Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Detection of Cognitive Diagnostic Model Misspecification using New Lancaster-Chesher Information Matrix Tests

Richard M. Golden [*] and Reyhaneh Hosseinpourkhoshkbari

COINS Lab, School of Behavioral and Brain Sciences, University of Texas at Dallas, 75080, USA
[*]Corresponding author. Email: golden@utdallas.edu

Abstract

Model specification tests play a crucial role in evaluating the appropriateness of probability models for estimation and inference. Existing methods for the detection of model misspecification such as the chi-square goodness-of-fit (GOF) tests and more recently the M2 statistic (Maydeu-Olivares & Joe, 2014; Maydeu-Olivares & Joe, 2005) tend to result in test statistics with excessive degrees of freedom for models with larger numbers of parameters. An alternative approach is based upon the Information Matrix (IM) equality. The IM equality asserts that if a probability model is correctly specified, the asymptotic covariance matrix of the maximum likelihood estimators can be asymptotically estimated using a methodology based upon either the first or second derivatives of the log-likelihood function. Using a contrapositive argument, White (1982) proposed a misspecification test methodology based upon comparing these two alternative covariance matrix estimators. Extending this work, Presnell and Boos (2004) showed how to develop a misspecification test which only requires one degree of freedom regardless of the complexity of the model or data. In this paper, we extend prior work and additionally apply methods of Golden et al. (2013, 2016) to derive and evaluate misspecification tests for Cognitive Diagnostic Models (CDMs) which only require 1 or 2 degrees of freedom regardless of model or data complexity. Analytic formulas for the tests are derived so they can be applied without requiring computationally intensive bootstrap simulation methods. Our simulation studies show the asymptotic statistical tests have good level (type 1 error) and power performance for CDM models and data which might be encountered in practice.

Keywords: misspecification, information matrix test, cognitive diagnostic model

1. Introduction

A correctly specified probability model is capable of perfectly representing the data generating process which generates observed data. The detection of model misspecification is critical for evaluating the validity of model-based measurement processes. When model misspecification is present, the interpretation of parameter estimates and the validity of inferences becomes questionable (Hausman, 1978). Therefore, methods for the assessment of the presence of model misspecification are desirable.

A variety of methods have been developed for the purpose of the assessment of model misspecification. For example, graphical residual diagnostic methods (Fox & Weisberg, 2018) are an important tool but do not provide an explicit procedure for supporting a hypothesis-driven decision process designed to determine whether or not a model should be classified as misspecified. Chi-square goodness-of-fit (GOF) tests (e.g., Hosmer et al., 1997) are commonly utilized for probability models with categorical response variables. Such tests compare the predicted probabilities of response categories to observed response frequencies in the observed data but such tests are not applicable

to continuous response variables. Moreover, the degrees of freedom for the chi-square GOF test statistic increase as an exponential function of the number of response categories. An extension of the chi-square GOF test statistic provides an improvement in statistical power by only examining the first and second moments of the categorical response distribution resulting in a chi-square GOF test statistic whose degrees of freedom increase as a quadratic function of the number of response categories. In practice, the number of response categories is relatively large so both of these methodologies face the challenge of test statistics with large variance which would be expected to have poor statistical power.

Using a different approach, White (1982, 1994) and Golden et al. (2013, 2016) introduced a misspecification detection methodology which, unlike chi-square GOF tests and graphical residual diagnostic methods, is based upon testing the Information Matrix (IM) equality. The IM equality states that the asymptotic covariance matrix of the maximum likelihood estimates can be estimated using either the Hessian or Outer Product Gradient (OPG) covariance matrix estimators when the probability model is correctly specified. Using a contrapositive argument, White (1982) noted that if the Hessian and OPG covariance matrix estimators are asymptotically different, then this is an indication of the presence of model misspecification.

Assume a Data Generating Process (DGP) exists which consists of n independent and identically distributed d-dimensional random vectors (observations) $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ with common probability density $p_e(\mathbf{x})$. The observed data sample is a realization of $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ which will be denoted as $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let $p(\mathbf{x}|\boldsymbol{\theta})$ denote a proposed probability model specification for the DGP. If the probability model $\mathcal{M} \equiv \{p(\mathbf{x}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{R}^q\}$ is correctly specified, then there exists a q-dimensional parameter vector $\boldsymbol{\theta}$ such that $p(\mathbf{x}|\boldsymbol{\theta}) = p_e(\mathbf{x})$ almost everywhere[1].

The maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}_n$ is defined as a strict local minimizer of the negative normalized log-likelihood function $\hat{\ell}_n(\boldsymbol{\theta}) \equiv -(1/n)\sum_{i=1}^n \log p(\mathbf{x}_i|\boldsymbol{\theta})$. The Hessian covariance matrix estimator, $\hat{\mathbf{A}}_n$, is defined as $\hat{\mathbf{A}}_n \equiv (1/n)\sum_{i=1}^n \hat{\mathbf{A}}^i$ where $\hat{\mathbf{A}}^i$ is second derivative of $-\log p(\mathbf{x}_i|\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}_n$. The OPG covariance matrix estimator, $\hat{\mathbf{B}}_n$, is defined as the inverse of the OPG matrix $(1/n)\sum_{i=1}^n \hat{\mathbf{g}}^i[\hat{\mathbf{g}}^i]^T$ where $\hat{\mathbf{g}}^i$ is the transpose of the gradient of $-\log p(\mathbf{x}_i|\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}_n$. Given appropriate regularity conditions, it can be shown that $\hat{\mathbf{A}}_n \to \mathbf{A}^*$ and $\hat{\mathbf{B}}_n \to \mathbf{B}^*$ with probability one as $n \to \infty$. If $\mathbf{A}^* \neq \mathbf{B}^*$, this implies the presence of model misspecification by the IM equality (White, 1982, 1987, 1994; Golden, 2020).

Presnell and Boos (2004) proposed that the comparison of $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ could be achieved by computing the "in-and-out-of-sample" IOS test statistic s_{IOS}. Let $\boldsymbol{\theta}_n^-(k)$ be the MLE obtained when the kth observation $\mathbf{x}_k$ is deleted from the data sample. The negative normalized log-likelihood evaluated at $\hat{\boldsymbol{\theta}}_n^-$ is denoted as $\hat{\ell}_n^-(\hat{\boldsymbol{\theta}}_n^-) = (k-1)^{-1}\sum_{i \neq k} \log p(\mathbf{x}_i|\boldsymbol{\theta}_n^-(i))$ and corresponds to an "out of sample" estimate of model fit. The IOS test statistic is then specified as:

$$\hat{s}_{ios} = -n\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) + n\hat{\ell}_n^-(\hat{\boldsymbol{\theta}}_n^-). \tag{1}$$

Presnell and Boos (2004) showed how this statistic is related to the IM testing methodology of White (1982, 1994) by proving that $\hat{s}_{ios}$ converges in probability to the quantity $\mathrm{Tr}([\mathbf{A}^*]^{-1}\mathbf{B}^*)$. If the model is correctly specified, then $\mathbf{A}^* = \mathbf{B}^*$ it then follows that $\mathrm{Tr}([\mathbf{A}^*]^{-1}\mathbf{B}^*) = q$. In addition, Presnell and Boos (2004) showed that the asymptotic distribution of the test statistic $\hat{s}_{ios}$ was Gaussian corresponding to a one degree of freedom chi-square statistic. Thus, Presnell and Boos (2004) provided the foundations of a one degree of freedom misspecification detection methodology for detecting model misspecification by testing the null hypothesis that $\hat{s}_{ios} \to q$.

Golden et al. (2013, 2016) proposed a generalization of the White (1982) Information Matrix testing framework termed the Generalized Information Matrix Test (GIMT) framework. The

[1] The "almost everywhere" condition is required when $\tilde{\mathbf{x}}_i$ is an absolutely continuous or mixed random vector but can be omitted when $\tilde{\mathbf{x}}_i$ is a discrete random vector.

fundamental concept of the GIMT framework involves comparing specific nonlinear functions of the Hessian and OPG covariance matrices. Thus, the previous work by White (1982) and Presnell and Boos (2004) can be viewed naturally as special cases of GIMTs. Golden et al. (2016) distinguishes between two types of GIMTs. Let H be a null hypothesis for a GIMT which means that if H is false then $\mathbf{A}^* \neq \mathbf{B}$. If H is formally equivalent to the null hypothesis that $\mathbf{A}^* = \mathbf{B}^*$ then H is called *non-directional* otherwise H is called *directional*.

In this paper we apply the GIMT theory developed by Golden et al. (2013, 2016) to develop and evaluate one directional GIMT and one non-directional GIMT for the detection of model misspecification. The first specification test is a directional GIMT which is closely related to the Presnell and Boos (2004) IOS test and the Golden et al. (2016) Robust Log GAIC GIMT. In particular, a test statistic for testing the null hypothesis that $\mathrm{Tr}([\mathbf{B}^*]^{-1}\mathbf{A}^*) = q$ is proposed. This test which is called the *Inverse Trace GAIC* (InvTraceGAIC) test is shown using the methods of Golden et al. (2016) to have a test statistic which is chi-square with only one degree of freedom regardless of model or data complexity. The second specification test is a non-directional GIMT whose null hypothesis is identical to the Composite Log GAIC GIMT described by Golden et al. (2016) but we will refer to this test as the *trace inverse-trace GAIC* (TrInvTr) in this paper.

In order to explicitly derive a GIMT using the theory of Golden et al. (2013, 2016), a specific probability model must be specified because the GIMT Theory requires both the first and second derivaties of the log-likelihood function. In this paper, the specified probability model is a Cognitive Diagnostic Model (Rupp and Templin, 2008; George and Robitzsch, 2015). Cognitive diagnostic models (CDMs) are a family of restricted latent class psychometric models designed to assess examinee skill mastery by incorporating prior knowledge of the relationship between latent skills and student responses. In addition, a CDM outputs the attribute distribution or the probability that an examinee will have a particular set of skills given the examinee's exam performance. Characteristics of latent skills in CDMs are defined through a Q-matrix that specifies which specific skills are relevant for answering a particular exam question. Furthermore, Q-matrix misspecification can affect parameter estimates and respondent classification accuracy (Rupp and Templin, 2008). In addition, the presence of model misspecification can affect semantic interpretation of the unobservable latent skills which is crucial since CDMs are used for diagnostic purposes.

In the first part of the paper, we provide our theoretical contribution which shows how to apply the theory of Golden et al. (2016) to develop these two new low degree of freedom model misspecification tests for CDMs. The theoretical contribution includes new explicit formulas which not only explicitly specify the new CDM-specific test statistics but also include estimates of their asymptotic covariance to support hypothesis testing.

In the second part of the paper, we evaluate our new GIMT misspecification tests in simulation studies involving a CDM in which misspecified and correctly specified CDMs are fit to a known DGP. The discrimination performance of the new GIMTs is reported using response operating curves (ROCs) and the performance of the analytic formulas in estimating Type 1 error probabilities ("p-values") is evaluated using p-value plots (Davidson and MacKinnon, 1998).

2. Theoretical Results

2.1 The New Generalized Information Matrix Tests

The null hypothesis of a Generalized Information Matrix Test (GIMT) is specified by

$$H_o : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_r \tag{2}$$

where the column vector-valued function $\mathbf{s}$ maps a q-dimensional matrix $\mathbf{A}$ and q-dimensional matrix $\mathbf{B}$ into an r-dimensional vector. The notation $\mathbf{0}_r$ denotes an r-dimensional vector of zeros.

Define the *selection test statistic* $\hat{\mathbf{s}}_n$ such that $\hat{\mathbf{s}}_n \equiv \mathbf{s}(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n)$. Let $\hat{\mathbf{C}}_s^n$ specify an estimator of the asymptotic covariance of $\hat{\mathbf{s}}_n$. To test the null hypothesis, the Wald test statistic

$$\hat{\mathcal{W}}_n = n\hat{\mathbf{s}}_n^T[\hat{\mathbf{C}}_s^n]^{-1}\hat{\mathbf{s}}_n \tag{3}$$

can be shown (Golden et al., 2016, Theorem 7) to have an asymptotic chi-square distribution with r degrees of freedom when H_o in 2 is true. When 2 is false, the power of the statistical test converges to one when the sample size n is sufficiently large. The assumptions of this theorem require that $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ are asymptotically positive definite.

Let the duplication matrix $\mathcal{D}_k : \mathcal{R}^{k(k+1)2} \rightarrow \mathcal{R}^{k^2}$ be defined such that $\mathcal{D}_k\mathbf{vech}(\mathbf{A}) = \mathbf{vec}(\mathbf{A})$ and the inverse duplication matrix $\mathcal{D}_k^\dagger$ be defined such that $\mathcal{D}_k^\dagger\mathbf{vec}(\mathbf{A}) = \mathbf{vech}(\mathbf{A})$. Let $\mathcal{D}_k^\otimes \equiv \mathbf{I}_2 \otimes \mathcal{D}_k$ where $\mathbf{I}_2$ is the two-dimensional identity matrix. Let $\mathcal{D}_k^{\otimes\dagger} = \mathbf{I}_2 \otimes \mathcal{D}_k^\dagger$.

Using the results of Theorem 6 and Theorem 8 of Golden et al. (2016), the asymptotic covariance matrix of $\hat{\mathbf{s}}_n$, $\hat{\mathbf{C}}_s^n$ is given by the formula:

$$\hat{\mathbf{C}}_s^n = (1/n)\sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_i\hat{\boldsymbol{\delta}}_i^T \tag{4}$$

where

$$\hat{\boldsymbol{\delta}}_i = \nabla\hat{\mathbf{s}}_n\mathcal{D}_k^\otimes\left(\hat{\mathbf{d}}_{i,n} - \ddot{\nabla}\hat{\mathbf{d}}_n[\hat{\mathbf{A}}_n]^{-1}\hat{\mathbf{g}}^i - \hat{\mathbf{d}}_n\right),\ \nabla\hat{\mathbf{s}}_n \equiv \left[\frac{d\mathbf{s}}{d\mathbf{A}}(\hat{\mathbf{A}}_n), \frac{d\mathbf{s}}{d\mathbf{B}}(\hat{\mathbf{B}}_n)\right], \tag{5}$$

$$\ddot{\nabla}\hat{\mathbf{d}}_n = \mathcal{D}_k^{\otimes\dagger}\begin{bmatrix} \frac{d\hat{\mathbf{B}}_n}{d\boldsymbol{\theta}} + n^{-1}\sum_{i=1}^{n} vec(\hat{\mathbf{A}}_n - \hat{\mathbf{B}}_n)\hat{\mathbf{g}}_i^T \\ \frac{d\hat{\mathbf{B}}_n}{d\boldsymbol{\theta}} \end{bmatrix}, \tag{6}$$

and

$$\frac{d\hat{\mathbf{B}}_n}{d\boldsymbol{\theta}} = n^{-1}\sum_{i=1}^{n}\left[\left(\hat{\mathbf{A}}_i \otimes \hat{\mathbf{g}}^i\right) + \left(\hat{\mathbf{g}}^i \otimes \hat{\mathbf{A}}_i\right)\right]. \tag{7}$$

Equations 4, 6, and 7 correspond to Equations 2, 3, and 4 in Golden et al. (2016) respectively. Equations 6 and 7 specify the Lancaster-Chesher formula (Lancaster, 1984) described in Golden et al. (2016).

2.1.1 Inverse-Trace GAIC

The Inverse-Trace GAIC GIMT is a statistical test for model misspecification which tests the null hypothesis

$$H_o : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = 0,\quad \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \log\left(q^{-1}[\mathrm{Tr}\left((\mathbf{B}^*)^{-1}\mathbf{A}^*\right)]\right). \tag{8}$$

When the null hypothesis in (8) is rejected, this implies that $\mathbf{A}^* \neq \mathbf{B}^*$ which indicates model misspecification by the IM Equality (White, 1982; Golden, 2020). This null hypothesis and its associated GIMT are referred to as directional (see Golden et al., 2016) since rejection of H_o implies $\mathbf{A}^* \neq \mathbf{B}^*$ but the converse of this statement does not necessarily hold. The Inverse-Trace GAIC null hypothesis is similar but not equivalent to the null hypotheses of the IOS statistical test (Presnell and Boos, 2004) and the Robust Log GAIC GIMT (Golden et al., 2016).

The selection statistic $\hat{\mathbf{s}}_n$ for the directional inverse-Trace GAIC test in 3 is defined by the formula:

$$\hat{\mathbf{s}}_n = \log\left(q^{-1}\,\mathrm{Tr}\left(\hat{\mathbf{B}}_n^{-1}\hat{\mathbf{A}}_n\right)\right). \tag{9}$$

Since $\hat{\mathbf{s}}_n$ is a scalar, the test statistic $\hat{\mathcal{W}}_n$ in 3 is a chi-square random variable with one degree of freedom under the null hypothesis by Theorem 7 of Golden et al. (2016).

To compute the covariance matrix $\hat{\mathbf{C}}_s^n$ in 3, the covariance matrix $\hat{\mathbf{C}}_n^s$ must be computed. Computation of the covariance matrix $\hat{\mathbf{C}}_n^s$ requires the derivative of $\hat{\mathbf{s}}_n \equiv \mathbf{s}(\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n)$. The derivative of $\mathbf{s}$ with respect to $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ for the directional inverse-trace GAIC in 12 is:

$$\nabla \mathbf{s}(\mathbf{A}, \mathbf{B}) = \left[\frac{d\mathbf{s}}{d\mathbf{A}}, \frac{d\mathbf{s}}{d\mathbf{B}}\right] \tag{10}$$

at $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ where

$$\frac{d\mathbf{s}}{d\mathbf{A}} = \left(\mathrm{Tr}\left(\mathbf{B}^{-1}\mathbf{A}\right)\right)^{-1}\left[\mathbf{vec}(\mathbf{B}^{-1})\right]^T \text{ and } \frac{d\mathbf{s}}{d\mathbf{B}} = -\left(\mathrm{Tr}\left(\mathbf{B}^{-1}\mathbf{A}\right)\right)^{-1}\left[\mathbf{vec}\left(\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}\right)\right]^T.$$

2.1.2 Trace Inverse-Trace GAIC

The Trace Inverse-Trace GAIC GIMT is a statistical test for model misspecification which tests the null hypothesis

$$H_o : \mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \mathbf{0}_2 \text{ where} \tag{11}$$

$$\mathbf{s}(\mathbf{A}^*, \mathbf{B}^*) = \left[\log\left(q^{-1}\,\mathrm{Tr}\left([\mathbf{A}^*]^{-1}\mathbf{B}^*\right)\right), \log\left(q^{-1}\,\mathrm{Tr}\left([\mathbf{B}^*]^{-1}\mathbf{A}^*\right)\right)\right]^T.$$

When the null hypothesis in 11 is rejected, this implies that $\mathbf{A}^* \neq \mathbf{B}^*$ which indicates model misspecification by the IM Equality (White, 1982; Golden, 2020). This null hypothesis and its associated GIMT are referred to as non-directional (see Golden et al., 2016) since it can be shown H_o holds if and only if $\mathbf{A}^* = \mathbf{B}^*$ (Cho and Phillips, 2018; also see Golden, 2020, Chapter 16, Theorem 16.3.2). The Trace Inverse-Trace GAIC null hypothesis is equivalent to the null hypothesis of the Composite Log GAIC GIMT (Golden et al., 2016) although the associated test statistic is slightly different.

In Equation 3, the selection statistic $\hat{\mathbf{s}}_n$ for the non-directional trace inverse-Trace GAIC test is defined by the formula:

$$\hat{\mathbf{s}}_n = \left[\log\left(q^{-1}\,\mathrm{Tr}\left(\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\right)\right), \log\left(q^{-1}\,\mathrm{Tr}\left(\hat{\mathbf{B}}_n^{-1}\hat{\mathbf{A}}_n\right)\right)\right]^T \tag{12}$$

Since $\hat{\mathbf{s}}_n$ is a two-dimensional vector, the test statistic $\hat{\mathbf{s}}_n$ is a chi-square random variable with two degrees of freedom under the null hypothesis by Theorem 7 of Golden et al. (2016).

2.2 Cognitive Diagnostic Probability Model

The Cognitive Diagnostic Probability model used in the mathematical derivations and simulation studies corresponds to the model specified by Hosseinpourkhoskbari and Golden (2024). For this model, the jth element of $\mathbf{x}_i$, x_{ij}, is the response of the ith student to the jth exam question which can take on the value of either zero or one.

In particular,

$$p(\mathbf{x}_i|\theta) = \sum_{\boldsymbol{\alpha}} p(\mathbf{x}_i|\boldsymbol{\alpha}, \theta)p(\boldsymbol{\alpha}) \tag{13}$$

where $\boldsymbol{\alpha}$ is a K-dimensional vector consisting of binary latent variables and the *proficiency model* $p(\boldsymbol{\alpha})$ is a known probability of a particular $\boldsymbol{\alpha}$. Let $\theta_{j,1}$ and $\theta_{j,2}$ be free parameters associated with the jth element of $\mathbf{x}_i$ respectively. Define $\theta = [\theta_{1,1}, \theta_{1,2}, \ldots, \theta_{q/2,1}, \theta_{q/2,2}]$. Let $q_{jk} = 1$ if the probability

distribution of the student's response to the jth item is functionally dependent upon the kth element of $\boldsymbol{\alpha}$ and let $q_{jk} = 0$ otherwise. The *evidence model* is defined as:

$$p(x_{ij} = 1|\boldsymbol{\alpha}, \boldsymbol{\theta}) = \left(1 + \exp[-\theta_{j,1}\psi_j(\boldsymbol{\alpha}) + \theta_{j,2}]\right)^{-1} \tag{14}$$

where ψ_j specifies a particular psychometric model for item j. For example, the DINA model for item j is specified by choosing: $\psi_j(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \alpha_k^{q_{jk}}$ (e.g., Rupp and Templin, 2008). The DINO model for item j is specified by choosing $\psi_j(\boldsymbol{\alpha}) = 1 - \prod_{k=1}^{k}(1 - \alpha_k)^{q_{jk}}$ (e.g., George and Robitzch, 2015).

3. Simulation Study

3.1 Data Set and Q matrix

The dataset extracted from Tatsuoka (1984) consists of a set of math problems designed to assess students' ability to solve problems involving fractions. In this case, we utilized the 'Fraction.1' dataset from the 'CDM' R package (George et al., 2016) containing the dichotomous responses of 536 middle school students over 15 fraction subtraction test items. The Q-matrix specifying which of five distinct skills were required to answer a particular test item was based upon the Q-matrix used by Tatsuoka (1984) and provided in George et al. (2016).

3.2 Methods

3.2.1 Methods: Type 1 Error Analytic Formula Performance

The CDM was fit to the original data set with a sample size of $n = 536$ by minimizing the negative normalized log-likelihood function. Then 500 simulated data sets each consisting of responses from $n = 134$ participants were generated from the fitted CDM. Next, the Lancaster-Chesher Wald test formulas whose derivations were presented in Section 2 were used to detect misspecification for the Inverse-Trace GAIC and Trace Inverse-Trace GAIC GIMTs for a wide range of significance levels. A p-value plot was then generated which plots the percentage of times the GIMT incorrectly rejected the null hypothesis of correct specification as a function of the significance level of the test.

3.2.2 Methods: Power Analyses using Alternative Models with Missing Skills

Three different ("misspecified") versions of the original fitted CDM used in the Type 1 Error Analyses (Section 3.2.1) were then defined by modifying the Q-matrix in Section 3.2.1 in three different ways. Specifically, the Q matrix in the original CDM used in the p-value analyses was modified by: (1) deleting skill 5, (2) deleting skills 2 and 4, and (3) deleting skills 1,3, and 4. These three different Q matrices corresponded to 3 alternative CDMs which will be referred to as "misspecified models" since they will be fit to data generated by the original CDM used in Section 3.2.1. The specific Q-matrix modifications were intentionally selected to represent different magnitudes of practically meaningful model misspecification. These changes were chosen intentionally to represent mild, moderate, and severe levels of misspecification.

In the original Q-matrix, Attribute 5 appears in only two items and plays a relatively small role in the assessment. Therefore, deleting Attribute 5 represents a mild form of misspecification. In contrast, Attributes 2 and 4 are mainly associated with the most complex items and are important for distinguishing higher-level mastery patterns. Removing both of these attributes leads to a much stronger distortion of the skill structure. The third case—deleting Attributes 1, 3, and 4—represents an intermediate level of misspecification which is less severe than deleting Attributes 2 and 4 but more severe than deleting Attribute 5.

Together, these three conditions create a spectrum of misspecification severity from weak to strong defined solely by the connectivity pattern of the Q matrix. This design allows us to evaluate

the performance of the GAIC-based tests in qualitatively distinct situations where an attribute is unintentionally omitted from the model.

The three misspecified CDMs and the original CDM used in the p-value analyses were then used to evaluate the discrimination performance of the Inverse-Trace GAIC and Trace Inverse-Trace GAIC Wald tests for misspecification detection. These results are reported using ROC curves. In particular, 500 pairs of simulated data samples of n = 134 data records were generated from the original CDM which had been fit to n = 536 data records.

The first member of each pair was fit to the original CDM ("correctly specified model") and the second member of each pair was fit to either the original CDM or a misspecified model. The two GIMTs were then used to classify the fitted model as correctly specified or misspecified for a wide range of significance levels.

These results were then reported using ROC curve plots where the true positive rate is plotted against the false positive rate for a range of different decision thresholds (i.e., significance levels). Additionally, the Area Under the ROC (AUROC) curve was computed which provides a measure of discrimination performance across decision thresholds.

3.3 Results and Discussion

3.3.1 Results: Type 1 Error Analytic Formula Performance

Figure 1 is a P-value plot which is designed to evaluate the performance of the two new GIMTs for controlling for Type 1 errors for various choices of significance levels. Figure 1 plots the Type 1 error of the Wald test in 3 as a function of the significance level. Given a sufficiently large sample size and a sufficiently large number of simulated bootstrap data sets, the Bootstrap and Theoretical P-values should agree.

Both the Inverse Trace GAIC GIMT and Trace Inverse-Trace GAIC GIMT provide reasonable estimates of the Type 1 error performance within the range of significance levels typically used in practice ($p < 0.1$). For the range of $p = 0.1$ through $p = 0.2$ the two-degree of freedom Trace Inverse-Trace GAIC GIMT formulas for computing p-values using 3 which have good agreement with the mathematical theory. Although the one-degree of freedom Inverse Trace GAIC GIMT deviates from its theoretical p-value predictions in the range of $p = 0.1$ through $p = 0.2$, the estimated p-values remain less than the desired Type 1 error cutoff (i.e., significance level) so the Type 1 error rate still remains correctly controlled at the desired significance level.

3.3.2 Results: Power Analysis using Alternative Models with Missing Skills

ROC curves generated using the methodology in Section 3.2.2 are presented in Figure 2 and Figure 3 respectively. In the presence of severe model misspecification associated with the removal of key skill 2 and key skill 4, both the Inverse-Trace GAIC and Trace Inverse-Trace GAIC show good misspecification detection performance (AUROC > 0.99). Deleting the skills 1, 3, and 4 corresponded to a situation where the Inverse-Trace GAIC GIMT (AUROC = 0.82) and Trace Inverse-Trace GAIC GIMT (AUROC = 0.77) were less effective at misspecification detection.

When the power analysis was performed by using the correctly specified model (no deleted skills) as the alternative hypothesis, then misspecification performance was at chance levels (AUROC = 0.5) for both GIMTs in Figures 2 and 3. In addition, chance discrimination performance (AUROC = 0.6) was obtained for both GIMTs when the alternative misspecified model was generated by deleting skill 5. Inspection of the Q matrix of the original CDM which generated the data showed that skill 5 was connected to fewer items.

4. General Discussion

Two entirely new tests for model misspecification for CDMs which have never been previously proposed or published in the existing applied statistics or psychometric literature were derived. These

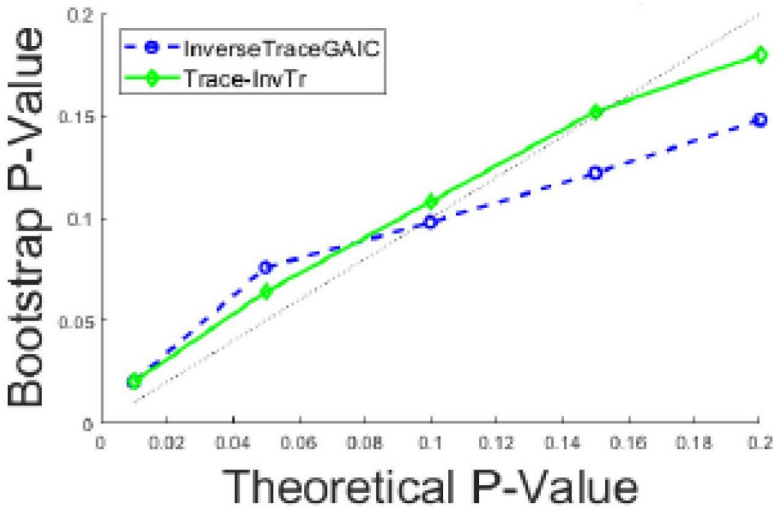

Figure 1. Simulated Bootstrap P-Values versus Analytic Formula Theoretical P-Values. The P-value plot shows that the Type 1 error rate is estimated by a bootstrap simulation resampling method (Bootstrap P-Value) has good agreement with the p-value theoretical formulas (which do not require bootstrap simulations) for typically used p-values (e.g., p = 0.05).

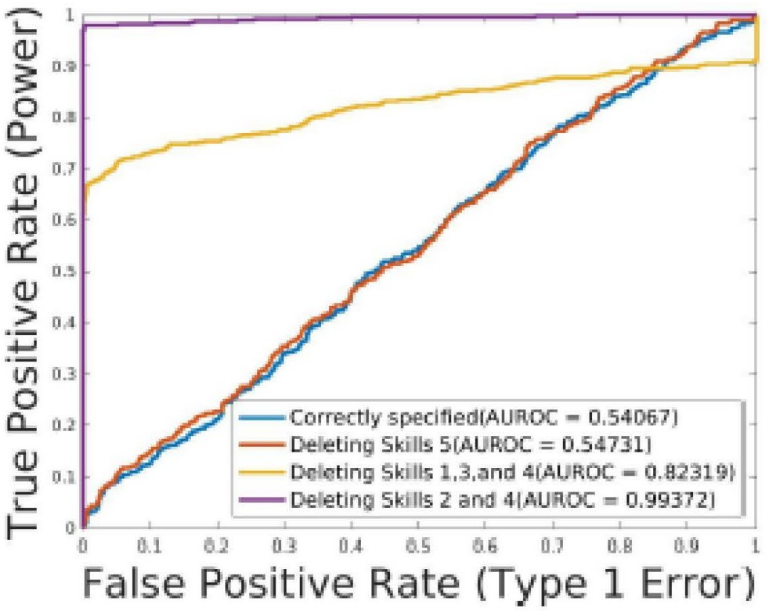

Figure 2. ROC Curves for investigating Inverse-Trace GAIC GIMT discrimination performance using alternative misspecified models with deleted skills. Discrimination performance was excellent when the alternative model involved deleting key skills 2 and 4. Moderate discrimination performance was obtained by deleting skills 1, 3, and 4. The Inverse Trace GAIC GIMT could not discriminate between the correctly specified model and a misspecified model generated by deleting the less relevant skill 5.

two new misspecification tests are Wald tests whose test statistics have a chi-square distribution with only 1 or 2 degrees of freedom regardless of data set size or model complexity. The Trace Inverse-Trace GAIC GIMT, in particular, is associated with a 2 degree of freedom test statistic and is a non-directional GIMT.

Next, the performance of these two new GIMTs were investigated in a series of simulation studies. First, the performance of the new analytic formulas for estimating p-values and reported as p-value plots. Second, the performance of the new analytic formulas for classifying models as misspecified or correctly specified was investigated and reported using ROC curves.

The new statistical tests showed good level (Type 1 error estimation) performance and power performance for a sample size of $n = 134$. Future work will futher investigate how the asymptotic behavior of the performance of the CDM GIMTs changes as sample size increases. Finally, we have focused on omnibus tests of model misspecification but we note this methodology can potentially be applied to new local diagnostic misspecification tests for individual task models such as 14 to inform model refinement.

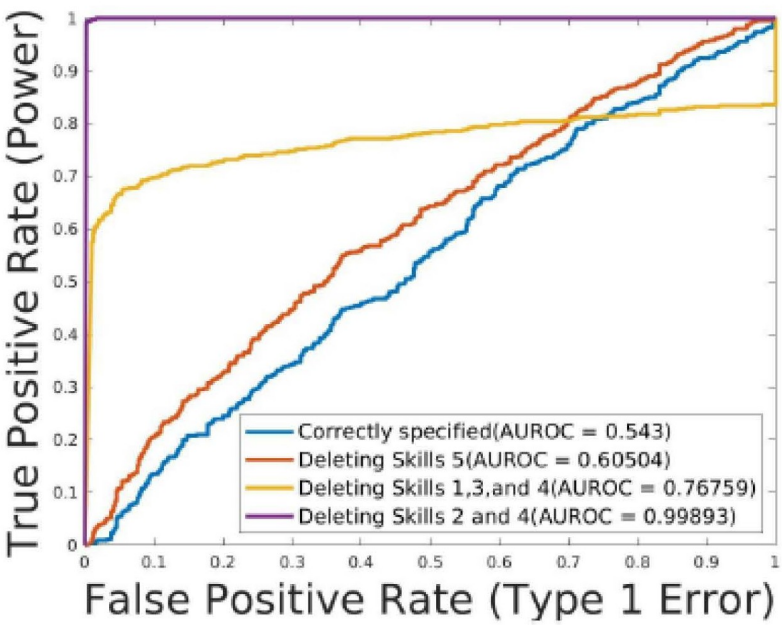

Figure 3. ROC Curves for investing the Trace Inverse-Trace GAIC GIMT discrimination performance using alternative misspecified models with deleted skills. Discrimination performance was excellent when the alternative model involved deleting key skills 2 and 4. Moderate discrimination performance was obtained by deleting skills 1, 3, and 4. The Inverse Trace GAIC GIMT could not discriminate between the correctly specified model and a misspecified model generated by deleting the less relevant skill 5.

Notes

Author Contributions: Both authors contributed equally to the theoretical work, empirical work, simulation work, and manuscript preparation. We thank Brian Junker for feedback on an earlier

version of the manuscript and his comment that GIMT methods for informing model refinement would be quite useful.

Competing Interests: None

References

Cho, J. S., & Phillips, P. C. (2018). Pythagorean generalization of testing the equality of two symmetric positive definite matrices. *Journal of Econometrics*, (2), 45–56. https://doi.org/https://doi.org/10.1016/j.jeconom.2017.05.020

Davidson, R., & MacKinnon, J. G. (1998). Graphical methods for investigating the size and power of hypothesis tests. *The Manchester School*, *66*(1), 1–26. https://doi.org/https://doi.org/10.1111/1467-9957.00086

Fox, J., & Weisberg, S. (2018). Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals. *Journal of Statistical Software*, *87*(9), 1–27. https://doi.org/10.18637/jss.v087.i09

George, A. C., & Robitzsch, A. (2015). Cognitive diagnosis models in r: A didactic. *The Quantitative Methods of Psychology*, *11*, 189–205. https://doi.org/10.20982/tqmp.11.3.p189

Golden, R., Henley, S. S., White, H., & Kashner, T. M. (2013). New directions in information matrix testing: Eigenspectrum tests. *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*, 145–177.

Golden, R., Henley, S. S., White, H., & Kashner, T. M. (2016). Generalized information matrix tests for detecting model misspecification. *Econometrics*, *4*(4), 46.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, 1251–1271.

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model [Hosmer, D W Hosmer, T Le Cessie, S Lemeshow, S eng Comparative Study Research Support, Non-U.S. Gov't Review England 1997/05/15 Stat Med. 1997 May 15;16(9):965-80. doi: 10.1002/(sici)1097-0258(19970515)16:9<965::aid-sim509>3.0.co;2-o.]. *Stat Med*, *16*(9), 965–80. https://doi.org/10.1002/(sici)1097-0258(19970515)16:9<965::aid-sim509>3.0.co;2-o

Hosseinpourkhoshkbari, R., & Golden, R. M. (2024). Assessment of misspecification in cdms using a generalized information matrix test. In H. Hwang, H. Wu, & T. Sweet (Eds.), *Quantitative psychology* (pp. 361–371). Springer Nature Switzerland.

Lancaster, T. (1984). The covariance matrix of the information matrix test. *Econometrica: Journal of the Econometric Society*, 1051–1053.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behav Res*, *49*(4), 305–28. https://doi.org/10.1080/00273171.2014.911075

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2n contingency tables. *Journal of the American Statistical Association*, *100*(471), 1009–1020. https://doi.org/10.1198/016214504000002069

Presnell, B., & Boos, D. D. (2004). The ios test for model misspecification [doi: 10.1198/016214504000000214]. *Journal of the American Statistical Association*, *99*(465), 216–227. https://doi.org/10.1198/016214504000000214

Rupp, A. A., & Templin, J. (2008). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the dina model. *Educational and Psychological Measurement*, *68*, 78–96. https://doi.org/10.1177/0013164407301545

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems. final report for nie-g-81-0002*. Illinois University, Urbana. Computer-Based Education Research Lab.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. https://doi.org/10.2307/1912526

White, H. (1994). *Estimation, inference, and specification analysis*. Cambridge University Press.