 0.13 to 0.16 for the larger sample sizes.

Table 2 shows the power with Γ_{norm} and the permutation test under the alternative hypothesis of block diagonal structure for statistical significance levels of $\alpha = 0.01$ and 0.05. As seen in Table 2, the statistical power was 1 for all tests with sample sizes of 100 and 1000.

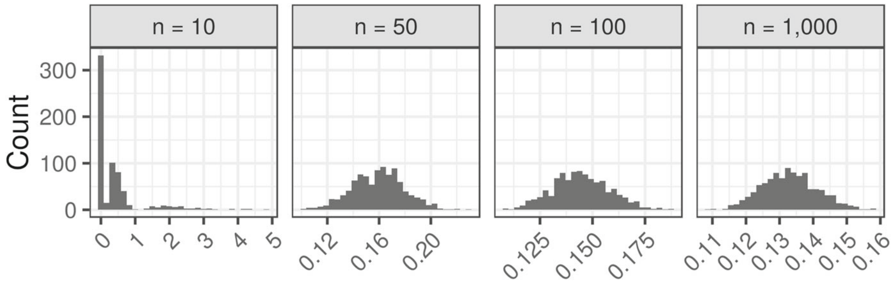


FIGURE 4.
Overall test for block diagonal scenario: RMSEA. For each sample size, we did 1000 simulations.

TABLE 2.
Statistical power in block diagonal scenario using Γ_{norm} in a permutation test for significance levels of $\alpha = 0.01$ and 0.05. 1000 simulations were run for each sample size.

		Block			
n	Overall	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$\alpha = 0.01$					
10	0.97	0.31	0.34	0.69	0.34
50	1.0	0.94	0.97	1.0	0.98
100	1.0	0.99	0.99	1.0	1.0
1000	1.0	1.0	1.0	1.0	1.0
$\alpha = 0.05$					
10	0.98	0.47	0.49	0.81	0.51
50	1.0	0.97	0.99	1.0	1.0
100	1.0	1.0	1.0	1.0	1.0
1000	1.0	1.0	1.0	1.0	1.0

Table 3 shows the percent of simulations with CFI and TLI above the cutoff value recommended by Hu and Bentler (1999) (0.95) as well as more liberal cutoff values noted by Hooper et al. (2008) (0.9 and 0.8). As can be seen in Table 3, the statistical power of TLI and CFI decreases as sample size increases, similar to the X_2 GOF test. Notably, the Type I power is at or near zero for both CFI and TLI for large sample sizes and cutoffs of 0.9 and 0.95.

Table 4 shows the percent of simulations with RMSEA below the cutoff values recommended by Steiger (2007) (0.07), as well as the alternative cutoff values recommended by Browne and Cudeck (1992) (0.05, 0.1). As can be seen in Table 4, the statistical power of RMSEA is low for all sample sizes, and is zero for all cutoffs at $n = 100$ and $n = 1000$.

We note that Steiger (1990) recommends using confidence intervals for RMSEA and concluding that the model fits if the lower bound is near zero and the upper bound is not too far above the cutoff (e.g., 0.07). In this simulation, because the point estimates are mostly above 0.07, the upper limits must also be above 0.07, so considering the confidence intervals would not be likely to change our conclusions.

We also note that the absolute correlation matrices shown in Fig. 2 are visually very similar to the matrices generated under a true CFA model shown in Figure S6 (with continuous outcome) and S8 (with discretized outcome) of "Web Appendix B in Supplementary material". However, in the simulations of this section, CFI, TLI, and RMSEA have low type I power, whereas under the true CFA generating process of "Web Appendix B in Supplementary material", CFI, TLI, and RMSEA have high type I power, even after discretization of the outcome. This indicates the

TABLE 3.
Type I power for block diagonal scenario with CFI and TLI: percent of simulation results above the cutoff value (CFI and TLI above the cutoff indicate good model fit).

		Cutoff		
Fit index	n	0.95	0.9	0.8
CFI	10	0.89	0.92	0.96
	50	0.016	0.30	0.97
	100	0.0	0.25	0.98
	1000	0.0	0.14	0.99
TLI	10	0.89	0.92	0.96
	50	0.0084	0.23	0.94
	100	0.0	0.16	0.96
	1000	0.0	0.073	0.98

TABLE 4.
Type I power for block diagonal scenario with RMSEA: percent of simulation results below the cutoff value (RMSEA below the cutoff indicates good model fit).

		Cutoff		
Fit index	n	0.05	0.07	0.1
RMSEA	10	0.51	0.51	0.51
	50	0.0	0.0	0.0021
	100	0.0	0.0	0.0
	1000	0.0	0.0	0.0

performance of CFI, TLI, and RMSEA may be robust to continuous versus discrete outcomes, but sensitive to other distributional assumptions of the CFA model. These other distributional assumptions are typically not of primary interest in construct validation.

5.2. *White Noise*

For this scenario, we generated the square root of the covariance matrix as $\Sigma_{t,ij}^{1/2} \sim N(0, 1)$ and set the covariance matrix to $\Sigma_t = \left(\Sigma_t^{1/2}\right)^T \Sigma_t^{1/2}$. The rest of the simulation is as described in Sect. 5.1.

Figure 5 shows the estimated Spearman's absolute correlation matrices A from a single simulation for sample sizes of $n = 10$, 100, and 1000.

Figure 6 shows the distribution of p values from a permutation test with Γ_{norm} and $B = 1000$ MC resamples, p values from the X_2 pattern hypothesis test, and CFI values from a CFA model. Figure 7 shows the distribution of RMSEA values. As seen in Figs. 6 and 7, the distribution of p values from Γ_{norm} is uniform, which is as expected under the null hypothesis. The p values from the X_2 statistic move from close to one to close to zero as the sample size increases, though some simulates gave p values close to 1 even for $n = 100$ and 1000. The CFI values cluster close to 1 for $n = 10$ and around 0.25 to 0.5 for $n = 100$ and $n = 1000$. The RMSEA values are near zero for $n = 10$, but center around 0.13 to 0.14 for $n = 100$ and $n = 1000$. In this scenario, the CFA model is misspecified, so small CFI values and large RMSEA values for $n = 100$ and 1000 indicate a low GOF false alarm rate.

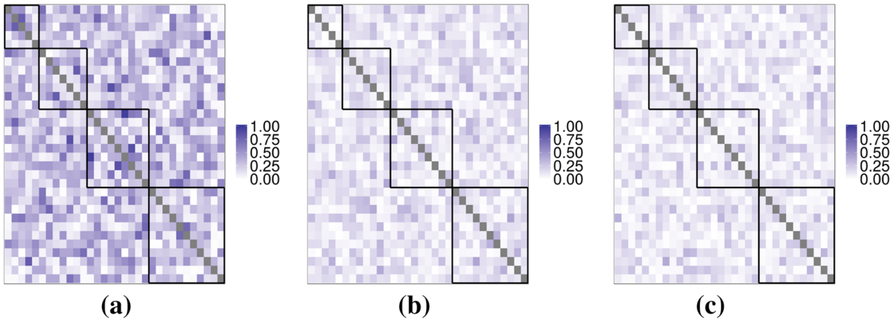


FIGURE 5.
White noise: estimated Spearman's correlation coefficient (absolute values) from a single simulation at sample sizes of $n = 10$, 100, and 1000. **a** $n = 10$. **b** $n = 100$. **c** $n = 1000$.

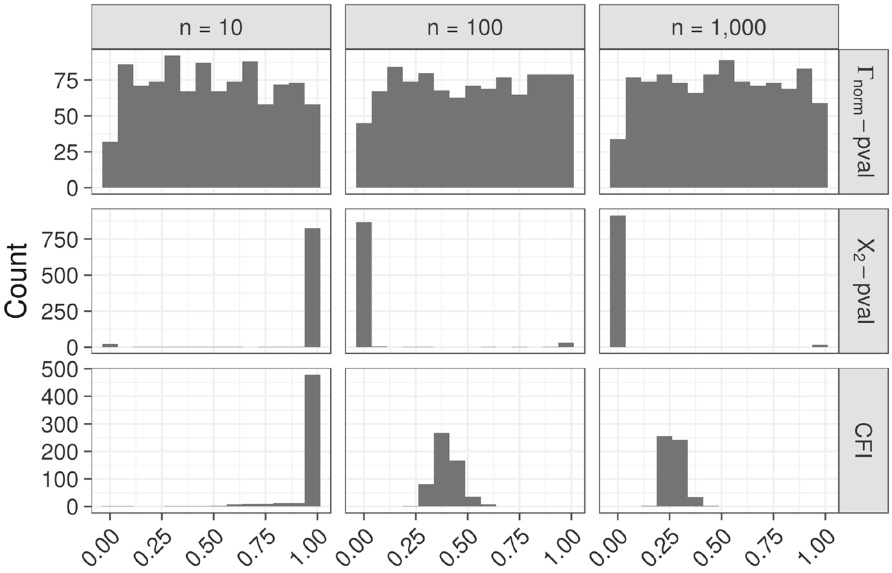


FIGURE 6.
Overall test in white noise scenario: permutation p values using Γ_{norm} and $B = 1000$ MC resamples, p values from the X_2 pattern hypothesis test, and CFI values from a CFA model. For each sample size, we did 1000 simulations. Results with TLI are similar to those for CFI and are not shown.

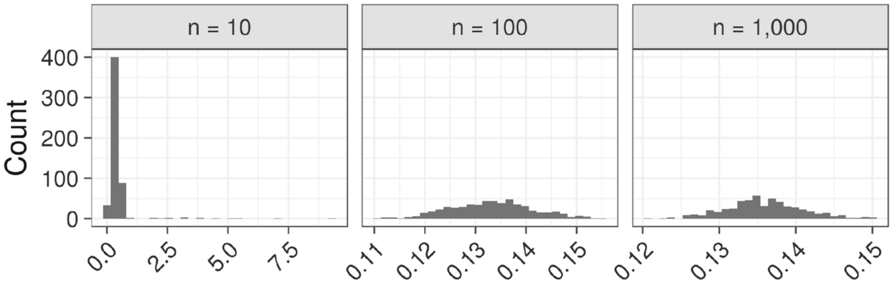


FIGURE 7.
Overall test in white noise scenario: RMSEA. For each sample size, we did 1000 simulations.

TABLE 5.
Type I error rates for white noise scenario using Γ_{norm} in a permutation test for significance levels of $\alpha = 0.01$ and 0.05. 1000 simulations were run for each sample size.

	n	Overall	Block-specific FWER
$\alpha = 0.01$	10	0.006	0.009
	100	0.008	0.007
	1000	0.006	0.015
$\alpha = 0.05$	10	0.046	0.047
	100	0.063	0.047
	1000	0.041	0.042

TABLE 6.
GOF false alarm rate for white noise scenario: Percent of simulation results above the cutoff value (CFI and TLI above the cutoff indicate good model fit).

		Cutoff		
Fit index	n	0.95	0.9	0.8
CFI	10	0.88	0.90	0.93
	100	0.0	0.0	0.0
	1000	0.0	0.0	0.0
TLI	10	0.88	0.90	0.93
	100	0.0	0.0	0.0
	1000	0.0	0.0	0.0

TABLE 7.
GOF false alarm rate for white noise scenario: percent of simulation results below the cutoff value (RMSEA below the cutoff indicates good model fit).

		Cutoff		
Fit index	n	0.05	0.07	0.1
RMSEA	10	0.061	0.061	0.061
	100	0.0	0.0	0.0
	1000	0.0	0.0	0.0

Table 5 shows the type I error rates for Γ_{norm} and the permutation test for statistical significance levels of $\alpha = 0.01$ and 0.05. As seen in Table 5, the error rates are near their nominal level for all sample sizes.

Table 6 shows the percent of simulations with CFI and TLI above the cutoff value recommended by Hu and Bentler (1999) (0.95), as well as the more liberal cutoff values noted by Hooper et al. (2008) (0.9 and 0.8). As seen in Table 6, the GOF false alarm rate is zero for all sample sizes larger than $n = 10$.

Table 7 shows the percent of simulations with RMSEA below the cutoff values recommended by Steiger (2007) (0.07), as well as the alternative cutoff values recommended by Browne and Cudeck (1992) (0.05, 0.1). As can be seen in Table 7, the GOF false alarm rate is zero for all cutoffs at $n = 100$ and $n = 1,000$.

TABLE 8.
Overall and block-specific tests with $B = 10{,}000$ MC resamples for the HRS big five personality traits questionnaire, controlling for family-wise error rate.

Block k	Interpretation	$\Gamma^{0}_{\text{norm},k}$	p-value
–	Overall	0.40	$<$ 0.0001
1	Neuroticism	0.55	0.0002
2	Extroversion	0.37	0.0025
3	Agreeableness	0.49	0.0003
4	Openness to experience	0.50	0.0002
5	Conscientiousness	0.21	0.11

6. Application

In this section, we continue our analysis of the HRS big five personality traits questionnaire described in Sect. 2.

Table 8 shows the results from a permutation test with $B = 10,000$ MC resamples.

As seen in Table 8, the permutation test provides evidence in favor of validating the extroversion, agreeableness, neuroticism, and openness blocks, but not the conscientiousness block. However, based on Fig. 1, the agreeableness, conscientiousness, and neuroticism blocks appear to be highly correlated with each other. In this case, we would recommend further discussions based on content area knowledge to better understand whether these blocks measure distinct underlying constructs in the HRS population. These results could potentially also help to inform future versions of the questionnaire.

6.1. Pattern Hypothesis Test and CFA

The p value from the pattern hypothesis test with X_2 gave a p value of $< 10^{-16}$, providing evidence against validating the construct. However, the large sample size in the HRS study leads to low type I power, making it unlikely that the pattern hypothesis test would provide evidence in favor of validating the big five personality traits.

Using the `lavaan` package (Rosseel 2012) for R (R Core Team 2017), we fit a CFA model with five latent factors (one for each of the five constructs, with each item loading onto its hypothesized factor). This gave a CFI of 0.91, a TLI of 0.90, and a RMSEA of 0.101 (90% confidence interval: 0.101, 0.102). Based on the recommendations of Hu and Bentler (1999) and Hooper et al. (2008), it is unclear whether the CFI and TLI values provide evidence for validating the construct. If we strictly followed the 0.95 cutoff recommended by Hu and Bentler (1999) for CFI and TLI, then neither metric would provide evidence in support of the constructs. Similarly, the RMSEA is larger than the cutoffs recommended by both Steiger (2007) and Browne and Cudeck (1992), so RMSEA also does not provide evidence in support of the constructs. However, as we found with Γ_{norm} and the permutation test, we likely have evidence in support of validating the constructs with the possible exception of the “conscientiousness" block.

7. Discussion

Directly testing hypotheses concerning the structure of A with the methods described in this paper, as opposed to implicitly testing the structure of A through a model-based approach, such as CFA, has both advantages and disadvantages. The tests of matrix structure presented in this paper allow for greater variety of matrices (e.g., A can be correlations or absolute correlations, in addition to covariances), have null hypotheses that are aligned with the scientific question, and make it possible to test each block separately in addition to the overall test. These nonparametric tests also address the challenge in CFA of determining whether poor fit (small GOF index) is due to incorrect assumptions on the distributions of the random variables (secondary interest), or an inaccurate attribution of test questions to latent variables (primary interest). This last benefit might be particularly important, as our simulations show that the performance of CLI, TLI, and RMSEA can be sensitive to the distributional assumptions of CFA (see discussion in Sect. 5.1).

However, CFA, and more generally, SEMs, allow for more flexible latent variable structures and can be used in subsequent analyses to study associations between latent variables and additional covariates. With this in mind, we view methods for directly testing the structure of A as being useful either by themselves when appropriate or to check the robustness of model-based approaches.

The simulation results suggest that the permutation test with Γ_{norm} maintains high power while controlling the type I error rate. In particular, the p values are uniformly distributed under the null hypothesis, so type I error rates can be estimated theoretically. In contrast, CLI, TLI, and RMSEA behave differently depending on the scenario, so it is not possible to theoretically estimate error rates, such as the GOF false alarm rate (type II errors, see Table (1)). This has the consequence that the known behavior of CLI, TLI, and RMSEA are restricted to simulation results and may not generalize to other settings.

In this paper, we focused on scenarios in which each observed variable loads onto no more than one latent factor, which implies a block diagonal structure in the covariance and correlation matrices. This constraint is commonly imposed on CFA models as well. However, the Γ_{norm} statistic and permutation test are not restricted to these scenarios, and in future work it could be beneficial to study the performance of these methods when testing more general matrix structures.

In future work, it may also be beneficial to investigate the use of the studentized difference in means (3) in place of Γ_{norm} in the permutation test. In our simulations, (3) gave nearly identical results as Γ_{norm} (results not shown), but due to the results of Chung and Romano (2013), we speculate that there may be scenarios in which (3) controls the type I error rate better than Γ_{norm}.

Finally, we note that we view these tests as single pieces of information that can be used in a larger decision-making process. This approach is consistent with the American Statistical Association's statement on p values (Wasserstein and Lazar 2016).

8. Supplementary Material

An R package `matrixTest` that implements the methods described in this paper is available at https://github.com/bdsegal/matrixTest, and code for reproducing all results in this paper is available at https://github.com/bdsegal/code-for-matrixTest-paper.

Acknowledgments

We thank Jacqui Smith, Philippa Clarke, and Trivellore Raghunathan for helpful discussion and feedback.

References

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, *42*, 815–824.

Beauducel, A., & Wittmann, W. (2005). Simulation study on fit indices in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, *12*(1), 41–75.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Bock, R. D., & Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika*, *31*(4), 507–534.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258.

Chung, E., & Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, *41*(2), 484–507.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indices to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*(3), 343–367.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.

Good, P. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses* (2nd ed.). New York: Springer.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2/3), 107–145.

Hawkins, D. L. (1989). Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. *The American Statistician*, *43*(4), 235–237.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods*, *6*(1), 53–60.

HRS (2016). Health and retirement study (core data release) public use dataset.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Hubert, L., & Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, *29*(2), 190–241.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs: Prentice Hall.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, *43*(4), 443–477.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York: Springer.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*(2), 209–220.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentlers (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341.

McDonald, R. P. (1974). Testing pattern hypotheses for covariance matrices. *Psychometrika*, *39*(2), 189–201.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Srivastava, J. N. (1966). On testing hypotheses regarding a class of covariance structures. *Psychometrika*, *31*(2), 147–164.

Steiger, J. H. (1980a). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioral Research*, *15*(3), 335–352.

Steiger, J. H. (1980b). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(5), 893–898.

Steiger, J. H., & Lind, J. (1980). Statistically-based tests for the number of common factors. Iowa City: Paper presented at the annual spring meeting of the Psychometric Society.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p values: Context, process, and purpose. *The American Statistician*, *70*(2), 193–242.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). New York, NY: Wiley.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*(1), 115–148.

Zaki, M. J., & Meira, W, Jr. (2014). *Data mining and analysis: Fundamental concepts and algorithms*. New York, NY: Cambridge University Press.

Manuscript Received: 12 MAY 2017
Published Online Date: 27 NOV 2018