Psychometrika (2025), **90**, 384–414
doi:10.1017/psy.2024.21

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

An Overall Test of Pairwise Mean Conditional Covariances in IRT

Jules L. Ellis[1], L. Andries van der Ark[2] and Klaas Sijtsma[3]

[1]Faculty of Psychology, Open Universiteit, Heerlen, Netherlands; [2]Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, Netherlands; [3]Department of Methodology and Statistics TSB, Tilburg University, Tilburg

Corresponding author: Jules L. Ellis; Email: jules.ellis@ou.nl

(Received 15 October 2024; accepted 17 October 2024; published online 3 January 2025)

The authors have no financial or proprietary interests in any material discussed in this article. No funds, grants, or other support were received.

Abstract

We study how the Conditioning on Added Regression Predictions (CARP) statistics from different item pairs can be aggregated into a single overall test of monotone homogeneity. As a pairwise statistic, we use the mean conditional covariance (MCC) or its standardized value (Z). We use three different estimates of the covariance matrix of the pairwise test statistics: (1) the covariance matrix of the MCCs, based on the sample moments; (2) the covariance matrix of the MCCs or Zs, based on bootstrapping; and (3) the covariance matrix of the Zs, equated to the identity matrix. We consider various aggregation methods, including (a) the chi-bar-square statistic; (b) the preselected standardized partial sum of pairwise statistics; (c) the product of preselected p-values; (d) the minimum of preselected p-values; and (e–h) the same statistics, but now conditioned on post-selecting only the negative values in the test sample. We study the Type 1 error rate and power of the ensuing 20 tests based on simulations. The tests with the highest power among the tests that control the Type I error rate are based on Z-statistics with the identity matrix: the conditional likelihood ratio test, the conditionalized product of p-values, the conditionalized sum of Z-values, and the preselected product of p-values.

Keywords: conditional association; monotone homogeneity model; monotone latent variable model; multidimensional measurement; unidimensional measurement

1. Introduction

In this paper, we develop new statistics for confirmatory tests of unidimensionality based on the nonparametric item response theory (IRT) model of monotone homogeneity (MH) (Mokken, 1971) with binary items. This model assumes that there is a unidimensional (i.e., real-valued) variable Θ such that the items are conditionally independent given Θ, and such that the item regressions on Θ are monotone increasing. Many parametric IRT models, such as the 2PL model and the Rasch model, are a special case of MH. We develop our statistical tests for the context where researchers have the theory or hypothesis that items of a certain specified set or category all have a monotone regression on the same latent variable, while the specific shape of the regressions is unspecified; that is, it does not have to be logistic, as in the 2PL or the Rasch model, or any other function, such as the normal ogive.

We assume that the objective of the test is to falsify the theory of unidimensionality. Thus, the objective is entirely aimed at fundamental theory development and not at the more pragmatic goal of building an efficient measurement tool.

Our objective of fundamental theory development rules out the possibility of assessing dimensionality with flexible parametric models such as latent class models (Douglas & Cohen, 2001; Van Onna, 2002; Vermunt, 2001) or monotone polynomial models (Falk & Cai, 2016). If such a parametric model is violated, it does not provide a convincing falsification of the theory because the violation may be caused by failure of the specific parametric assumptions. Conversely, if such a restrictive parametric model is not rejected, one may wonder whether maybe the statistical test used was not sensitive to some assumptions (e.g., van den Wollenberg, 1982).

The new test statistics will be based on the recently developed Conditioning on Added Regression Predictions (CARP) statistics of Ellis and Sijtsma (2023). The CARP testing approach is a generalization of Rosenbaum's (1984) case 5, in which one tests nonnegativity of the covariance of an item pair, conditionally on decile groups defined by the sum score on the other items. The generalization Ellis and Sijtsma proposed using a weighted sum score instead of a simple sum score, where the weights are based on regression analysis in a training sample. Ellis and Sijtsma (2023) argued that this is currently the only known partial test of conditional association (see next section) that can detect multidimensionality within monotone IRT models. This is the reason why we focus on this test statistic.

An important limitation of the CARP test is that it pertains to a single item pair. Generally, a test has many item pairs, and it would seem logical to apply a test to each item pair, but hitherto it has not been studied how such pairwise tests can be compounded into a single test statistic. The same is true for Rosenbaum's case 5 test. For example, if a psychological test consists of 10 items, the CARP tests would yield 45 p-values, one for each item pair. The main question of this article is: How can the pairwise CARP statistics be aggregated into a single omnibus test?

The next section provides some background information about the CARP tests. After the specification of the hypothesis and the relevant pairwise statistics (mean conditional covariances (MCCs) and their Z-values), we consider various methods to estimate the covariance matrix of the pairwise statistics. These estimated matrices are used in the theory of order restricted statistical inference (Robertson et al., 1988) and multiple testing (Davidov, 2011; Ellis et al., 2018) to compound them into overall statistics in 20 different ways. We will compare the mathematical structure of some of these aggregated statistics to the most prominent competitor, which is the DETECT index (Zhang & Stout, 1999a, 1999b). Next, we use Monte Carlo simulations to study the Type 1 error rates and power of the ensuing tests and select the best tests.

2. Conditional association and CARP tests

Rosenbaum (1984) showed that MH implies that the item score variables have the property of conditional association, which means that any two increasing functions of any subtest have a nonnegative covariance conditionally upon any function of the items that were not included in the subtest. Holland and Rosenbaum (1986) generalized this result to non-binary items. Clarke and Yuan (2001) and De Gooijer and Yuan (2011) developed statistical tests for conditional association, but it is well known that a full test of conditional association is not feasible for realistic sizes of item sets because of the large number of tested conditions and the sparseness of the relevant response patterns. Several authors have therefore focussed on what Ligtvoet (2022) recently called "partial tests of conditional association." These tests include MTP2 (Bartolucci & Forcina, 2000; Ellis, 2015), nonnegative partial correlations (Ellis, 2014), nonnegative correlations (Mokken, 1971), and increasing item-rest regressions ("manifest monotonicity"; Junker & Sijtsma, 2000). Ellis and Sijtsma (2023) showed that all these conditions are insensitive to violations of unidimensionality if the data are generated by multiple latent variables that are independent or MTP2. They developed the CARP test,

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

which includes Rosenbaum's case 5 as a special case. This is currently the only known partial test of conditional association that can detect such violations. That is the reason why we focus on the CARP test.

3. Definitions

3.1. Definitions of variables

Assume that the item scores are binary manifest variables. Let variable X_i represent the scores (1 = positive, 0 = negative) a random subject obtained on the i-th item, and denote the full vector of item scores as $\boldsymbol{X} = (X_1, \ldots, X_J)$. For each item pair (i,j), we assume that there is a discrete variable R_{ij} that is used for creating groups in which the conditional covariances are computed. We call the R_{ij}s the conditioning variables, and we assume that they attain integer values ranging from 1 to $\max R_{ij}$. For example, in Case 2 of Rosenbaum (1984), R_{ij} is defined as the sum score on the other items; that is, the items X_k with $k \neq i,j$. Then, we have $R_{ij} = \left(\sum_{k=1}^{J} X_k\right) - X_i - X_j$; we call this the *pairwise rest score*. In Case 5 of Rosenbaum (1984), R_{ij} consists of deciles of the pairwise rest score. In the CARP test of Ellis and Sijtsma (2023), R_{ij} consists of deciles of a weighted sum score on the other items, with weights estimated from a training sample. Let $\boldsymbol{R}$ be the vector of all R_{ij} $(i,j = 1, \ldots, J; i \neq j)$.

3.2. The hypothesis

The null hypothesis of interest is

$$\mathrm{H}_0 : Cov\left(X_i, X_j | R_{ij} = s\right) \geq 0 \text{ for all } i,j = 1, \ldots, J; s = 1, \ldots, \max R_{ij}.$$

However, the version of the Mantel–Haenszel statistic Rosenbaum (1984) and Ellis and Sijtsma (2023) used for testing H_0 is rather based on a weighted mean of sample covariances, and it would be more precise to say that the null hypothesis is

$$\mathrm{H}_0 : \sum_{s=1}^{\max R_{ij}} P\left(R_{ij} = s\right) Cov\left(X_i, X_j | R_{ij} = s\right) \geq 0 \text{ for all } i,j = 1, \ldots, J.$$

3.3. Definition of the pairwise statistics

In this section, we define two sample statistics per item pair (i,j) that we aggregate later. First, we define the mean of conditional covariances that estimates the quantity $P\left(R_{ij} = s\right) Cov\left(X_i, X_j | R_{ij} = s\right)$ in the null hypothesis. Second, we define the Z-statistic, which is the standardized version of the first statistic. The formal definition is the following.

Assume that there are N i.i.d. copies of $\boldsymbol{X}$, denoted by $\boldsymbol{X}^{(n)} = \left(X_1^{(n)}, \ldots, X_J^{(n)}\right)$; $n = 1,2, \ldots, N$. $\boldsymbol{X}^{(n)}$ contains the scores of the nth subject in the sample. Let $I_{ijs}^{(n)} := 1\left[R_{ij}^{(n)} = s\right]$ denote the indicator function for the event $R_{ij} = s$ in subject n. That is, $1\left[R_{ij}^{(n)} = s\right] = 1$ if $R_{ij}^{(n)} = s$, and $1\left[R_{ij}^{(n)} = s\right] = 0$ otherwise. Let $N_{ijs} = \sum_{n=1}^{N} I_{ijs}^{(n)}$ denote the number of subjects with $R_{ij}^{(n)} = s$. The conditional covariance in the subgroup with $R_{ij}^{(n)} = s$ is given by

$$C_{ijs} = \frac{\sum_{n=1}^{N} X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}} - \frac{\sum_{n=1}^{N} X_i^{(n)} I_{ijs}^{(n)}}{N_{ijs}} \frac{\sum_{n=1}^{N} X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}}.$$

The version of the Mantel–Haenszel statistic Rosenbaum (1984) used is based on standardization of

$$C_{ij+} = \sum_{s=1}^{S} N_{ijs} C_{ijs}.$$

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

We refer to C_{ij+} as the MCC. The standardization Rosenbaum used is based on the variance estimate

$$V_{ij} = \sum_{s=1}^{S} \frac{\sum_{n=1}^{N} X_i^{(n)} I_{ijs}^{(n)} \sum_{n=1}^{N} \left(1 - X_i^{(n)}\right) I_{ijs}^{(n)} \sum_{n=1}^{N} X_j^{(n)} I_{ijs}^{(n)} \sum_{n=1}^{N} \left(1 - X_j^{(n)}\right) I_{ijs}^{(n)}}{N_{ijs}^2 \left(N_{ijs} - 1\right)}.$$

The Z-statistic is then defined as

$$Z_{ij} = \frac{C_{ij+} + 0.5}{\sqrt{V_{ij}}}.$$

The term 0.5 is a continuity correction.

4. Estimation of the covariance matrix of the pairwise test statistics

In this section, we discuss the estimation of the covariance matrix of the C_{ij+}s and the covariance matrix of the Z_{ij}s. These covariance matrices are conceptually like the asymptotic covariance matrices in structural equation modelling (SEM) because they estimate the covariance across all possible samples. However, because the asymptotic covariance matrices in SEM are typically derived from the model and typically pertain to model estimates rather than conditional covariances, we further refrain from focusing on the apparent similarity. Next, we delineate three estimation methods.

4.1. Estimation based on sample moments

The equation for MCC contains only sums of products and products of sums, divided by the N_{ijs}. We worked out a formula for $Cov\left(C_{ijs}, C_{klt}\right)$, assuming that N_{ijs} and N_{klt} are fixed values rather than random variables (see Appendix). This new equation uses only moments of the variables $X_i^{(n)} I_{ijs}^{(n)}$, $X_j^{(n)} I_{ijs}^{(n)}$, $X_k^{(n)} I_{klt}^{(n)}$, $X_l^{(n)} I_{klt}^{(n)}$, and their products. By substituting the corresponding sample moments, one obtains an estimate $\widehat{Cov}\left(C_{ijs}, C_{klt}\right)$ for $Cov\left(C_{ijs}, C_{klt}\right)$. Next, the required covariance is estimated as $\widehat{Cov}\left(C_{ij+}, C_{kl+}\right) := \sum_s \sum_t N_{ijs} N_{klt} \widehat{Cov}\left(C_{ijs}, C_{klt}\right)$.

It should be noted that in the IRT application of testing the MH model, the N_{ijs} are not fixed. By doing as if the N_{ijs} are fixed anyway, we ignore the possibility of a correlation between C_{ijs} and N_{ijs}. This might still entail a better approximation than assuming the identity matrix, and we will use the simulation studies to decide whether this approximation is useful.

4.2. Estimation based on bootstrapping

4.2.1. Bootstrapping of the MCCs

In this approach, we resample N rows of the data matrix with replacement and compute the MCC for each item pair (i,j) in the resample. Denote the MCC of item pair (i,j) in a resample as C_{ij+}^*. We resample $n_{\text{resample}} = 1000$ times, thus constructing a matrix of n_{resample} rows and $J\left(J-1\right)/2$ columns, in which each row m contains the C_{ij+}^* values of the m-th resample. Next, we compute the covariance matrix of the C_{ij+}^*s.

4.2.2. Bootstrapping of the Zs

In this approach, we resample N rows of the data matrix with replacement and compute the Z-statistic for each item pair (i,j) in the resample. Denote the Z-value of item pair (i,j) in a resample as Z_{ij}^*. We resample $n_{\text{resample}} = 1000$ times, thus constructing a matrix of n_{resample} rows and $J\left(J-1\right)/2$ columns, in which each row m contains the Z_{ij}^* values of the m-th resample. Next, we compute the covariance matrix of the Z_{ij}^*s.

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

4.3. Estimation with the identity matrix

In this approach, the third method, we simply assume that the asymptotic covariance matrix of the Z_{ij}s is the identity matrix. This is comparable to the diagonally weighted least squares (DWLS) method often used for polychoric correlations in ordinal factor analysis (Li, 2015). The reason why we suspect that the identity matrix may work well is that under MH, each conditioning group with a fixed pairwise rest score has only a small remaining variance of the latent variable, which implies that the response variables are close to independent, and independent Bernoulli variables have covariances that are asymptotically uncorrelated (Anderson & Goodman, 1957).

5. Aggregation of the pairwise statistics

In this section, we consider the lower-diagonal matrix of C_{ij+}s or Z_{ij}s as a vector $\boldsymbol{y}$ in $\mathbb{R}^{J(J-1)/2}$. Let $\boldsymbol{W}$ be the covariance matrix of $\boldsymbol{y}$ as estimated in any method of the previous section. We consider various methods to aggregate $\boldsymbol{y}$ into an omnibus test.

5.1. Distance to the nonnegative cone

The theory in this section is based on order restricted statistical inference (Robertson et al., 1988). We project $\boldsymbol{y}$ onto the nonnegative cone in $\mathbb{R}^{J(J-1)/2}$, defined by $\mathcal{C} = \left\{ \boldsymbol{x} \in \mathbb{R}^{\frac{J(J-1)}{2}} | x_1 \geq 0, x_2 \geq 0, \ldots, x_{J(J-1)/2} \geq 0 \right\}$. We define the projection $\boldsymbol{y}^*$ as the point in $\mathcal{C}$ that minimizes the squared Mahalanobis distance $\left(\boldsymbol{y}-\boldsymbol{y}^*\right)^T \boldsymbol{W}^{-1} \left(\boldsymbol{y}-\boldsymbol{y}^*\right)$. The chi-bar-square statistic is then defined as this squared distance, i.e.

$$\bar{\chi}^2 := \left(\boldsymbol{y}-\boldsymbol{y}^*\right)^T \boldsymbol{W}^{-1} \left(\boldsymbol{y}-\boldsymbol{y}^*\right).$$

This result serves as the overall test statistic in the decision rules discussed in the next section. The following algorithm was used to obtain $\boldsymbol{y}^*$ and $\bar{\chi}^2$:

1. Obtain the Cholesky decomposition $\boldsymbol{W}^{-1} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$, and let $\boldsymbol{A} = \boldsymbol{B}^{-1}$.
2. Compute $\boldsymbol{z} = \boldsymbol{B}\boldsymbol{y}$. If $\boldsymbol{W}$ is the covariance matrix of $\boldsymbol{y}$, then the covariance matrix of $\boldsymbol{z}$ is the identity matrix.
3. Project $\boldsymbol{z}$ onto the cone $\boldsymbol{A}\boldsymbol{x} \geq 0$ with the function coneA of the R-package **coneproj** (Liao & Meyer, 2014). Let $\boldsymbol{z}^*$ be the result of this projection; then $\boldsymbol{y}^* = \boldsymbol{A}\boldsymbol{z}^*$.
4. Compute $\bar{\chi}^2 = \left\| \boldsymbol{z} - \boldsymbol{z}^* \right\|^2$, where $\|\cdot\|$ is the Euclidian distance.

5.2. Preselected standardized partial sum of pairwise statistics

In their CARP statistics, Ellis and Sijtsma (2023) split the total sample of subjects into a training sample and a test sample. In our treatment hitherto, all statistics were computed with the test sample. However, one can compute C_{ij+} in the training sample as well; denote this as $C_{ij+}^{\text{training}}$. Let T_- be the set of pairs (i,j) with $C_{ij+}^{\text{training}} < 0$ and $i > j$, and denote its size as $|\, T_- \,|$. We add the corresponding values of $\boldsymbol{y}$ and divide the sum by its standard error. More specifically, let $y_{\frac{(i-1)(i-2)}{2}+j} = C_{ij+}$ for all $J \geq i > j \geq 1$, and define $\boldsymbol{y}_-$ as the subvector of $\boldsymbol{y}$ containing the elements $y_{\frac{(i-1)(i-2)}{2}+j}$ for which $(i,j) \in T_-$. Let $\boldsymbol{W}_-$ be the submatrix of $\boldsymbol{W}$, corresponding to the elements of $\boldsymbol{y}_-$; Ellis and Sijtsma (2023) defined the omnibus statistic as

$$Z_{\text{PS}} := \frac{\sum \boldsymbol{y}_-}{\sqrt{\sum \boldsymbol{W}_-}},$$

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

where $\sum$ indicates the sum over all elements in the following vector or matrix. Here, we call this omnibus test statistic the preselected standardized partial sum of pairwise statistics. For clarity, if $\boldsymbol{y}$ contains the Z_{ij}s and $\boldsymbol{W}$ is assumed to be the identity matrix, then

$$Z_{\text{PS}} = \frac{\sum \{Z_{ij} | (i,j) \in T_-\}}{\sqrt{|T_-|}}.$$

5.3. Conditionalized multiple testing procedures

Ellis et al. (2018) argued that multiple testing procedures for interval hypotheses can be enhanced with the following general adaptation: Pick some real number $\lambda \in (0,1]$, and select only the p-values with $p_i < \lambda$, and divide each p_i by λ; then, apply an ordinary multiple testing procedure, such as the Bonferroni correction or the Benjamini-Hochberg correction, to the resulting set. That is, apply an ordinary multiple testing procedure to the set of corrected p-values $\{p_i/\lambda : p_i < \lambda\}$. Ellis et al. argue that this procedure often controls the Type 1 error, certainly with independent p-values, and also with non-independent p-values if the number of p-values is large and the Bonferroni-correction is used. They show that this procedure increases the power if the p-values are supra-uniform; that is, if most p-values are higher than would be expected in a uniform distribution. We expect that the latter condition is often fulfilled in the situation under investigation here, where the p-values test whether conditional covariances are nonnegative in the context of MH. If MH holds and the item response functions are not flat, then the conditional covariances will be positive, yielding supra-uniform p-values.

Ellis et al. (2018) investigated their conditionalization procedure with the multiple testing methods Davidov (2011) discussed in the context of independent p-values, and we will apply some of these multiple testing methods here. Davidov recommended using the I_+ statistics, but the method that he labelled "normal" achieved similar power (Davidov, p. 2439–2440). The latter method means that the p-values are converted to standard normal Z-statistics and added, and this method is the natural candidate if the test statistics underlying the p-values are normal. Applied to the present situation with the conditionalization rule of Ellis et al. (2018) and $\lambda = 0.5$, this amounts to the following. Let S be the set of pairs $i > j$ with $Z_{ij} < 0$ in the test sample, and denote its size as $|S|$. A test statistic based on the conditionalized sum is

$$Z_{\text{CS}} := \frac{\sum \left\{\Phi^{-1}(2\Phi(Z_{ij})) | (i,j) \in S\right\}}{\sqrt{|S|}},$$

where S is the number of pairs $i < j$ with $Z_{ij} < 0$. However, the statistic that was most powerful in the simulations, Ellis et al. reported was the product of p-values, which Davidov attributed to Fisher. In the present situation, after a log-transformation of the conditionalized product, we obtain,

$$Q_{\text{CP}} := -2 \sum \{\log(2\Phi(Z_{ij})) | (i,j) \in S\},$$

which is compared to a chi-square distribution with $df = 2|S|$.

Finally, we consider the Bonferroni correction because it is so easy and well-known despite the general consensus that more powerful alternatives exist. In conditionalized form, this amounts to

$$p_{\text{CB}} = \min \{|S| \Phi(Z_{ij}) | (i,j) \in S\}.$$

Note that all three statistics, Z_{CS}, Q_{CP}, and p_{CB}, ignore the correlations of the Z_{ij}s. In Z_{CS} and Q_{CP}, it is implicitly assumed that the Z_{ij}s are uncorrelated, and we are not certain that they control the Type 1 error in correlated cases. However, the correlations between pairwise CARP statistics might be so small that it hardly affect the distribution, as we argued in the section where we proposed the identity matrix. Therefore, we study these statistics despite the uncertainty about possibly correlated Z_{ij}s.

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

5.4. Preselection with multiple testing procedures

The idea of a preselection of item pairs based on the training sample can also be applied to multiple testing procedures, like the conditionalization principle Ellis et al. (2018) discussed. Suppose the data sample is randomly split into a training sample and a test sample. Let the training data matrix be denoted as $\boldsymbol{D}_1$ and the test data matrix as $\boldsymbol{D}_2$, where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are independent. Suppose that for each subset L of pairs of variables in $\boldsymbol{D}_2$ there is a multiple testing procedure M_L that controls the Type I error when it is applied to L. More precisely, M_L is a function that is applied to $\boldsymbol{D}_2$ and only uses the pairs of variables in L, resulting in a 1 (reject) or 0 (no reject) decision, with $P\left(M_L\left(\boldsymbol{D}_2\right)=1\right)\leq\alpha$ if the null hypothesis is true, where α is the nominal level of significance. Suppose that we chose L as a function of $\boldsymbol{D}_1$; let this function be denoted as Λ with range $R\left(\Lambda\right)$. An example of this would be that Λ selects the pairs with negative conditional covariances and M_L uses the p-values $p_{ij}=\Phi\left(Z_{ij}\right)$ of $(i,j)\in L$ with the Bonferroni correction. That is, $\Lambda\left(\boldsymbol{D}_1\right)=\left\{(i,j):C_{ij+}^{\text{training}}<0\text{ and }i>j\right\}$ and $M_L=1\iff\min\left\{p_{ij}|L|\right\}\leq\alpha$. Then $M_{\Lambda\left(\boldsymbol{D}_1\right)}\left(\boldsymbol{D}_2\right)$ is the procedure that applies the Bonferroni correction to p-values in the test sample using only pairs that have a negative conditional covariance in the training sample. Such procedures control the Type I error rate: Since $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are independent, the conditional distribution of $\boldsymbol{D}_2$ given $\Lambda\left(\boldsymbol{D}_1\right)$ is the same as the unconditional distribution of $\boldsymbol{D}_2$, so the rejection rate is

$$P\left(M_{\Lambda\left(\boldsymbol{D}_1\right)}\left(\boldsymbol{D}_2\right)=1\right)=\mathbb{E}\left(P\left(M_{\Lambda\left(\boldsymbol{D}_1\right)}\left(\boldsymbol{D}_2\right)=1|\Lambda\left(\boldsymbol{D}_1\right)\right)\right)$$

$$=\sum_{L\in R(\Lambda)}P\left(M_L\left(\boldsymbol{D}_2\right)=1\right)P\left(\Lambda\left(\boldsymbol{D}_1\right)=L\right)\leq\sum_{L\in R(\Lambda)}\alpha P\left(\Lambda\left(\boldsymbol{D}_1\right)=L\right)=\alpha.$$

Thus, we may calculate p-values for conditional covariances in the test data using only the pairs that have a negative conditional covariance in the training sample and then apply a multiple testing procedure to this selection as if it were the entire test data set from the outset. Applying this procedure to the statistics of the previous section, we obtain the following results. Let T_- be the set of pairs (i,j) with $C_{ij+}^{\text{training}}<0$ and $i>j$. Then

$$Q_{\text{PP}}:=-2\sum\left\{\log\left(\Phi\left(Z_{ij}\right)\right)|\left(i,j\right)\in T_-\right\},$$

with $df=2\mid T_-\mid,\mid$ where $\mid T_-\mid$ is the size of T_-. Similarly,

$$p_{\text{PB}}:=\min\left\{|T_-|\Phi\left(Z_{ij}\right)|\left(i,j\right)\in T_-\right\}.$$

6. Decision rules

6.1. Decision rules: the LR test and the conditional LR test

In this section, we consider two decision rules based on $\overline{\chi}^2$. The first decision rule uses the unconditional distribution of $\overline{\chi}^2$. The second decision rule uses the conditional distribution of $\overline{\chi}^2$, given the dimensionality of the boundary hyperplane that contains $\boldsymbol{y}^*$. If $\boldsymbol{W}$ is the identity matrix, then this dimensionality is equal to the number of negative MCCs. Both the C_{ij+}s and the Z_{ij}s have an asymptotic multivariate normal distribution as $N\to\infty$ (Browne, 1984, proposition 2). Therefore, we assume a multivariate normal distribution for $\boldsymbol{y}$, which is either the vector of C_{ij+}s or the vector of Z_{ij}s.

First, we consider the likelihood ratio (LR) test. Using this test, we reject the null hypothesis if $\overline{\chi}^2$ exceeds a critical level (Robertson et al., 1988). Under the least favorable case of the null hypothesis, where $Cov\left(X_i,X_j|R_{ij}=s\right)=0$ for all $i,j=1,\ldots,J;s=1,\ldots,\max R_{ij}$, the distribution of $\overline{\chi}^2$ is a weighted average of chi-square distributions (Robertson et al., 1988):

$$P\left(\overline{\chi}^2>c\right)=\sum_{r=1}^{\frac{J(J-2)}{2}}P\left(\chi_r^2>c\right)P\left(df\left(\boldsymbol{y}^*\right)=r\right).$$

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

We estimated the weights $P\left(df\left(\boldsymbol{y}^*\right)=r\right)$ by drawing from a multivariate normal distribution with a covariance matrix $\boldsymbol{W}$, and counting for each draw how many coordinates are negative. We used 10^5 draws. Denote the estimated probability of getting r negative coordinates as $\tilde{p}_r$. The p-value for the observed chi-bar-squared is obtained as

$$p_{\text{LR}} := \sum_{r=1}^{\frac{J(J-2)}{2}} P\left(\chi_r^2 > \bar{\chi}_{obs}^2\right)\tilde{p}_r.$$

The null hypothesis is rejected if $p_{\text{LR}} < \alpha$.

Second, we consider the conditional test based on Wollan and Dykstra (1986). Ellis et al. (2018) generalized the conditionalization principle to other multiple testing procedures with one-sided hypotheses and demonstrated that conditionalization achieves a strong gain in power if most null hypotheses are true, a situation that can be expected here. We also discuss other conditional tests; therefore, we call this test the conditional likelihood (CL) ratio test.

Let $dh\left(\boldsymbol{y}^*\right)$ be the dimensionality of the boundary hyperplane on which $\boldsymbol{y}$ is projected, and let $df\left(\boldsymbol{y}^*\right) = \frac{J(J-1)}{2} - dh\left(\boldsymbol{y}^*\right)$. Wollan and Dykstra explain that the conditional distribution of $\bar{\chi}^2$ given $df\left(\boldsymbol{y}^*\right) = r$ is a chi-square distribution with r degrees of freedom if $r > 0$. Let $\chi_r^2(\alpha)$ be the right-sided critical value for nominal significance level α in a chi-square distribution with r degrees of freedom; that is, $P\left(\chi_r^2 > \chi_r^2(\alpha)\right) = \alpha$. In the conditional test, we reject the null hypothesis if both $df\left(\boldsymbol{y}^*\right) > 0$ and $\bar{\chi}^2 > \chi_{df(\boldsymbol{y}^*)}^2(\alpha)$. Assuming a multivariate normal distribution, the Type 1 error rate of the conditional test is less than α, because, as pointed out by Wollan and Dykstra,

$$\begin{aligned} P(\text{reject H}_0) &= \sum_{r=1}^{\frac{J(J-2)}{2}} P\left(\bar{\chi}^2 > \chi_r^2(\alpha) \,|\, df\left(\boldsymbol{y}^*\right) = r\right) P\left(df\left(\boldsymbol{y}^*\right) = r\right) \\ &= \sum_{r=1}^{\frac{J(J-2)}{2}} \alpha P\left(df\left(\boldsymbol{y}^*\right) = r\right) = \alpha\left(1 - P\left(df\left(\boldsymbol{y}^*\right) = 0\right)\right). \end{aligned}$$

The event $df\left(\boldsymbol{y}^*\right) = 0$ corresponds to $\boldsymbol{y}^* = \boldsymbol{y}$, which would happen if all C_{ij+} or Z_{ij} are nonnegative. Wollan and Dykstra continue to estimate this factor, but this probability is small for five items or more, and therefore it is ignored here, consistent with Ellis et al. (2018). In sum, we define

$$p_{\text{CL}} := P\left(\chi_{df(\boldsymbol{y}^*)}^2 > \bar{\chi}_{obs}^2\right) \quad \text{if } df\left(\boldsymbol{y}^*\right) > 0$$

$$p_{\text{CL}} := 1 \quad \text{if } df\left(\boldsymbol{y}^*\right) = 0$$

6.2. Decision rules: other tests

The p-values of the other tests are computed using

$$p_{\text{PS}} := \Phi\left(Z_{\text{PS}}\right);$$

$$p_{\text{CS}} := \Phi\left(Z_{\text{CS}}\right);$$

$$p_{\text{PP}} := G_{2|T_-|}^2\left(Q_{\text{PP}}\right);$$

$$p_{\text{CP}} := G_{2|S|}^2\left(Q_{\text{CP}}\right),$$

where Φ is the standard normal cumulative distribution function and G_k^2 is the chi-square survival function with k degrees of freedom. The corrected p-values p_{CB} and p_{PB} are used without further correction.

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

7. Comparison with competing methods

We compare our statistics with the two prominent alternative methods for the proposed test, which are the DIMTEST or DETECT procedures for analysis of essential dimensionality. DIMTEST and DETECT are two procedures based on Stout's (1987) theory of essential dimensionality (also Stout et al. 1996). DIMTEST has a confirmative approach that tests unidimensionality, whereas DETECT was created as an explorative approach that divides the set of test items into clusters that are associated with different dimensions. Li et al. (2017, p. 210) summarize DIMTEST as follows:

> The DIMTEST procedure (Nandakumar, Yu, Li, & Stout, 1998; Stout, 1987; Stout, Froelich, & Gao, 2001) is often used to test the null hypothesis that an exam is locally independent and unidimensional. It does this by dividing the test into two subtests (an assessment subtest called AT and a partitioning subtest called PT) and testing whether there are any local dependencies among the AT items, conditioned on the score on the partitioning test. DIMTEST has been widely studied for dichotomous item exams and has good power when AT and PT are chosen well (e.g., Froelich & Habing, 2008). If AT and PT are chosen poorly (e.g., both are random samples of items), the procedure will have power near 0.

Like a non-aggregated CARP test, DIMTEST uses conditional covariances, but whereas a CARP test rejects unidimensionality if the conditional covariances are negative, DIMTEST rejects unidimensionality if the conditional covariances are too high. Both CARP tests and DIMTEST divide the set of items into a partitioning test and an assessment test first, but CARP tests restrict the assessment test to a pair of items. Our new aggregated CARP (ACARP) test avoids this problem of selecting an assessment test by aggregating the individual CARP tests across item pairs. DIMTEST handles this problem by splitting the test based on factor analysis in a training sample, and the procedure can be improved further with bootstrapping (Froelich & Habing, 2008). DIMTEST is based on the statistic

$$T = \sum_{\substack{i,j \in AT \\ i \neq j}} \mathbb{E}\left(Cov\left(X_i, X_j | \Theta\right)\right).$$

A sample estimate $\widehat{T}$ of T is obtained by replacing Θ with an estimate $\widehat{\Theta}$ based on the partitioning test, usually the sum score. Stout (1987) argued that $Cov\left(X_i, X_j | \Theta\right) = 0$ if the test is unidimensional, and, therefore, $Cov\left(X_i, X_j | \widehat{\Theta}\right) \approx 0$ if $\widehat{\Theta}$ is a good estimate of Θ. A high value of $\widehat{T}$ means that the test is not unidimensional. This may be due to multi-dimensionality or lack of local independence. Several adjustments and improvements of DIMTEST have been suggested to estimate or reduce the bias in $\widehat{T}$ caused by $\widehat{\Theta}$ being a fallible estimate of Θ (e.g., Kieftenbeld & Nandakumar, 2015), especially if the number of items is small.

DETECT is a method to cluster items based on their conditional covariances. The clustering is based on Zhang and Stout's (1999) theory of conditional covariances for tests with a simple structure. The procedure produces a clustering of the items, and a unidimensional test should result in one cluster that contains all items. This is an explorative method, and not a statistical significance test. For a given partition $\mathcal{P}$ of the set of test items (that is, $\{1, \ldots, J\}$ in our notation), the theoretical DETECT index is

$$D := \frac{1}{J(J-1)} \sum_{\substack{i,j \in TT \\ i \neq j}} \delta_{ij}^{\mathcal{P}} \mathbb{E}\left(Cov\left(X_i, X_j | \Theta\right)\right),$$

where TT is the set of all items in the test, and $\delta_{ij}^{\mathcal{P}} = 1$ if i and j are elements of the same cluster in $\mathcal{P}$, and $\delta_{ij}^{\mathcal{P}} = -1$ otherwise. DETECT searches for the partition that maximizes D using an estimate $\widehat{\Theta}$ that replaces Θ, leading to a sample estimate $\widehat{D}$ of D. Many adjustments and improvements of DETECT have

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

been suggested to estimate or reduce the bias in $\widehat{D}$ caused by $\widehat{\Theta}$ being a fallible estimate of Θ (Roussos & Ozbek, 2006; Zhang, 2007), especially if the number of items is small.

Considering the relationship between DIMTEST and DETECT, we study how DETECT could be used as a confirmatory test of unidimensionality. It would then be logical to define $\mathcal{P}$ as one cluster that contains all items, and the theoretical DETECT index for this would be

$$D_1 := \frac{1}{J(J-1)} \sum_{\substack{i,j \in TT \\ i \neq j}} \mathbb{E}\left(Cov\left(X_i, X_j | \Theta\right)\right).$$

Aside from the factor $1/J(J-1)$, one could describe D_1 as an instance of T where both the assessment test and the partitioning test contain all items. A high value of D_1 would lead to the conclusion that the test is not unidimensional.

For the sake of comparison, a single CARP test would test whether $\mathbb{E}\left(Cov\left(X_i, X_j | \widehat{\Theta}_i + \widehat{\Theta}_j\right)\right) \geq 0$, where $\widehat{\Theta}_i$ and $\widehat{\Theta}_j$ are predictors of X_i and X_j, respectively, based on the items excluding i and j. The Z_{PS} statistic based on the pairwise MCCs C_{ij+}, as defined earlier, can be considered a sample estimate of the theoretical index

$$\zeta_{PS} := \frac{\sum_{(i,j) \in T_-} \mathbb{E}\left(Cov\left(X_i, X_j | \widehat{\Theta}_i + \widehat{\Theta}_j\right)\right)}{\sqrt{\sum \boldsymbol{\Omega}_-}},$$

where T_- consists of the pairs for which $\mathbb{E}\left(Cov\left(X_i, X_j | \widehat{\Theta}_i + \widehat{\Theta}_j\right)\right) < 0$ and $i < j$ in the training sample, and $\boldsymbol{\Omega}_-$ is the covariance matrix used for normalization. A negative value of ζ_{PS} would lead to the conclusion that the test is not unidimensional. The DIMTEST index $\widehat{T}$ and the DETECT index $\widehat{D}_1$ thus have a structure that is very similar to the Z_{PS} statistic before the latter is normalized with $\sqrt{\sum \boldsymbol{\Omega}_-}$. The main difference is that they are computed using different pairs (i,j). While DIMTEST would use pairs with *high* conditional covariances in the training sample, Z_{PS} would use pairs with *low* conditional covariances in the training sample. Note that if $\widehat{\Theta}_i$ and $\widehat{\Theta}_j$ are poor estimates, $\mathbb{E}\left(Cov\left(X_i, X_j | \widehat{\Theta}_i + \widehat{\Theta}_j\right)\right)$ must still be nonnegative, and therefore ACARP does not require bias corrections in order to control the Type I error rate.

DIMTEST, DETECT, and the ACARP tests developed here are closely related. The main difference is the choice of the targeted item pairs and the conditioning variable and the implications that this has for the sign of the covariances. In DIMTEST and DETECT, the conditioning variable is supposed to capture the partitioning test, and unidimensionality is rejected if the conditional covariances in the assessment test are high. In the ACARP tests, one would rather combine covariances of pairs from different dimensions; the conditioning variables are supposed to predict the assessment items, and unidimensionality is rejected if the conditional covariances are negative. For example, if the test has two dimensions A and B, DIMTEST would use A as the assessment test and B as the partitioning test, or conversely; but Z_{PS} would use pairs (i,j) with $i \in A$ and $j \in B$.

8. Simulation study I: preliminary selection of test methods

We investigated whether the Type 1 error rate is under control in typical IRT cases, and we compared our test methods on statistical power. In the first simulation study, we aimed to make a preliminary selection of the most promising test methods, which we investigated further in the second and third simulation studies. We used J items and a logistic model,

$$P\left(X_i = 1 | \Theta_1, \Theta_2\right) = \left(1 + \exp\left(-\left(\alpha_{i1}\Theta_1 + \alpha_{i2}\Theta_2 + \beta_i\right)\right)\right)^{-1},$$

where (Θ_1, Θ_2) has a bivariate standard normal distribution with correlation 0. Denote the number of items that load on dimensions 1 and 2 as J_1 and J_2, respectively, so that $J_1 + J_2 = J$.

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

For the first simulations, we used $J = 10$. We studied three possible dimensionality cases:

- Dimensionality 0: In this case, $\alpha_{i1} = \alpha_{i2} = 0$ for $1 \leq i \leq J$
- Dimensionality 1: In this case, $\alpha_{i1} > 0$ and $\alpha_{i2} = 0$ for $1 \leq i \leq J$
- Dimensionality 2: In this case, $\alpha_{i1} > 0$ and $\alpha_{i2} = 0$ for $1 \leq i \leq J_1$, and $\alpha_{i2} > 0$ and $\alpha_{i1} = 0$ for $J_1 < i \leq J_2$. We used $J_1 = \text{ceiling}(J/2)$.

The methods discussed allow several ways to obtain a p-value: based on C_{ij+} or Z_{ij}; aggregated with LR, CL, PS, PP, PB, CS, CP, or CB; and their covariance matrix estimated with the sample moments or bootstrapping or set to the identity matrix. We adopt the following convention to name the tests with four-letter acronyms: The first letter indicates the pairwise statistic (Z for the Z_{ij} and M for the MCCs, C_{ij+}); the second letter indicates the covariance matrix (B for bootstrapping, M for moments, I for identity matrix, and N for none); the last two letters indicate the aggregation method (LR, CL, PS, PP, PB, CS, CP, or CB). For example, ZICS is based on Z_{ij}s with the identity matrix and aggregation with the conditionalized sum. An asterisk will be used to indicate a group of tests; for example, ZI** is the group of tests based on the Z_{ij}s with the identity matrix. Not all combinations are reasonable: The identity matrix is only reasonable for the Z_{ij}s but not for the C_{ij+}s, and the sample moments and bootstrap method make sense only for LR, CL, and PS. The remaining 20 relevant combinations are displayed in the first column of Table 1. In addition to these tests, we studied the DETECT index $\widehat{D}_1$.

The simulations were conducted in R, and the code is provided on the Open Science Framework (https://osf.io/hyuzm/). Statistical testing was done at a nominal level of significance $\alpha = 0.05$. We programmed the CARP tests with the training sample size equal to 30% of the total sample. For DETECT, we used the confirmatory DETECT function conf.detect of the **sirt** R-package (Robitzsch, 2022), with all items in one cluster; this gives $\widehat{D}_1$. We rejected unidimensionality if $\widehat{D}_1 > 0.20$, as recommended in the **sirt** documentation and Roussos and Ozbek (2006, p. 220).

DETECT had a rejection rate of 0 in all circumstances. In the Discussion, we reflect on this result. Tables 1 and 2 show the rejection rates for all other methods. Table 1 shows the rejection rates with all $\alpha_{id} \in \{0,1\}, \beta_i = 0$ for 1000 samples of 1000 subjects. All 1000 samples in a column are generated with the same parameters. Table 2 shows the rejection rates if the α_{id} that are not constrained to be zero have distribution $\alpha_{id} \sim Uniform(0,2)$ and the $\beta_i \sim Uniform(-2,2)$. The 1000 samples in a column of Table 2 all have different parameters, and all contain 1000 subjects.

We conclude that only the following combinations keep the Type 1 error rate under control in both dimensionality 0 and dimensionality 1, at least in the above cases:

- If the covariance matrix is replaced by the identity matrix: all aggregation methods based on the pairwise Z_{ij}-statistics.
- If the covariance matrix is estimated from sample moments: all aggregation methods based on pairwise C_{ij+} or Z_{ij}-statistics.
- If the covariance matrix is estimated by bootstrapping: only PS, based on pairwise C_{ij+} or Z_{ij}-statistics.

The tests *BCL have a Type I error rate that significantly exceeds 0.05 in Table 1 ($p = 0.0006$). If we omit these tests, and compare the other tests that use a covariance matrix (CL, LR, and PS) across the different versions (M or Z; bootstrap, moments, or identity), then the tests based on the pairwise Z_{ij}-statistics with the identity matrix have the highest power. The tests where the covariance matrix was based on the sample moments had the lowest power, and we conclude that this method has no advantages.

The maximum discrimination parameter $\alpha_{id} = 1.0$ in the simulations of Tables 1 and 2 is rather low. For a broader view, we also conducted simulations with discrimination parameters $\alpha_{id} = 1.7$ (medium) and $\alpha_{id} = 7.0$ (extremely high) if the item loads on dimension d, using 100 simulations of 1000 subjects per case. Figure 1 shows the plots of the p-values (Schweder & Spjøtvoll, 1982) for the cases with

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table 1. Rejection rates for various tests, based on 1000 samples with fixed item parameters

Pairwise statistic	Estimation covariance matrix	Aggregation method	Rejection rate dimensionality 0	Rejection rate dimensionality 1	Rejection rate dimensionality 2
M	bootstrap	CL	0.074	0.000	0.557
M	bootstrap	LR	0.062	0.000	0.182
M	bootstrap	PS	0.028	0.000	0.596
Z	bootstrap	CL	0.061	0.000	0.529
Z	bootstrap	LR	0.055	0.000	0.168
Z	bootstrap	PS	0.028	0.000	0.592
Z	identity	CL	0.044	0.000	0.554
Z	identity	CP	0.043	0.000	0.584
Z	identity	CS	0.036	0.000	0.587
Z	identity	LR	0.023	0.000	0.336
Z	identity	PP	0.029	0.000	0.755
Z	identity	PS	0.026	0.000	0.686
M	moments	CL	0.042	0.000	0.190
M	moments	LR	0.028	0.000	0.000
M	moments	PS	0.023	0.000	0.312
Z	moments	CL	0.054	0.000	0.349
Z	moments	LR	0.033	0.000	0.000
Z	moments	PS	0.026	0.000	0.368
Z	none	CB	0.027	0.000	0.242
Z	none	PB	0.038	0.000	0.296

Note: M = MCC (C_{ij+}); Z = pairwise Z-statistic (Z_{ij}); LR = likelihood ratio; CL = conditional likelihood ratio; CS = conditional sum; CP = conditional product; CB = conditional Bonferroni; PS = preselected sum; PP = preselected product; and PB = preselected Bonferroni. The item parameters were fixed to $\alpha_{id} \in \{0,1\}, \beta_i = 0$. Each of the 1000 samples contained 1000 subjects.

dimensionality 0 and 1, and Figure 2 shows these plots for dimensionality 2. If the *p*-values have a uniform distribution, they lay on the diagonal line $y = x$ in the plot. These plots confirm the conclusions of Tables 1 and 2: in cases with dimensionality 0 ($\alpha_{i1} = \alpha_{i2} = 0$), all tests produce *p*-values that are approximately uniformly distributed or slightly higher. In cases with dimensionality 1, ($\alpha_{i1} > 0, \alpha_{i2} = 0$), all tests produce *p*-values that higher than uniform, and this effect increases with the discrimination parameter. This is to be expected because the population values of the conditional covariances are positive in unidimensional cases with $\alpha_{i1} > 0$. The power is generally lowest if the covariance matrix is estimated with the moments method. With bootstrapping, each Z***-test produces *p*-values that are very close to the *p*-values of the corresponding M***-test. However, the power of the tests based on the identity matrix matches or outperforms the power of the corresponding tests based on bootstrapping. The next simulation study will therefore focus on the tests based on the identity matrix.

9. Simulation study II: comparison of aggregation methods

In this second simulation study, we focussed on the tests that use the identity matrix as the covariance matrix of the *Z*-statistics. The goal was to determine which aggregation methods (CL, LR, PS, CS, PP, CP, PB, and CB) have the highest power and whether this depends on the number of items, number of subjects, and discrimination parameters. The goal was furthermore to determine whether there are

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table 2. Rejection rates for various tests, based on 1000 samples with random item parameters

Pairwise statistic	Estimation covariance matrix	Aggregation method	Rejection rate dimensionality 0	Rejection rate dimensionality 1	Rejection rate dimensionality 2
M	bootstrap	CL	0.055	0.010	0.293
M	bootstrap	LR	0.038	0.000	0.125
M	bootstrap	PS	0.012	0.000	0.232
Z	bootstrap	CL	0.055	0.009	0.288
Z	bootstrap	LR	0.038	0.000	0.123
Z	bootstrap	PS	0.011	0.000	0.217
Z	identity	CL	0.024	0.009	0.289
Z	Identity	CP	0.023	0.008	0.292
Z	identity	CS	0.030	0.010	0.264
Z	identity	LR	0.012	0.000	0.160
Z	identity	PP	0.016	0.003	0.306
Z	identity	PS	0.009	0.000	0.248
M	moments	CL	0.017	0.006	0.065
M	moments	LR	0.008	0.000	0.000
M	moments	PS	0.006	0.000	0.078
Z	moments	CL	0.039	0.021	0.179
Z	moments	LR	0.014	0.000	0.000
Z	moments	PS	0.010	0.000	0.112
Z	none	CB	0.031	0.020	0.168
Z	none	PB	0.025	0.011	0.190

Note: M = MCC (C_{ij+}); Z = pairwise Z-statistic (Z_{ij}); LR = likelihood ratio; CL = conditional likelihood ratio; CS = conditionalized sum; CP = conditionalized product; CB = conditionalized Bonferroni; PS = preselected sum; PP = preselected product; and PB = preselected Bonferroni. The item parameters had distribution $\alpha_{i1} = \alpha_{i2} = 0$ (dimensionality 0), $\alpha_{i1} \sim Uniform(0,2)$, $\alpha_{i2} = 0$ (dimensionality 1), or $\alpha_{i1}, \alpha_{i2} \sim Uniform(0,2)$ (dimensionality 2), and $\beta_i \sim Uniform(-2,2)$. Each of the 1000 samples contained 1000 subjects.

cases with unexpected low power. We investigate the effects of the number of items (J), sample size (N), and discrimination parameter (α_{id}) on the rejection rates, with fixed item difficulty $\beta_i = 0$, using 100 simulations per combination. The rejection rates of the two-dimensional cases with low discrimination parameters, $\alpha_{id} \in \{0,1\}$, are shown in Figure 3.

For medium-valued discrimination parameters, $\alpha_{id} \in \{0,1.7\}$, the estimated power was generally 0.99 or 1.00 even with $N = 500$ and $J = 10$, except for PB and CB. We did not display these excellent power rates in a figure because they do not help discern any pattern. For low discrimination parameters, $\alpha_{id} \in \{0,1\}$, and $N = 2{,}000$, the power was usually about .90 or higher, with the exception of ZIPB and ZICB. The power differences between the tests are more pronounced for $\alpha_{id} \in \{0,1\}$ and $N = 500$ or $N = 1000$. There we see that the power of ZICL, ZICS, and ZICP increases with the number of items, and that the power of these three tests is generally the highest, except that the power of ZIPP is higher if the number of items is small. The power of ZIPB and ZICB is generally among the lowest and tends to remain low if the number of items increases with $N = 500$ or $N = 1000$. The power of ZIPS tends to decrease with the number of items if $N = 500$ or $N = 1000$. The power of ZILR tends to remain low if the number of items increases with $N = 500$, but slowly increases if $N = 1000$. In sum, the highest power is observed for ZICL, ZICS, ZICP, and sometimes ZIPP. Therefore, we will focus on these tests in the next section.

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

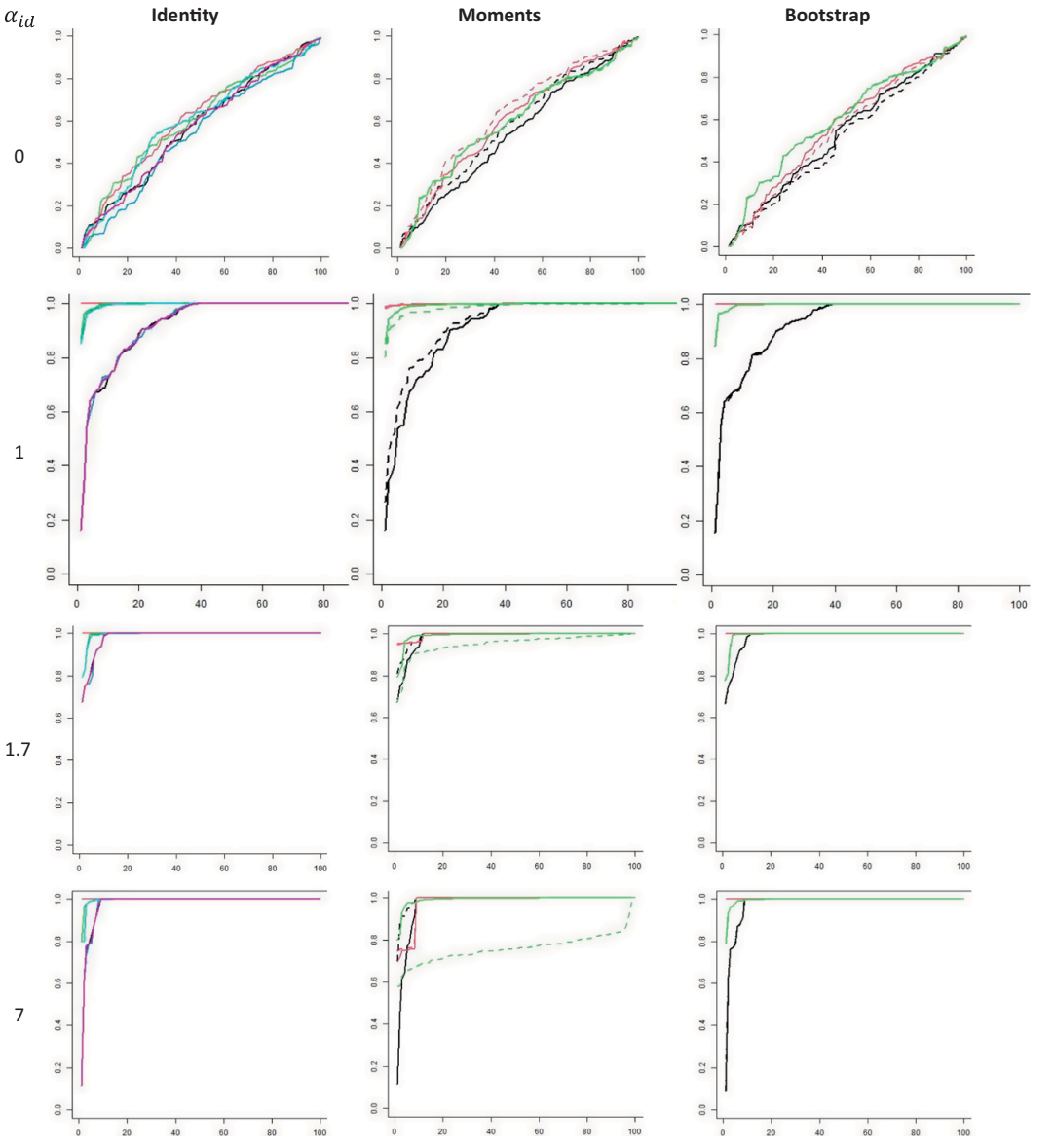

Figure 1. *Plots of p-values showing the Type I error rates.*
Note: The vertical axis is the *p*-value and the horizontal axis is the rank of the *p*-value. Dashed curve = M***, solid curve = Z***, black = CT, red = LR, green = PS, blue = CS, light blue = PP, and magenta = CP.

10. Simulation study III: type I error rates

In Simulation Study I, we investigated the Type I error rate only for $J = 10$ items and a sample size of $N = 1000$ subjects. The present section discusses the Type I error rate more thoroughly, with simulations with varying J and N, but only for the ZI** tests, which were selected in Simulation Study I, and we focus especially on the tests ZICL, ZICS, ZICP, and ZIPP, based on their power in Simulation Study II. We label tests with $J \leq 10$ small and tests with $J > 10$ large. Further, we consider $N \leq 1000$ small and $N > 1000$ large.

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

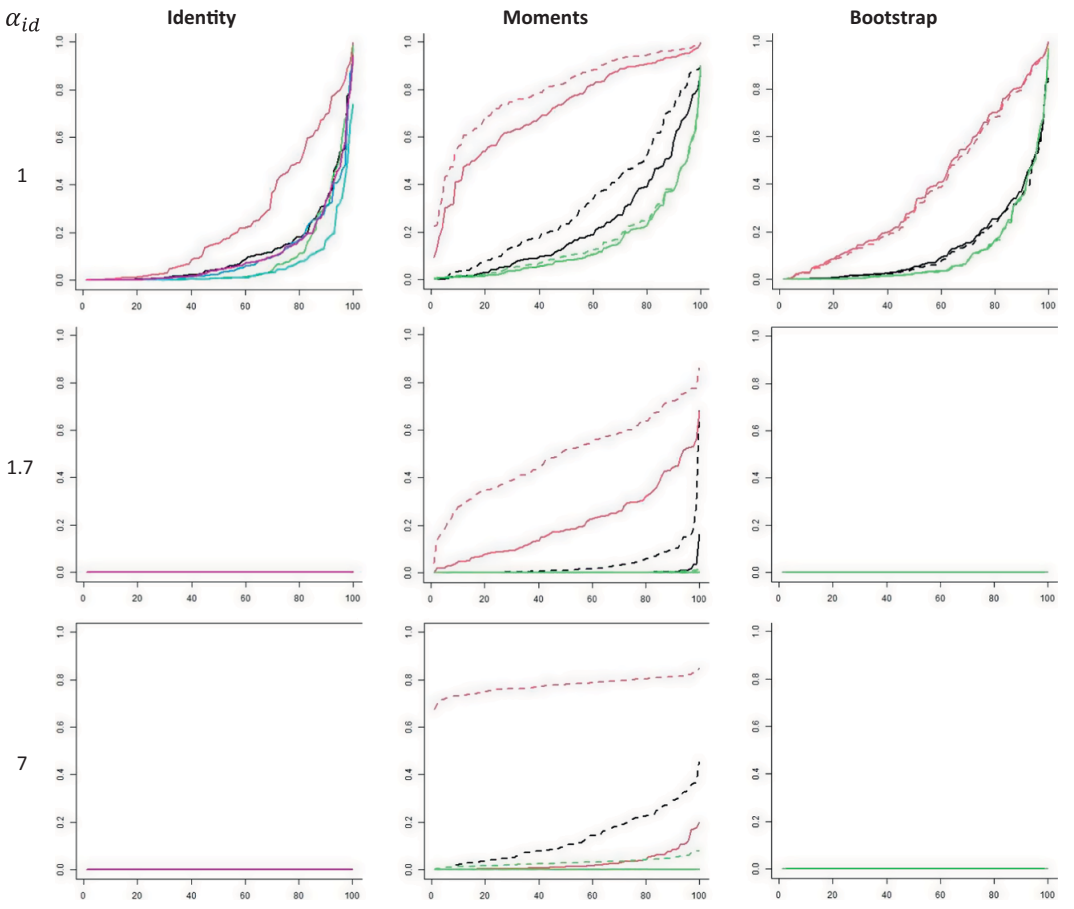

Figure 2. *Plots of p-values showing the power.*
Note: The vertical axis is the *p*-value and the horizontal axis is the rank of the *p*-value. Dashed curve = M***, solid curve = Z***, black = CT, red = LR, green = PS, blue = CS, light blue = PP, and magenta = CP.

10.1. Zero-dimensional cases

We investigated the effects of the number of items (J), sample sizes (N) on the rejection rates for a nominal significance level of 5%, with fixed discrimination parameters $\alpha_{id} = 0$ and fixed item difficulties $\beta_i = 0$, using 100 simulations per combination. For the number of items, we used all small values ($J = 3,4,5,6,7,8,9,10$) and two large values ($J = 20,30$). For the number of subjects, we used several small values ($N = 250,500,750,1000$) and two large values ($N = 2000, N = 10^4$). For each of the test statistics ZICL, ZICS, ZICP, and ZIPP, the cumulative distribution of rejection counts was larger than the cumulative binomial distribution with $n = 100, \pi = 0.05$; that is, the rejection rates were smaller than expected under the binomial distribution. The highest rejection rates (0.07, 0.08, 0.09, and 0.10) were mostly observed with $J = 3,4,5,6$ and $N = 10^4$. Therefore a second simulation was conducted with these values of J and N, but with 1000 simulations per combination. The rejection rates for ZICL, ZICS, ZICP, and ZIPP varied between 0.041 and 0.061, and none were significantly greater than 0.05. We also studied this for cases where the $J, N, \alpha_{id}, \beta_i$ were chosen randomly and independently from uniform distributions with J between 3 and 30, N between 250 and 10^4, and $\beta_i \in (-2,2)$. We sampled 100 cases of $(J, N, \boldsymbol{\beta})$, and generated 100 data sets with a zero-dimensional model for each case. The cumulative distribution of rejection counts was larger than the cumulative binomial distribution with $n = 100, \pi = 0.05$ for all ZI** tests except ZIPB. ZIPB had two cases with rejection rates of 0.11.

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

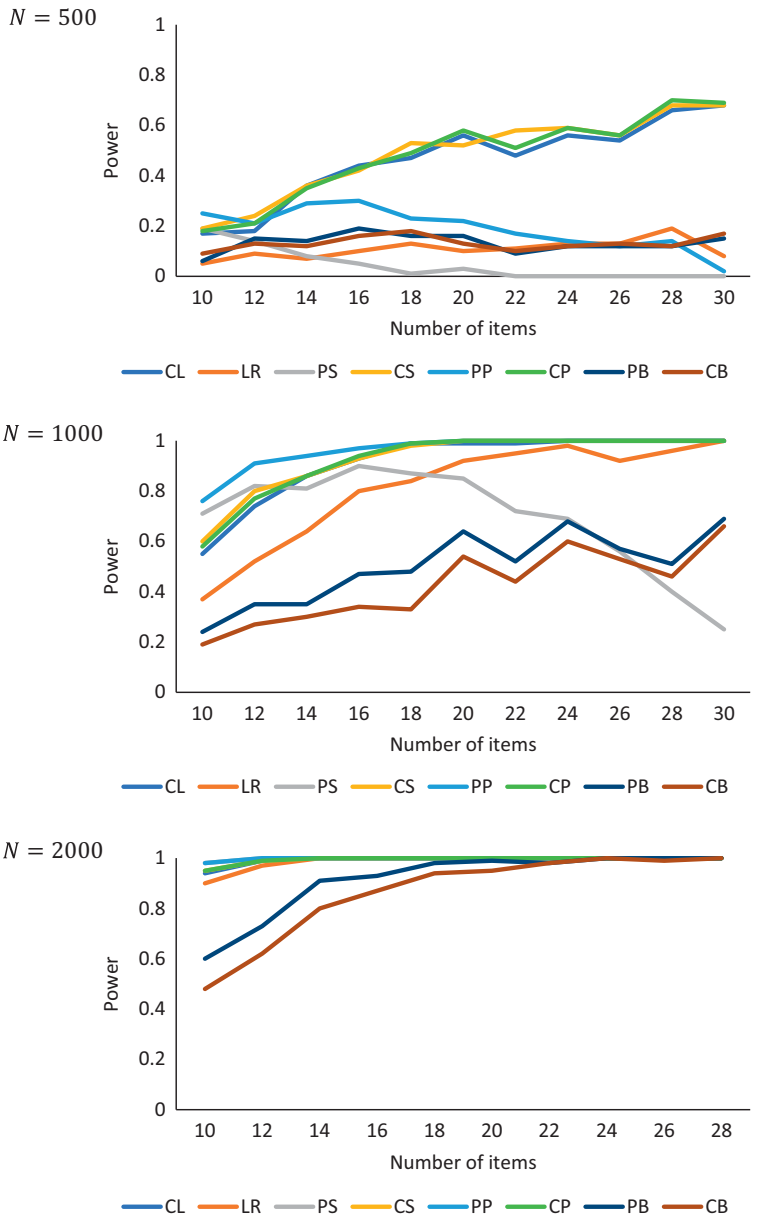

Figure 3. *Rejection rates as a function of the number of items and sample sizes.*

We conclude that the Type I error rate is under control for the tests ZICL, ZICS, ZICP, and ZIPP in these cases.

10.2. Unidimensional cases

We investigated the effects of the number of items (J), sample sizes (N) on the rejection rates for a nominal significance level of 5%, with fixed discrimination parameters $\alpha_{i1} = 1, \alpha_{i2} = 0$ and fixed item difficulties $\beta_i = 0$, using 100 simulations per combination. For the number of items, we used all small values $(J = 3,4,5,6,7,8,9,10)$ and two large values $(J = 20,30)$. For the number of subjects, we used

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table 3. *Distribution of rejection rates in various settings of J and N with unidimensional models*

J		*N*										
min	max	min	max	Rejection rate	ZICP	ZICL	ZICS	ZIPP	ZILR	ZIPS	ZICB	ZIPB
3	30	250	10^4	0.00	73	71	70	93	99	93	41	60
				0.01	18	19	22	5	1	4	25	25
				0.02	7	8	8	1	0	2	19	10
				0.03	2	1	0	1	0	1	10	4
				0.04	0	1	0	0	0	0	2	0
				0.05	0	0	0	0	0	0	2	0
				0.06	0	0	0	0	0	0	0	0
				0.07	0	0	0	0	0	0	0	0
				0.08	0	0	0	0	0	0	1	0
				0.09	0	0	0	0	0	0	0	1
10	10	10^3	10^4	0.00	46	44	47	79	100	92	36	49
				0.01	29	29	24	17	0	6	22	20
				0.02	12	11	16	2	0	0	18	13
				0.03	6	7	9	2	0	2	12	13
				0.04	5	6	3	0	0	0	3	3
				0.05	1	1	0	0	0	0	7	1
				0.06	1	2	1	0	0	0	2	1
10	20	10^3	10^4	0.00	71	67	65	95	100	96	31	47
				0.01	16	16	25	3	0	2	18	27
				0.02	5	10	6	0	0	2	25	16
				0.03	5	4	2	1	0	0	15	6
				0.04	2	1	1	1	0	0	3	2
				0.05	0	1	0	0	0	0	5	1
				0.06	0	0	0	0	0	0	3	0
				0.07	1	1	1	0	0	0	0	1
3	10	10^2	10^3	0.00	40	39	43	78	95	82	29	49
				0.01	36	36	32	12	5	11	28	29
				0.02	12	12	12	6	0	5	22	13
				0.03	5	4	7	3	0	2	11	4
				0.04	3	5	3	1	0	0	5	4
				0.05	0	1	0	0	0	0	3	0
				0.06	2	1	2	0	0	0	0	0
				0.07	2	1	1	0	0	0	2	0
				0.08	0	0	0	0	0	0	0	1
				0.09	0	1	0	0	0	0	0	0

Note: In each setting of J and N, 100 parameter cases were generated with $J \sim Uniform(min\,J, max\,J)$, $N \sim Uniform(min\,N, max\,N)$, $\alpha_{i1} \sim Uniform(0,2)$, $\alpha_{i2} = 0$, and $\beta_i \sim Uniform(-2,2)$. In each of these 100 parameter cases, 100 samples were simulated. Each cell shows the number of parameter cases with the rejection rate specified in that row. For example, for ZICP, there were 18 cases out of 100 that had a rejection rate of 0.01 over 100 samples in the first setting.

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

several small values ($N = 250, 500, 750, 1000$) and two large values ($N = 2000, N = 10^4$). The highest rejection rate was 0.01. We also studied this for cases where the $J, N, \alpha_{i1}, \beta_i$ were chosen randomly and independently from uniform distributions with J between 3 and 30, N between 250 and 10^4, $\alpha_{i1} \in (0, 2)$, and $\beta_i \in (-2, 2)$. We sampled 100 cases of $(J, N, \boldsymbol{\beta})$, and generated 100 data sets with a unidimensional model for each case. The rejection rates are given in the first nine rows of Table 3. The rejection rates of ZICP, ZICL, ZICS, and ZIPP were at most 0.04. The cumulative distribution of rejection counts of each of the tests was larger than the cumulative binomial distribution with $n = 100, \pi = 0.05$, that is, the rejection rates were smaller than expected under the binomial distribution. Table 3 also shows the rejection rates in various other settings of J and N, with similar conclusions.

11. Simulation study IV: without continuity correction

The previous simulations were conducted with the Z_{ij}-statistics corrected for continuity with the addition of a term 0.5 in the numerator, as proposed by Rosenbaum (1984) and adopted by Ellis and Sijtsma (2023). Many other continuity corrections exist (Andrés et al., 2024), and we are not sure that a continuity correction is necessary for the sample sizes ordinarily found in IRT. Therefore, we repeated the simulation studies of the ZI-statistics without continuity correction. The simulations of Table 1 and Table 2 are repeated in Table 4 and Table 5 without continuity correction.

As was to be expected, the rejection rates were now generally larger than in Tables 1 and 2. In Table 4, most Type I error rates (Dimensionality 0) were now 0.05 or slightly higher, and the power (Dimensionality 2) was substantially higher than in Table 1. In Table 5, all Type I error rates are below 0.05, and the power is still larger than in Table 2. The power rates in Tables 4 and 5 are low, but note that these results were obtained for low discrimination parameters ($\alpha_{id} \leq 1$). We also repeated Simulation Study III without continuity correction, and our conclusion is that the rejection rates were dominated by a binomial distribution with probability 0.05 in all cases, meaning that the Type I error rate is under control. The rejection rates of these versions of ZICL, ZICP, and ZICS are close to 0.05 in the zero-dimensional cases.

Table 6 shows the power rates for the simulations underlying Figure 3, but now repeated without continuity correction, with the positive discrimination parameters set to 1 (low). If the objective is to have power > 0.90, then all four tests achieved this goal with $N = 2000$ and $J \geq 10$, and also with $N = 1000$ and $J \geq 14$, but not with $N = 500$. However, if the positive discrimination parameters are equal to 1.7 (medium), then the power rates were 1.00 even with $N = 500$ and $J \geq 10$. We did not display these excellent power rates in a table because they were all 1.00.

Table 4. Rejection rates for various tests without continuity correction with fixed item parameters

Pairwise statistic	Covariance matrix	Aggregation method	Rejection rate dimensionality 0	Rejection rate dimensionality 1	Rejection rate dimensionality 2
Z	identity	CL	0.058	0.000	0.657
Z	identity	CP	0.058	0.000	0.686
Z	identity	CS	0.057	0.000	0.684
Z	identity	LR	0.052	0.000	0.447
Z	identity	PP	0.053	0.000	0.835
Z	identity	PS	0.050	0.000	0.756
Z	none	CB	0.039	0.000	0.278
Z	none	PB	0.051	0.000	0.334

Note: Z = pairwise Z-statistic (Z_{ij}); LR = likelihood ratio; CL = conditional likelihood ratio; CS = conditional sum; CP = conditional product; CB = conditional Bonferroni; PS = preselected sum; PP = preselected product; and PB = preselected Bonferroni. The item parameters were fixed to $\alpha_{id} \in \{0, 1\}, \beta_i = 0$. Each rate is based on 1000 samples of 1000 subjects.

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table 5. Rejection rates for various tests without continuity correction with random item parameters

Pairwise statistic	Covariance matrix	Aggregation method	Rejection rate dimensionality 0	Rejection rate dimensionality 1	Rejection rate dimensionality 2
Z	identity	CL	0.045	0.007	0.354
Z	Identity	CP	0.048	0.008	0.361
Z	identity	CS	0.047	0.009	0.332
Z	identity	LR	0.038	0.000	0.212
Z	identity	PP	0.035	0.002	0.381
Z	identity	PS	0.039	0.000	0.300
Z	none	CB	0.035	0.024	0.200
Z	none	PB	0.038	0.013	0.221

Note: Z = pairwise Z-statistic (Z_{ij}); LR = likelihood ratio; CL = conditional likelihood ratio; CS = conditionalized sum; CP = conditionalized product; CB = conditionalized Bonferroni; PS = preselected sum; PP = preselected product; and PB = preselected Bonferroni. The item parameters had distribution $\alpha_{i1} = \alpha_{i2} = 0$ (dimensionality 0), $\alpha_{i1} \sim Uniform\,(0,2)$, $\alpha_{i2} = 0$ (dimensionality 1), or $\alpha_{i1}, \alpha_{i2} \sim Uniform\,(0,2)$ (dimensionality 2), and $\beta_i \sim Uniform\,(-2,2)$. Each rate is based on 1000 samples of 1000 subjects.

Table 6. Rejection rates for various tests without continuity correction with low discrimination parameters

J	$N = 500$				$N = 1000$				$N = 2000$			
	CL	CS	PP	CP	CL	CS	PP	CP	CL	CS	PP	CP
10	0.30	0.31	0.37	0.31	0.66	0.77	0.89	0.72	0.95	0.98	1.00	0.96
12	0.35	0.39	0.32	0.37	0.86	0.87	0.93	0.89	1.00	0.99	1.00	1.00
14	0.41	0.43	0.37	0.40	0.92	0.91	0.94	0.92	1.00	1.00	1.00	1.00
16	0.62	0.61	0.51	0.65	0.98	0.98	0.99	0.98	1.00	1.00	1.00	1.00
18	0.61	0.60	0.38	0.64	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00
20	0.69	0.64	0.35	0.68	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
22	0.68	0.71	0.33	0.71	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
24	0.74	0.78	0.29	0.81	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
26	0.78	0.74	0.20	0.81	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
28	0.84	0.83	0.23	0.84	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
30	0.81	0.78	0.15	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: Z = pairwise Z-statistic (Z_{ij}); CL = conditional likelihood ratio; CS = conditional sum; CP = conditional product; and PP = preselected product. The item parameters were fixed to $\alpha_{id} \in \{0,1\}$, $\beta_i = 0$. Each rate is based on 100 samples.

12. Conclusions and discussion

We conclude that the pairwise CARP tests Ellis and Sijtsma (2023) proposed can best be aggregated with four of the tests developed here: ZICL, ZICP, ZICS, and ZIPP. These tests control the Type I error rate in a wide variety of test lengths and sample sizes, and their power against two-dimensional alternatives is larger than the power of other aggregate statistics that we studied. ZIPP had the greatest relative power if there were less than 18 items with $N = 1000$, but not in most other cases. Further investigations are needed to determine whether this is also true for alternatives with more than two dimensions.

The Type I error rates of the ZI-tests are well below the nominal rate of 0.05, and this suggests that improvement is possible. The pairwise Z-statistic, as defined by Ellis and Sijtsma (2023), includes a continuity correction that might be too conservative. Based on our simulations, we conclude that the continuity correction may be abandoned for the sample sizes we studied ($N \geq 500$). Without continuity

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

correction, the Type I error rate is still under control, and the power increases. The Type I error rates of ZICL, ZICP, and ZICS are then close to 0.05 in the zero-dimensional cases of Tables 4 and 5.

The Type I error rate in unidimensional cases is far below 0.05, even without continuity correction, but this does not imply that the tests are too conservative. As an analogy, consider the elementary normal-theory one-sided Z-test for a mean μ with known variance σ^2. For the null hypothesis $\mu \geq 0$ and sample mean $\overline{X}$ one would use $Z = \left(\overline{X}/\sigma\right)\sqrt{N}$ and reject the null hypothesis if $\Phi(Z) < \alpha$. If real data were generated with $\mu > 0$, then $P\left(\Phi(Z) < \alpha\right) < \alpha$, meaning that the Type I error rate is less than α. This is usually not viewed as a sign that something is wrong with the one-sided Z-test. The situation in our case is similar because we have a one-sided test for a mean, but in our case it is a mean of conditional covariances. In the unidimensional case, this mean is positive, which reduces the Type I error rate.

If the discrimination parameters equal 1 and the intercepts equal 0, the power rates of ZICL, ZICP, ZICS, and ZIPP are well above 0.90 for $N = 2000$, regardless of whether the continuity correction is used. For these item parameters, if there are at least 14 items and the continuity correction is abandoned, the power is also above 0.90 for $N = 1000$, but the power is substantially below 0.90 for $N = 500$ for all studied test lengths between $J = 10$ and $J = 30$. We emphasize that these power rates were obtained for low discrimination parameters. We consider discrimination parameters of 1 as low because we did not use the general factor 1.7 in our parametrization (unlike e.g., Roussos & Ozbek, 2006). If the positive discrimination parameters equal 1.7 (medium), then the power rates of ZICL, ZICP, ZICS, and ZIPP are 1.00, even with $N = 500$ and all investigated test lengths from 10 to 30, based on simulations using 100 samples.

We also compared our statistics with the DETECT index, applied in a confirmatory manner, using the criterion $D_1 < 0.20$. To our surprise, despite the theoretical similarity of this index to the ZIPS statistic, this index appeared to lack discriminatory power, as it never rejected the hypothesis of unidimensionality. This is a puzzling result, seemingly at odds with the positive evaluations reported by the index's creators. While we have concerns about the validity of our results for DETECT, we were unable to identify any errors in our code. We believe this issue warrants further investigation.

Our study provides three new statistics for a confirmatory test of unidimensionality in monotone IRT models, and they seem to outperform older methods—at least in the cases we simulated. Still, the power of these methods is somewhat disappointing for sample size $N = 1000$ and discrimination parameter 1, and better methods may be possible. A simple improvement might be found in the size of the training sample, which was set at 30% in all our analyses. Furthermore, aggregation of different splits into training samples and test samples might be useful. Finally, it would be worthwhile to investigate which of the four tests can be recommended as most powerful under various circumstances.

Data availability statement. The simulated data and the code that generated it, are available in the Open Science Framework repository at https://osf.io/hyuzm/

References

Anderson, T. W., & Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, *28*(1), 89–110. https://doi.org/10.1214/aoms/1177707039

Andrés, A. M., Hernández, M. Á., & Moreno, F. G. (2024). The Yates, Conover, and Mantel statistics in 2 × 2 tables revisited (and extended). *Statistica Neerlandica*, *78*(2), 334–356. https://doi.org/10.1111/stan.12320

Bartolucci, F., & Forcina, A. (2000). A likelihood ratio test for MTP_2within binary variables. *The Annals of Statistics*, *28*, 1206–1218. https://doi.org/10.1214/aos/1015956713

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Clarke, B., & Yuan, A. (2001). Manifest characterization and testing for certain latent properties. *The Annals of Statistics*, *29*(3), 876–898. https://doi.org/10.1214/aos/1009210693

Davidov, O. (2011). Combining p-values using order-based methods. *Computational Statistics & Data Analysis*, *55*(7), 2433–2444. https://doi.org/10.1016/j.csda.2011.01.024

De Gooijer, J.G., & Yuan, A. (2011). Some exact tests for manifest properties of latent trait models. *Computational Statistics & Data Analysis*, *55*, 34–44. https://doi.org/10.1016/j.csda.2010.04.022

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*, 234–243. https://doi.org/10.1177/01466210122032046

Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika*, *79*, 303–316. https://doi.org/10.1007/s11336-013-9341-5

Ellis, J. L. (2015). MTP2 and partial correlations in monotone higher-order factor models. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research. The 78th annual meeting of the psychometric society* (pp. 261–272). Springer. https://doi.org/10.1007/978-3-319-07503-7_16

Ellis, J. L., Pecanka, J., & Goeman, J. J. (2018). Gaining power in multiple testing of interval hypotheses via conditionalization. *Biostatistics*, *21 (* 2 *)*, e65–e79. https://doi.org/10.1093/biostatistics/kxy042

Ellis, J. L. & Sijtsma, K. (2023). A test to distinguish monotone homogeneity from monotone multifactor models. *Psychometrika*, *88* (2), 387–412. https://doi.org/10.1007/s11336-023-09905-w

Falk, C. F., & Cai, L. (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika*, *81*, 434–460. https://doi.org/10.1007/s11336-014-9428-7

Froelich, A. G., & Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. *Applied Psychological Measurement*, *32*(2), 138–155. https://doi-org.ezproxy.elib11.ub.unimaas.nl/10.1177/0146621607300421

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, *14*, 1523–1543. https://doi.org/10.1214/aos/1176350174

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*(1), 65–81. https://doi.org/10.1177/01466216000241004

Kieftenbeld, V., & Nandakumar, R. (2015). Alternative hypothesis testing procedures for DIMTEST. *Applied Psychological Measurement*, *39*(6), 480–493. https://doi.org/10.1177/0146621615577618

Li, C.-H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Li, T., Habing, B., & Roussos, L. (2017). Conditional covariance-based subtest selection for polytomous DIMTEST. *Applied Psychological Measurement*, *41*(3), 209–226. https://doi.org/10.1177/0146621616681130

Liao, X., & Meyer, M. C. (2014). coneproj: An R package for the primal or dual cone projections with routines for constrained regression. *Journal of Statistical Software*, *61*(12). https://doi.org/10.18637/jss.v061.i12

Ligtvoet, R. (2022). Incomplete tests of conditional association for the assessment of model assumptions. *Psychometrika*. https://doi.org/10.1007/s11336-022-09841-1

Mokken, R. J. (1971). *A theory and procedure of scale-analysis*. The Hague: Mouton.

Robertson, T., Wright, F. T. & Dykstra, R. L. (1988). *Order restricted statistical inference*. Chichester: Wiley.

Robitzsch, A. (2022). Package 'sirt'. https://cran.r-project.org/web/packages/sirt/sirt.pdf

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*(3), 425–435. https://doi.org/10.1007/bf02306030

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, *43*(3), 215–243. https://doi.org/10.1111/j.1745-3984.2006.00014.x

Schweder, T., & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, *69*(3), 493. https://doi.org/10.2307/2335984

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617. https://doi.org/10.1007/bf02294821

Stout, W., Habing, B., Douglas, J., Hae Rim Kim, Roussos, L., & Jinming Zhang. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*(4), 331–354. https://doi.org/10.1177/014662169602000403

Van den Wollenberg, A.L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123–140. https://doi.org/10.1007/BF02296270

Van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika*, *67*, 519–538.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement*, *25*, 283–294.

Wollan, P. C. & Dykstra, R. L. (1986). Conditional tests with an order restriction as a null hypothesis. In R. L. Dykstra, T. Robertson, and F. T. Wright (Eds), *Advances in Order Restricted Statistical Inference* (pp. 279–295). Springer-Verlag.

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika*, *72*(1), 69–91. https://doi.org/10.1007/s11336-004-1257-7

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, *64*(2), 129–152. https://doi.org/10.1007/bf02294532

Zhang, J., & Stout, W. (1999b). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249. https://doi.org/10.1007/BF02294536

A. Appendix

Proposition A1. *If Z is independent of (X,Y) then $Cov(X,YZ) = Cov(X,Y)\,\mathbb{E}(Z)$.*

Proof. $Cov(X,YZ) = \mathbb{E}(XYZ) - \mathbb{E}(X)\mathbb{E}(YZ) = \mathbb{E}(XY)\mathbb{E}(Z) - \mathbb{E}(X)\mathbb{E}(Y)\mathbb{E}(Z) = Cov(X,Y)\,\mathbb{E}(Z)$. □

A1. The covariance of two sample covariances

In asymptotic distribution free (ADF) SEM, it is common to obtain the covariance matrix of sample covariances as the sample grows to infinity, which is called the asymptotic covariance matrix (e.g., Browne, 1984). However, it is also possible to obtain an exact formula for the covariance of two sample covariances for finite N, provided that the variables have finite fourth moments.

We develop the covariance of two sample covariances, assuming finite fourth moments of the involved variables. Denote the variables to be studied as $\boldsymbol{X} = (X_1, X_2, \ldots, X_J)$. Suppose that a random sample of N subjects is drawn, and denote the score of subject n on variable i as $X_i^{(n)}$, and let $\boldsymbol{X}^{(n)} = \left(X_1^{(n)}, \ldots, X_J^{(n)}\right)$, the score pattern of subject n. We assume that the N subjects are drawn independently, and that therefore their score patterns $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(N)}$ are independent, and that each $\boldsymbol{X}^{(n)}$ has the same multivariate distribution as $\boldsymbol{X}$. Thus, the $\boldsymbol{X}^{(n)}$ are independent copies of $\boldsymbol{X}$. We study two sample covariances, given by

$$C_{ij}^N = \frac{\sum_{n=1}^N X_i^{(n)} X_j^{(n)}}{N} - \frac{\sum_{n=1}^N X_i^{(n)}}{N}\frac{\sum_{n=1}^N X_j^{(n)}}{N}$$

$$C_{kl}^N = \frac{\sum_{n=1}^N X_k^{(n)} X_l^{(n)}}{N} - \frac{\sum_{n=1}^N X_k^{(n)}}{N}\frac{\sum_{n=1}^N X_l^{(n)}}{N}$$

Using the summation rules for covariances, the covariance of these sample covariances is

$$Cov\left(C_{ij}^N, C_{kl}^N\right) = Cov\left(\frac{\sum_{n=1}^N X_i^{(n)} X_j^{(n)}}{N} - \frac{\sum_{n=1}^N X_i^{(n)}}{N}\frac{\sum_{n=1}^N X_j^{(n)}}{N}, \frac{\sum_{n=1}^N X_k^{(n)} X_l^{(n)}}{N} - \frac{\sum_{n=1}^N X_k^{(n)}}{N}\frac{\sum_{n=1}^N X_l^{(n)}}{N}\right) =$$

$$N^{-2}Cov\left(\sum_{n=1}^N X_i^{(n)} X_j^{(n)}, \sum_{n=1}^N X_k^{(n)} X_l^{(n)}\right) - N^{-3}Cov\left(\sum_{n=1}^N X_i^{(n)} X_j^{(n)}, \sum_{n=1}^N X_k^{(n)} \sum_{n=1}^N X_l^{(n)}\right)$$
$$- N^{-3}Cov\left(\sum_{n=1}^N X_k^{(n)} X_l^{(n)}, \sum_{n=1}^N X_i^{(n)} \sum_{n=1}^N X_j^{(n)}\right) + N^{-4}Cov\left(\sum_{n=1}^N X_i^{(n)} \sum_{n=1}^N X_j^{(n)}, \sum_{n=1}^N X_k^{(n)} \sum_{n=1}^N X_l^{(n)}\right).$$

We will now develop the four terms in this sum, which will be called the four *main terms*. For the first main term, we need

$$Cov\left(\sum_{n=1}^N X_i^{(n)} X_j^{(n)}, \sum_{n=1}^N X_k^{(n)} X_l^{(n)}\right) = \sum_{n=1}^N \sum_{m=1}^N Cov\left(X_i^{(n)} X_j^{(n)}, X_k^{(m)} X_l^{(m)}\right).$$

This sum has N^2 terms, but if $n \neq m$ then $Cov\left(X_i^{(n)} X_j^{(n)}, X_k^{(m)} X_l^{(m)}\right) = 0$, and in the remaining N terms with $n = m$, $Cov\left(X_i^{(n)} X_j^{(n)}, X_k^{(m)} X_l^{(m)}\right) = Cov\left(X_i^{(n)} X_j^{(n)}, X_k^{(n)} X_l^{(n)}\right) = Cov\left(X_i X_j, X_k X_l\right)$. Therefore, we obtain

$$Cov\left(\sum_{n=1}^N X_i^{(n)} X_j^{(n)}, \sum_{n=1}^N X_k^{(n)} X_l^{(n)}\right) = \sum_{n=1}^N \sum_{m=1}^N Cov\left(X_i^{(n)} X_j^{(n)}, X_k^{(m)} X_l^{(m)}\right) = NCov\left(X_i X_j, X_k X_l\right)$$

For the second main term, we need

$$Cov\left(\sum_{n=1}^N X_i^{(n)} X_j^{(n)}, \sum_{n=1}^N X_k^{(n)} \sum_{n=1}^N X_l^{(n)}\right) = \sum_{n=1}^N \sum_{m=1}^N \sum_{r=1}^N Cov\left(X_i^{(n)} X_j^{(n)}, X_k^{(m)} X_l^{(r)}\right).$$

Similar to the previous term, the outcome of the generic covariance term in this sum, $Cov\left(X_i^{(n)} X_j^{(n)}, X_k^{(m)} X_l^{(r)}\right)$, depends on which of the indices $n, m, and\ r$ are equal. To keep track of this, we constructed Table A1, in which all possible truth values of the equalities $n = m$, $n = r$, and $m = r$ are listed. Each row contains one combination of truth values and the number of terms with that combination. Each row also contains the intermediate expression of the covariance, where the true equalities of that

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table A1. *Development of the term* $\sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{r=1}^{N} Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(m)}X_l^{(r)}\right)$

$n=m$	$n=r$	$m=r$	Count	Intermediate expression	Final expression
0	0	0	$N(N-1)(N-2)$	$Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(m)}X_l^{(r)}\right)$	0
0	0	1	$N(N-1)$	$Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(m)}X_l^{(m)}\right)$	0
0	1	0	$N(N-1)$	$Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(m)}X_l^{(n)}\right)$	$Cov\left(X_iX_j,X_l\right)\mathbb{E}\left(X_k\right)$
0	1	1	0		
1	0	0	$N(N-1)$	$Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(n)}X_l^{(r)}\right)$	$Cov\left(X_iX_j,X_k\right)\mathbb{E}\left(X_l\right)$
1	0	1	0		
1	1	0	0		
1	1	1	N	$Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(n)}X_l^{(n)}\right)$	$Cov\left(X_iX_j,X_kX_l\right)$

row are substituted in the generic expression $Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(m)}X_l^{(r)}\right)$, and the final expression, where this is intermediate expression is rewritten in terms of the central moments of $\boldsymbol{X}$. A few examples may clarify this:

In the second row, we consider the generic terms $Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(m)}X_l^{(r)}\right)$ with $n\neq m,n\neq r,m=r$. There are $N(N-1)$ such terms, and substitution of the equality $m=r$ leads to the intermediate expression $Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(m)}X_l^{(m)}\right)$. Since $n\neq m$, $X_i^{(n)}X_j^{(n)}$ is independent of $X_k^{(m)}X_l^{(m)}$, and therefore this covariance is 0, which is the final expression.

In the third row, we consider the generic terms $Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(m)}X_l^{(r)}\right)$ with $n\neq m,n=r,m\neq r$. There are $N(N-1)$ such terms, and substitution of the equality $n=r$ leads to $Cov\left(X_i^{(n)}X_j^{(n)},X_k^{(m)}X_l^{(n)}\right)$. Since $n\neq m$, $X_i^{(n)}X_j^{(n)}$ is independent of $X_k^{(m)}$ but not of $X_l^{(n)}$. Here, we can use proposition A1, and obtain $Cov\left(X_iX_j,X_l\right)\mathbb{E}\left(X_k\right)$.

In the fourth row, we consider the terms with $n\neq m,n=r,m=r$. This is a logical contradiction, and therefore there are 0 of such terms.

Summarizing from Table A1, we obtain

$$Cov\left(\sum_{n=1}^{N}X_i^{(n)}X_j^{(n)},\sum_{n=1}^{N}X_k^{(n)}\sum_{n=1}^{N}X_l^{(n)}\right)=N(N-1)\,Cov\left(X_iX_j,X_l\right)\mathbb{E}\left(X_k\right)+N(N-1)\,Cov\left(X_iX_j,X_k\right)\mathbb{E}\left(X_l\right)+NCov\left(X_iX_j,X_kX_l\right).$$

For the third main term, we obtain analogously,

$$Cov\left(\sum_{n=1}^{N}X_k^{(n)}X_l^{(n)},\sum_{n=1}^{N}X_i^{(n)}\sum_{n=1}^{N}X_j^{(n)}\right)=N(N-1)\,Cov\left(X_kX_l,X_j\right)\mathbb{E}\left(X_i\right)+N(N-1)\,Cov\left(X_kX_l,X_i\right)\mathbb{E}\left(X_j\right)+NCov\left(X_iX_j,X_kX_l\right).$$

For the fourth term, we need to develop

$$Cov\left(\sum_{n=1}^{N}X_i^{(n)}\sum_{n=1}^{N}X_j^{(n)},\sum_{n=1}^{N}X_k^{(n)}\sum_{n=1}^{N}X_l^{(n)}\right)=Cov\left(\sum_{n=1}^{N}\sum_{m=1}^{N}X_i^{(n)}X_j^{(m)},\sum_{q=1}^{N}\sum_{r=1}^{N}X_k^{(q)}X_l^{(r)}\right)=\sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{q=1}^{N}\sum_{r=1}^{N}Cov\left(X_i^{(n)}X_j^{(m)},X_k^{(q)}X_l^{(r)}\right).$$

The generic expression for the covariances in this sum is $Cov\left(X_i^{(n)}X_j^{(m)},X_k^{(q)}X_l^{(r)}\right)$. There are now four indices, $n,m,q,$ *and* r and therefore there are six equalities that may or may not be true: $m=n,q=n,q=m,r=n,r=m,$ *and* $r=q$. All possible truth values of these equalities are listed in Table A2, together with their count, the intermediate expression, and the final expression. For example, in the third row, we consider the case where $r=m$, while all other indices are unequal. There are $N(N-1)(N-2)$ such terms. Substitution of $r=m$ leads to $Cov\left(X_i^{(n)}X_j^{(m)},X_k^{(q)}X_l^{(r)}\right)=Cov\left(X_i^{(n)}X_j^{(m)},X_k^{(m)}X_l^{(r)}\right)$, and then proposition A1 leads to $Cov\left(X_i^{(n)}X_j^{(m)},X_k^{(m)}X_l^{(r)}\right)=$

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

$\mathbb{E}\left(X_i^{(n)}\right) Cov\left(X_j^{(m)}, X_k^{(m)}\right) \mathbb{E}\left(X_l^{(r)}\right) = Cov\left(X_j, X_l\right) \mathbb{E}\left(X_i\right) \mathbb{E}\left(X_k\right)$. In row 13 of Table A2, we used the following reasoning:

$$Cov\left(X_i^{(n)} X_j^{(m)}, X_k^{(m)} X_l^{(n)}\right) = \mathbb{E}\left(X_i^{(n)} X_j^{(m)} X_k^{(m)} X_l^{(n)}\right) - \mathbb{E}\left(X_i^{(n)} X_j^{(m)}\right) \mathbb{E}\left(X_k^{(m)} X_l^{(n)}\right) =$$

$$\mathbb{E}\left(X_i^{(n)} X_l^{(n)}\right) \mathbb{E}\left(X_j^{(m)} X_k^{(m)}\right) - \mathbb{E}\left(X_i^{(n)}\right) \mathbb{E}\left(X_j^{(m)}\right) \mathbb{E}\left(X_k^{(m)}\right) \mathbb{E}\left(X_l^{(n)}\right) =$$

$$\mathbb{E}\left(X_i X_l\right) \mathbb{E}\left(X_j X_k\right) - \mathbb{E}\left(X_i\right) \mathbb{E}\left(X_j\right) \mathbb{E}\left(X_k\right) \mathbb{E}\left(X_l\right)$$

A similar argument leads to the final expression in row 19 of Table A2.

Taking all four main terms together, we obtain:

$$\begin{aligned} Cov\left(C_{ij}^N, C_{kl}^N\right) = {} & N^{-2}\, NCov\left(X_i X_j, X_k X_l\right) \\ & - N^{-3}\left(N(N-1)\, Cov\left(X_i X_j, X_l\right) \mathbb{E}\left(X_k\right) + N(N-1)\, Cov\left(X_i X_j, X_k\right) \mathbb{E}\left(X_l\right) + NCov\left(X_i X_j, X_k X_l\right)\right) \\ & - N^{-3}\left(N(N-1)\, Cov\left(X_k X_l, X_j\right) \mathbb{E}\left(X_i\right) + N(N-1)\, Cov\left(X_k X_l, X_i\right) \mathbb{E}\left(X_j\right) + NCov\left(X_i X_j, X_k X_l\right)\right) + \\ & N^{-4} \left(\begin{array}{c} N(N-1)(N-2) \left(\begin{array}{c} Cov\left(X_j, X_l\right) \mathbb{E}\left(X_i\right) \mathbb{E}\left(X_k\right) + Cov\left(X_i, X_l\right) \mathbb{E}\left(X_j\right) \mathbb{E}\left(X_k\right) + \\ Cov\left(X_j, X_k\right) \mathbb{E}\left(X_i\right) \mathbb{E}\left(X_l\right) + Cov\left(X_i, X_k\right) \mathbb{E}\left(X_j\right) \mathbb{E}\left(X_l\right) \end{array} \right) \\ + N(N-1) \left(\begin{array}{c} Cov\left(X_k X_l, X_j\right) \mathbb{E}\left(X_i\right) + Cov\left(X_k X_l, X_i\right) \mathbb{E}\left(X_j\right) + \\ Cov\left(X_i X_j, X_l\right) \mathbb{E}\left(X_k\right) + Cov\left(X_i X_j, X_k\right) \mathbb{E}\left(X_l\right) \end{array} \right) + \\ N(N-1)\left(\mathbb{E}\left(X_i X_l\right) \mathbb{E}\left(X_j X_k\right) + \mathbb{E}\left(X_j X_l\right) \mathbb{E}\left(X_i X_k\right) - 2\mathbb{E}\left(X_i\right) \mathbb{E}\left(X_j\right) \mathbb{E}\left(X_k\right) \mathbb{E}\left(X_l\right)\right) + \\ NCov\left(X_i X_j, X_k X_l\right) \end{array} \right) \end{aligned}$$

A2. The covariance of two sums of conditional covariances with fixed weights

In the test proposed by Rosenbaum's (1984) case 2, the covariance of two items is considered conditionally on the sum of the other items. The statistics used by Rosenbaum is a standardization of a weighted sum of covariances, where the subgroup sizes act as weights. We will now develop a formula for the covariance of a weighted sum of conditional covariances. Let R_{ij} be the variable that is used for the conditioning of pair $\left(X_i, X_j\right)$. In the case of Rosenbaum's case 2, R_{ij} would be the sum of the other items, but for our purposes it is sufficient to assume that the range of R_{ij} is some finite set. The null hypothesis would now be that $Cov\left(X_i, X_j | R_{ij} = s\right) \geq 0$ for each s in the range of R_{ij}. Denote the corresponding sample covariance as

$$C_{ijs} = \widehat{cov}\left(X_i, X_j | R_{ij} = s\right).$$

Let N_{ijs} be the number of subjects in the subsample with $R_{ij} = s$. Rosenbaum's statistics is based on standardization of

$$C_{ij+} = \sum_s N_{ijs} C_{ijs}.$$

The goal of this section is to obtain a formula for $Cov\left(C_{ij+}, C_{kl+}\right)$. As a first step,

$$Cov\left(C_{ij+}, C_{kl+}\right) = \sum_s \sum_t Cov\left(N_{ijs} C_{ijs}, N_{klt} C_{klt}\right).$$

We will now first develop $Cov(C_{ijs}, C_{klt})$. In this section, we will assume that N_{ijs} and N_{klt} are fixed numbers instead of random variables. Then $\sum_s \sum_t Cov(N_{ijs} C_{ijs}, N_{klt} C_{klt}) = \sum_s \sum_t N_{ijs} N_{klt} Cov(C_{ijs}, C_{klt})$, so a formula for $Cov\left(C_{ijs}, C_{klt}\right)$ would solve the problem. In Rosenbaum's (1984) application, the N_{ijs} and N_{klt} are actually random variables, as discussed in the main text.

Let $I_{ijs}^{(n)} = 1\left[R_{ij}^{(n)} = s\right]$, the indicator function for the event $R_{ij} = s$ in subject n. That is, $1\left[R_{ij}^{(n)} = s\right] = 1$ if $R_{ij}^{(n)} = s$, and $1\left[R_{ij}^{(n)} = s\right] = 0$ otherwise, so that $N_{ijs} = \sum_{n=1}^N I_{ijs}^{(n)}$. Then

$$C_{ijs} = \frac{\sum_{n=1}^N X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}} - \frac{\sum_{n=1}^N X_i^{(n)} I_{ijs}^{(n)}}{N_{ijs}} \frac{\sum_{n=1}^N X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}}$$

$$C_{klt} = \frac{\sum_{n=1}^N X_k^{(n)} X_l^{(n)} I_{klt}^{(n)}}{N_{klt}} - \frac{\sum_{n=1}^N X_k^{(n)} I_{klt}^{(n)}}{N_{klt}} \frac{\sum_{n=1}^N X_l^{(n)} I_{klt}^{(n)}}{N_{klt}}$$

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table A2. *Development of the term* $\sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{q=1}^{N}\sum_{r=1}^{N} Cov\left(X_i^{(n)}X_j^{(m)},X_k^{(q)}X_l^{(r)}\right)$

	$m=n$	$q=n$	$q=m$	$r=n$	$r=m$	$r=q$	Count	Final expression
1	0	0	0	0	0	0	$N(N-1)(N-2)(N-3)$	0
2	0	0	0	0	0	1	$N(N-1)(N-2)$	0
3	0	0	0	0	1	0	$N(N-1)(N-2)$	$Cov\left(X_j,X_l\right)\mathbb{E}\left(X_i\right)\mathbb{E}\left(X_k\right)$
4	0	0	0	0	1	1	0	
5	0	0	0	1	0	0	$N(N-1)(N-2)$	$Cov\left(X_i,X_l\right)\mathbb{E}\left(X_j\right)\mathbb{E}\left(X_k\right)$
6	0	0	0	1	0	1	0	
7	0	0	0	1	1	0	0	
8	0	0	0	1	1	1	0	
9	0	0	1	0	0	0	$N(N-1)(N-2)$	$Cov\left(X_j,X_k\right)\mathbb{E}\left(X_i\right)\mathbb{E}\left(X_l\right)$
10	0	0	1	0	0	1	0	
11	0	0	1	0	1	0	0	
12	0	0	1	0	1	1	$N(N-1)$	$Cov\left(X_kX_l,X_j\right)\mathbb{E}\left(X_i\right)$
13	0	0	1	1	0	0	$N(N-1)$	$\mathbb{E}\left(X_iX_l\right)\mathbb{E}\left(X_jX_k\right)-\mathbb{E}\left(X_i\right)\mathbb{E}\left(X_j\right)\mathbb{E}\left(X_k\right)\mathbb{E}\left(X_l\right)$
14	0	0	1	1	0	1	0	
15	0	0	1	1	1	0	0	
16	0	0	1	1	1	1	0	
17	0	1	0	0	0	0	$N(N-1)(N-2)$	$Cov\left(X_i,X_k\right)\mathbb{E}\left(X_j\right)\mathbb{E}\left(X_l\right)$
18	0	1	0	0	0	1	0	
19	0	1	0	0	1	0	$N(N-1)$	$\mathbb{E}\left(X_jX_l\right)\mathbb{E}\left(X_iX_k\right)-\mathbb{E}\left(X_i\right)\mathbb{E}\left(X_j\right)\mathbb{E}\left(X_k\right)\mathbb{E}\left(X_l\right)$
20	0	1	0	0	1	1	0	
21	0	1	0	1	0	0	0	
22	0	1	0	1	0	1	$N(N-1)$	$Cov\left(X_kX_l,X_i\right)\mathbb{E}\left(X_j\right)$
23	0	1	0	1	1	0	0	
24	0	1	0	1	1	1	0	
25	0	1	1	0	0	0	0	
26	0	1	1	0	0	1	0	
27	0	1	1	0	1	0	0	
28	0	1	1	0	1	1	0	
29	0	1	1	1	0	0	0	
30	0	1	1	1	0	1	0	

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table A2. Continued

	$m=n$	$q=n$	$q=m$	$r=n$	$r=m$	$r=q$	Count	Final expression
31	0	1	1	1	1	0	0	
32	0	1	1	1	1	1	0	
33	1	0	0	0	0	0	$N(N-1)(N-2)$	0
34	1	0	0	0	0	1	$N(N-1)$	0
35	1	0	0	0	1	0	0	
36	1	0	0	0	1	1	0	
37	1	0	0	1	0	0	0	
38	1	0	0	1	0	1	0	
39	1	0	0	1	1	0	$N(N-1)$	$Cov\left(X_iX_j,X_l\right)\mathbb{E}\left(X_k\right)$
40	1	0	0	1	1	1	0	
41	1	0	1	0	0	0	0	
42	1	0	1	0	0	1	0	
43	1	0	1	0	1	0	0	
44	1	0	1	0	1	1	0	
45	1	0	1	1	0	0	0	
46	1	0	1	1	0	1	0	
47	1	0	1	1	1	0	0	
48	1	0	1	1	1	1	0	
49	1	1	0	0	0	0	0	
50	1	1	0	0	0	1	0	
51	1	1	0	0	1	0	0	
52	1	1	0	0	1	1	0	
53	1	1	0	1	0	0	0	
54	1	1	0	1	0	1	0	
55	1	1	0	1	1	0	0	
56	1	1	0	1	1	1	0	
57	1	1	1	0	0	0	$N(N-1)$	$Cov\left(X_iX_j,X_k\right)\mathbb{E}\left(X_l\right)$
58	1	1	1	0	0	1	0	
59	1	1	1	0	1	0	0	
60	1	1	1	0	1	1	0	
61	1	1	1	1	0	0	0	
62	1	1	1	1	0	1	0	
63	1	1	1	1	1	0	0	
64	1	1	1	1	1	1	N	$Cov\left(X_iX_j,X_kX_l\right)$

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Therefore $Cov\left(C_{ijs}, C_{klt}\right)$ can be developed analogously to the one-sample case of the previous section. We start with rewriting it into four main terms.

$$
\begin{aligned}
Cov\left(C_{ijs}, C_{klt}\right) = & Cov\left(\frac{\sum_{n=1}^{N} X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}}, \frac{\sum_{m=1}^{N} X_k^{(m)} X_l^{(m)} I_{klt}^{(m)}}{N_{klt}}\right) \\
& - Cov\left(\frac{\sum_{n=1}^{N} X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}}, \frac{\sum_{m=1}^{N} X_k^{(m)} I_{klt}^{(m)}}{N_{klt}} \frac{\sum_{r=1}^{N} X_l^{(r)} I_{klt}^{(r)}}{N_{klt}}\right) \\
& - Cov\left(\frac{\sum_{n=1}^{N} X_i^{(n)} I_{ijs}^{(n)}}{N_{ijs}} \frac{\sum_{n=1}^{N} X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}}, \frac{\sum_{n=1}^{N} X_k^{(n)} X_l^{(n)} I_{klt}^{(n)}}{N_{klt}}\right) \\
& + Cov\left(\frac{\sum_{n=1}^{N} X_i^{(n)} I_{ijs}^{(n)}}{N_{ijs}} \frac{\sum_{n=1}^{N} X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}}, \frac{\sum_{m=1}^{N} X_k^{(m)} I_{klt}^{(m)}}{N_{klt}} \frac{\sum_{r=1}^{N} X_l^{(r)} I_{klt}^{(r)}}{N_{klt}}\right) \\
= & \frac{1}{N_{ijs} N_{klt}} \sum_{n=1}^{N} \sum_{m=1}^{N} Cov\left(X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}, X_k^{(m)} X_l^{(m)} I_{klt}^{(m)}\right) \\
& - \frac{1}{N_{ijs} N_{klt}^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{r=1}^{N} Cov\left(X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}, X_k^{(m)} I_{klt}^{(m)} X_l^{(r)} I_{klt}^{(r)}\right) \\
& - \frac{1}{N_{ijs}^2 N_{klt}} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{r=1}^{N} Cov\left(X_k^{(n)} X_l^{(n)} I_{klt}^{(n)}, X_i^{(m)} I_{ijs}^{(m)} X_j^{(r)} I_{ijs}^{(r)}\right) \\
& + \frac{1}{N_{ijs}^2 N_{klt}^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{q=1}^{N} \sum_{r=1}^{N} Cov\left(X_i^{(n)} I_{ijs}^{(n)} X_j^{(m)} I_{ijs}^{(m)}, X_k^{(q)} I_{klt}^{(q)} X_l^{(r)} I_{klt}^{(r)}\right).
\end{aligned}
$$

Using the results of the previous section, we can express these terms as covariance of the variables $X_i I_{ijs}$, $X_j I_{ijs}$, $X_k I_{klt}$, and $X_l I_{klt}$ and their products. The first main term is

$$
\begin{aligned}
Cov\left(\frac{\sum_{n=1}^{N} X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}}, \frac{\sum_{m=1}^{N} X_k^{(m)} X_l^{(m)} I_{klt}^{(m)}}{N_{klt}}\right) &= \frac{1}{N_{ijs} N_{klt}} \sum_{n=1}^{N} \sum_{m=1}^{N} Cov\left(X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}, X_k^{(m)} X_l^{(m)} I_{klt}^{(m)}\right) \\
&= \frac{N}{N_{ijs} N_{klt}} Cov\left(X_i X_j I_{ijs}, X_k X_l I_{klt}\right).
\end{aligned}
$$

The second main term is

$$
\begin{aligned}
-Cov\left(\frac{\sum_{n=1}^{N} X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}}{N_{ijs}}, \frac{\sum_{m=1}^{N} X_k^{(m)} I_{klt}^{(m)}}{N_{klt}} \frac{\sum_{r=1}^{N} X_l^{(r)} I_{klt}^{(r)}}{N_{klt}}\right) &= -\frac{1}{N_{ijs} N_{klt}^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{r=1}^{N} Cov\left(X_i^{(n)} X_j^{(n)} I_{ijs}^{(n)}, X_k^{(m)} I_{klt}^{(m)} X_l^{(r)} I_{klt}^{(r)}\right) \\
&= -\frac{1}{N_{ijs} N_{klt}^2} \left(N(N-1) Cov\left(X_i X_j I_{ijs}, X_l I_{klt}\right) \mathbb{E}\left(X_k I_{klt}\right)\right. \\
&\quad + N(N-1) Cov\left(X_i X_j I_{ijs}, X_k I_{klt}\right) \mathbb{E}\left(X_l I_{klt}\right) \\
&\quad \left. + N Cov\left(X_i X_j I_{ijs}, X_k X_l I_{klt}\right)\right).
\end{aligned}
$$

This is developed in Table A3.

The third main term is analogously,

$$
\begin{aligned}
& - \frac{1}{N_{ijs}^2 N_{klt}} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{r=1}^{N} Cov\left(X_k^{(n)} X_l^{(n)} I_{klt}^{(n)}, X_i^{(m)} I_{ijs}^{(m)} X_j^{(r)} I_{ijs}^{(r)}\right) \\
= & - \frac{1}{N_{ijs}^2 N_{klt}} \left(N(N-1) Cov\left(X_k X_l I_{klt}, X_j I_{ijs}\right) \mathbb{E}\left(X_i I_{ijs}\right)\right. \\
& \left. + N(N-1) Cov\left(X_k X_l I_{klt}, X_i I_{ijs}\right) \mathbb{E}\left(X_j I_{ijs}\right) + N Cov\left(X_i X_j I_{ijs}, X_k X_l I_{klt}\right)\right).
\end{aligned}
$$

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table A3. *Development of the term* $\sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{r=1}^{N} Cov\left(X_i^{(n)}X_j^{(n)}I_{ijs}^{(n)},X_k^{(m)}I_{klt}^{(m)}X_l^{(r)}I_{klt}^{(r)}\right)$

$n=m$	$n=r$	$m=r$	Count	Intermediate term	Final term
0	0	0	$N(N-1)(N-2)$	$Cov\left(X_i^{(n)}X_j^{(n)}I_{ijs}^{(n)},X_k^{(m)}I_{klt}^{(m)}X_l^{(r)}I_{klt}^{(r)}\right)$	0
0	0	1	$N(N-1)$	$Cov\left(X_i^{(n)}X_j^{(n)}I_{ijs}^{(n)},X_k^{(m)}I_{klt}^{(m)}X_l^{(m)}I_{klt}^{(m)}\right)$	0
0	1	0	$N(N-1)$	$Cov\left(X_i^{(n)}X_j^{(n)}I_{ijs}^{(n)},X_k^{(m)}I_{klt}^{(m)}X_l^{(n)}I_{klt}^{(n)}\right)$	$Cov\left(X_iX_jI_{ijs},X_lI_{klt}\right)\mathbb{E}\left(X_kI_{klt}\right)$
0	1	1	0		
1	0	0	$N(N-1)$	$Cov\left(X_i^{(n)}X_j^{(n)}I_{ijs}^{(n)},X_k^{(n)}I_{klt}^{(n)}X_l^{(r)}I_{klt}^{(r)}\right)$	$Cov\left(X_iX_jI_{ijs},X_kI_{klt}\right)\mathbb{E}\left(X_lI_{klt}\right)$
1	0	1	0		
1	1	0	0		
1	1	1	N	$Cov\left(X_i^{(n)}X_j^{(n)}I_{ijs}^{(n)},X_k^{(n)}X_l^{(n)}I_{klt}^{(n)}\right)$	$Cov\left(X_iX_jI_{ijs},X_kX_lI_{klt}\right)$

The fourth main term is

$$\frac{1}{N_{ijs}^2N_{klt}^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{q=1}^{N}\sum_{r=1}^{N}Cov\left(X_i^{(n)}I_{ijs}^{(n)}X_j^{(m)}I_{ijs}^{(m)},X_k^{(q)}I_{klt}^{(q)}X_l^{(r)}I_{klt}^{(r)}\right)$$

$$=\frac{1}{N_{ijs}^2N_{klt}^2}\left(\begin{array}{c} N(N-1)(N-2)\left(\begin{array}{c} Cov\left(X_jI_{ijs},X_lI_{klt}\right)\mathbb{E}\left(X_iI_{ijs}\right)\mathbb{E}\left(X_kI_{klt}\right)\\ +Cov\left(X_iI_{ijs},X_lI_{klt}\right)\mathbb{E}\left(X_jI_{ijs}\right)\mathbb{E}\left(X_kI_{klt}\right)\\ +Cov\left(X_jI_{ijs},X_kI_{klt}\right)\mathbb{E}\left(X_iI_{ijs}\right)\mathbb{E}\left(X_lI_{klt}\right)\\ +Cov\left(X_iI_{ijs},X_kI_{klt}\right)\mathbb{E}\left(X_jI_{ijs}\right)\mathbb{E}\left(X_lI_{klt}\right)\end{array}\right)\\ +\\ N(N-1)\left(\begin{array}{c} Cov\left(X_kX_lI_{klt},X_jI_{ijs}\right)\mathbb{E}\left(X_iI_{ijs}\right)\\ +Cov\left(X_kX_lI_{klt},X_iI_{ijs}\right)\mathbb{E}\left(X_jI_{ijs}\right)\\ +Cov\left(X_iX_jI_{ijs},X_lI_{klt}\right)\mathbb{E}\left(X_kI_{klt}\right)\\ +Cov\left(X_iX_jI_{ijs},X_kI_{klt}\right)\mathbb{E}\left(X_lI_{klt}\right)\end{array}\right)\\ +\\ N(N-1)\left(\begin{array}{c} \mathbb{E}\left(X_iI_{ijs}X_lI_{klt}\right)\mathbb{E}\left(X_jI_{ijs}X_kI_{klt}\right)+\mathbb{E}\left(X_jI_{ijs}X_lI_{klt}\right)\mathbb{E}\left(X_iI_{ijs}X_kI_{klt}\right)\\ -2\mathbb{E}\left(X_iI_{ijs}\right)\mathbb{E}\left(X_jI_{ijs}\right)\mathbb{E}\left(X_kI_{klt}\right)\mathbb{E}\left(X_lI_{klt}\right)\end{array}\right)\\ +\\ NCov\left(X_iX_jI_{ijs},X_kX_lI_{klt}\right)\end{array}\right)$$

This is developed in Table A4.

Let us now summarize the results. For arbitrary variables V,W,X,Y write

$$\gamma_1(V,W,X,Y)=Cov(VW,XY)$$

$$\gamma_2(V,W,X,Y)=Cov(VW,Y)\mathbb{E}(X)+Cov(VW,X)\mathbb{E}(Y)$$

$$\gamma_3(V,W,X,Y)=\gamma_2(X,Y,V,W)=Cov(XY,W)\mathbb{E}(V)+Cov(XY,V)\mathbb{E}(W)$$

$$\gamma_4(V,W,X,Y)=Cov(V,X)\mathbb{E}(W)\mathbb{E}(Y)+Cov(V,Y)\mathbb{E}(W)\mathbb{E}(X)+Cov(W,X)\mathbb{E}(V)\mathbb{E}(Y)+Cov(W,Y)\mathbb{E}(V)\mathbb{E}(X)$$

$$\begin{aligned}\gamma_5(V,W,X,Y)&=Cov(VW,X)\mathbb{E}(Y)+Cov(VW,Y)\mathbb{E}(X)+Cov(V,XY)\mathbb{E}(W)+Cov(W,XY)\mathbb{E}(V)\\&=\gamma_2(V,W,X,Y)+\gamma_3(V,W,X,Y)\end{aligned}$$

$$\gamma_6(V,W,X,Y)=\mathbb{E}(VY)\mathbb{E}(WX)-\mathbb{E}(V)\mathbb{E}(W)\mathbb{E}(X)\mathbb{E}(Y)+\mathbb{E}(WY)\mathbb{E}(VX)-\mathbb{E}(V)\mathbb{E}(W)\mathbb{E}(X)\mathbb{E}(Y)$$

and

$$T_1(V,W,X,Y,N)=N\gamma_1(V,W,X,Y)$$

$$T_2(V,W,X,Y,N)=N(N-1)\gamma_2(V,W,X,Y)+N\gamma_1(V,W,X,Y)$$

$$T_3(V,W,X,Y,N)=N(N-1)\gamma_3(V,W,X,Y)+N\gamma_1(V,W,X,Y)$$

$$\begin{aligned}T_4(V,W,X,Y,N)&=N(N-1)(N-2)\gamma_4(V,W,X,Y)\\&\quad+N(N-1)\gamma_5(V,W,X,Y)+N(N-1)\gamma_6(V,W,X,Y)+N\gamma_1(V,W,X,Y).\end{aligned}$$

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table A4. *Development of* $\sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{q=1}^{N}\sum_{r=1}^{N} Cov\left(X_i^{(n)}I_{ijs}^{(n)}X_j^{(m)}I_{ijs}^{(m)},X_k^{(q)}I_{klt}^{(q)}X_l^{(r)}I_{klt}^{(r)}\right)$

	$m=n$	$q=n$	$q=m$	$r=n$	$r=m$	$r=q$	Count	Final
1	0	0	0	0	0	0	$N(N-1)(N-2)(N-3)$	0
2	0	0	0	0	0	1	$N(N-1)(N-2)$	0
3	0	0	0	0	1	0	$N(N-1)(N-2)$	$Cov\left(X_jI_{ijs},X_lI_{klt}\right)\mathbb{E}\left(X_iI_{ijs}\right)\mathbb{E}\left(X_kI_{klt}\right)$
4	0	0	0	0	1	1	0	
5	0	0	0	1	0	0	$N(N-1)(N-2)$	$Cov\left(X_iI_{ijs},X_lI_{klt}\right)\mathbb{E}\left(X_jI_{ijs}\right)\mathbb{E}\left(X_kI_{klt}\right)$
6	0	0	0	1	0	1	0	
7	0	0	0	1	1	0	0	
8	0	0	0	1	1	1	0	
9	0	0	1	0	0	0	$N(N-1)(N-2)$	$Cov\left(X_jI_{ijs},X_kI_{klt}\right)\mathbb{E}\left(X_iI_{ijs}\right)\mathbb{E}\left(X_lI_{klt}\right)$
10	0	0	1	0	0	1	0	
11	0	0	1	0	1	0	0	
12	0	0	1	0	1	1	$N(N-1)$	$Cov\left(X_kX_lI_{klt},X_jI_{ijs}\right)\mathbb{E}\left(X_iI_{ijs}\right)$
13	0	0	1	1	0	0	$N(N-1)$	$\mathbb{E}\left(X_jI_{ijs}X_kI_{klt}\right)\mathbb{E}\left(X_iI_{ijs}X_lI_{klt}\right)-\mathbb{E}\left(X_iI_{ijs}\right)\mathbb{E}\left(X_jI_{ijs}\right)\mathbb{E}\left(X_kI_{klt}\right)\mathbb{E}\left(X_lI_{klt}\right)$
14	0	0	1	1	0	1	0	
15	0	0	1	1	1	0	0	
16	0	0	1	1	1	1	0	
17	0	1	0	0	0	0	$N(N-1)(N-2)$	$Cov\left(X_iI_{ijs},X_kI_{klt}\right)\mathbb{E}\left(X_jI_{ijs}\right)\mathbb{E}\left(X_lI_{klt}\right)$
18	0	1	0	0	0	1	0	
19	0	1	0	0	1	0	$N(N-1)$	$\mathbb{E}\left(X_jI_{ijs}X_lI_{klt}\right)\mathbb{E}\left(X_iI_{ijs}X_kI_{klt}\right)-\mathbb{E}\left(X_iI_{ijs}\right)\mathbb{E}\left(X_jI_{ijs}\right)\mathbb{E}\left(X_kI_{klt}\right)\mathbb{E}\left(X_lI_{klt}\right)$
20	0	1	0	0	1	1	0	
21	0	1	0	1	0	0	0	
22	0	1	0	1	0	1	$N(N-1)$	$Cov\left(X_kX_lI_{klt},X_iI_{ijs}\right)\mathbb{E}\left(X_jI_{ijs}\right)$
23	0	1	0	1	1	0	0	
24	0	1	0	1	1	1	0	
25	0	1	1	0	0	0	0	
26	0	1	1	0	0	1	0	
27	0	1	1	0	1	0	0	
28	0	1	1	0	1	1	0	

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Table A4. Continued

	$m=n$	$q=n$	$q=m$	$r=n$	$r=m$	$r=q$	Count	Final
29	0	1	1	1	0	0	0	
30	0	1	1	1	0	1	0	
31	0	1	1	1	1	0	0	
32	0	1	1	1	1	1	0	
33	1	0	0	0	0	0	$N(N-1)(N-2)$	0
34	1	0	0	0	0	1	$N(N-1)$	0
35	1	0	0	0	1	0	0	
36	1	0	0	0	1	1	0	
37	1	0	0	1	0	0	0	
38	1	0	0	1	0	1	0	
39	1	0	0	1	1	0	$N(N-1)$	$Cov\left(X_i I_{ijs} X_j, X_l I_{klt}\right)\mathbb{E}\left(X_k I_{klt}\right)$
40	1	0	0	1	1	1	0	
41	1	0	1	0	0	0	0	
42	1	0	1	0	0	1	0	
43	1	0	1	0	1	0	0	
44	1	0	1	0	1	1	0	
45	1	0	1	1	0	0	0	
46	1	0	1	1	0	1	0	
47	1	0	1	1	1	0	0	
48	1	0	1	1	1	1	0	
49	1	1	0	0	0	0	0	
50	1	1	0	0	0	1	0	
51	1	1	0	0	1	0	0	
52	1	1	0	0	1	1	0	
53	1	1	0	1	0	0	0	
54	1	1	0	1	0	1	0	
55	1	1	0	1	1	0	0	
56	1	1	0	1	1	1	0	
57	1	1	1	0	0	0	$N(N-1)$	$Cov\left(X_i X_j I_{ijs}, X_k I_{klt}\right)\mathbb{E}\left(X_l I_{klt}\right)$
58	1	1	1	0	0	1	0	
59	1	1	1	0	1	0	0	
60	1	1	1	0	1	1	0	
61	1	1	1	1	0	0	0	
62	1	1	1	1	0	1	0	
63	1	1	1	1	1	0	0	
64	1	1	1	1	1	1	N	$Cov\left(X_i X_j I_{ijs}, X_k X_l I_{klt}\right)$

https://doi.org/10.1017/psy.2024.21 Published online by Cambridge University Press

Then the result of the previous section can be written as

$$Cov\left(C_{VW}^{N}, C_{XY}^{N}\right) = N^{-2}T_1\left(V,W,X,Y,N\right) - N^{-3}T_2\left(V,W,X,Y,N\right) - N^{-3}T_3\left(V,W,X,Y,N\right) + N^{-4}T_4\left(V,W,X,Y,N\right)$$

The result of the present section can be obtained with $V = X_i I_{ijs}$, $W = X_j I_{ijs}$, $X = X_k I_{klt}$, $Y = X_l I_{klt}$:

$$\begin{aligned} Cov\left(C_{ijs}, C_{klt}\right) = & \frac{1}{N_{ijs}N_{klt}} T_1\left(X_i I_{ijs}, X_j I_{ijs}, X_k I_{klt}, X_l I_{klt}, N\right) \\ & - \frac{1}{N_{ijs}N_{klt}^2} T_2\left(X_i I_{ijs}, X_j I_{ijs}, X_k I_{klt}, X_l I_{klt}, N\right) \\ & - \frac{1}{N_{ijs}^2 N_{klt}} T_3\left(X_i I_{ijs}, X_j I_{ijs}, X_k I_{klt}, X_l I_{klt}, N\right) \\ & + \frac{1}{N_{ijs}^2 N_{klt}^2} T_4\left(X_i I_{ijs}, X_j I_{ijs}, X_k I_{klt}, X_l I_{klt}, N\right). \end{aligned}$$

Thus

$$\begin{aligned} Cov\left(N_{ijs}C_{ijs}, N_{klt}C_{klt}\right) = & T_1\left(X_i I_{ijs}, X_j I_{ijs}, X_k I_{klt}, X_l I_{klt}, N\right) \\ & - \frac{1}{N_{klt}} T_2\left(X_i I_{ijs}, X_j I_{ijs}, X_k I_{klt}, X_l I_{klt}, N\right) \\ & - \frac{1}{N_{ijs}} T_3\left(X_i I_{ijs}, X_j I_{ijs}, X_k I_{klt}, X_l I_{klt}, N\right) \\ & + \frac{1}{N_{ijs}N_{klt}} T_4\left(X_i I_{ijs}, X_j I_{ijs}, X_k I_{klt}, X_l I_{klt}, N\right). \end{aligned}$$

Cite this article: Ellis, J. L., van der Ark, L. A. and Sijtsma, K. (2025). An Overall Test of Pairwise Mean Conditional Covariances in IRT. *Psychometrika*, 384–414. https://doi.org/10.1017/psy.2024.21