Three New Studies on Model-data Fit for Latent Variable Models in Educational Measurement

Zhuangzhuang Han

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Three New Studies on Model-data Fit for Latent Variable Models in Educational Measurement

Zhuangzhuang Han

This dissertation encompasses three studies on issues of model-data fit methods for latent variable models implemented in modern educational measurement. The first study proposes a new statistic to test the mean-difference of the ability distributions estimated based on the responses of a group of examinees, which can be used to detect aberrant responses of a group of test-takers. The second study is a review of the current model-data fit indexes used for cognitive diagnostic models. Third study introduces a modified version of an existing item fit statistic so that the modified statistic has a known chi-square distribution. Lastly, a discussion of the three studies is given, including the studies' limitations and thoughts on the direction of future research.

Contents

List of Figures

List of Tables

Acknowledgements

I want to first express my profuse gratitude for Dr. Matthew Johnson, my advisor. It is your tireless support and patient guidance that sparked my research interests. Your personal attention via countless one-to-one advising on academic questions and even future career choices is deeply felt by me and crucially important to my growth, especially as an international student. I will endeavor to become a caring and encouraging person like you.

I owe thanks to, and feel indebted to, other professors who helped me during the doctorate program, especially, Dr. Lawrence DeCarlo and Dr. Young-Sun Lee. Dr. DeCarlo, thank you for sharing your perspectives and inspiring me via your weekly seminar. Dr. Lee, thank you for being such as a resource for me and standing by me all the time. I would also like to thank the members on my dissertation committee, Dr. Ye Wang and Dr. Sandip Sinharay. Without your insightful and constructive suggestions, the paper would not have made it this far.

A special thanks to Dr. Bryan Keller, Dr. Qiwei He, and Dr. Matthias von Davier whom I had the honor to work with. I am grateful for the research opportunities they provided me for the tremendous amount of guidance they offered on my own research, and for encouraging me with their strong work ethic and professional attitude.

Accompanied by great peers such as Dr. Xiang Liu and Dr. Huacheng Li, I have enjoyed this wonderful journey. I am grateful for their advices, experiences, and thoughts ranging from conducting academic research to searching for jobs.

Lastly, this wonderful journey would not have been possible without the unconditional support of my beloved fiancée, Xinyi He, and my dear parents, Huandi Liu and Jing Han.

To my family

Chapter 1

Introduction

This dissertation is comprised of three separate studies sharing a common theme: model-data fit of latent variable models (LVMs) implemented in educational measurement. Modern educational measurement relies heavily on LVMs. Responses are collected via test items as a major form of measurement instrument designed to measure one or several intended underlying cognitive constructs. LVMs serve as a useful tool for modeling the distribution of response data. As an outcome, inference procedures can be made upon the fitted models, such as estimating ability levels (or attribute/skill profiles) of respondents, ranking (or classifying) respondents, evaluating item characteristics (e.g., item difficulty and discrimination), adaptively selecting items matching a respondent's ability level to enhance the test efficiency, and so forth. Success and accuracy of these inferences hinge on the extent to which LVMs employed adequately describe responses. Model-data fit methods are developed to access such adequacy, to evidence the validity of the inference procedures and their applications.

Assessing model-data fit is a multi-facet procedure in the sense that misfit stems from different sources. Typically, LVMs involve a number of restrictive assumptions: dimensionality of latent variables, response functions specified to relate the probability of answering an item with a particular response to a level of ability, and local independence (that assumes responses are independent of each other given the level of ability). Violating any of the aforementioned assumptions could result in model-data misfit. Plus, misfit can be viewed and investigated from different perspectives, leading to useful applications in practice, especially in educational and psychological measurement. For example, model-data fit can be assessed at either a item-level or a person-level, allowing for checking “local” misfit of an individual item and identifying abnormal response vectors of participants respectively. As a result, even though a considerable amount of research on model fit for LMVs have been studied in educational measurement and other disciplines, the body of literature is continuously growing and the relevant topics keep updating, suggesting that there is still room for new studies. The three subsections that follow offer a brief introduction to my doctoral research, organized in chronological order.

1.1 The First Study: A Wald Test to Detect Mean-shift in A Group Ability Distribution

Within the framework of item response theory (IRT), person-fit analysis plays a substantive role in identifying aberrant response patterns at individual level. However, limited attention is given to detect an aberration in a set of responses from a group of test-takers.

To fill the gap, the first study proposes a Wald-type statistic measuring a standardized mean-shift in the group ability distribution estimated from responses for a group of test-takers, and tests whether the responses present any abnormality. To obtain the mean-shift, test items are separated into two sets, "clean" and suspected items, based on external information such as whether the items have been overused, leaked to the public or fraudulently compromised by test administrators. Given the precalibrated item parameters, the means of ability distributions, $\hat{\mu}_{\bar{\mathcal{S}}}$ and $\hat{\mu}_{\mathcal{S}}$, are estimated from responses to "clean" and suspected sets of items respectively. The wald statistic is written as

$$Z_g = \frac{\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}}}{\sigma\left(\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}}\right)}.$$

The computation of $\sigma(\hat{\mu}_{\bar{\mathcal{S}}} - \hat{\mu}_{\mathcal{S}})$ that takes into account the generic dependency between $\hat{\mu}_{\mathcal{S}}$ and $\hat{\mu}_{\bar{\mathcal{S}}}$ is derived out in detail.

Simulation studies show that the Type-1 error rate of the Wald statistic under various conditions (created using different group sizes, degrees of item quality, and so on) is close to the nominal level. The feasibility of the test is further discussed by studies of power and analyses in real datasets.

1.2 The Second Study: Global- and Item-level Fit Indexes for Cognitive Diagnostic Models

Cognitive diagnostic models (CDMs), also regarded as a type of restricted latent class model, have gained prominence in educational and psychological measurement. One of the benefits of CDMs is that the models provide a parametric framework, through which inferences about what skills a test-taker has can be made—that is, assigning a test-taker into an attribute profile (i.e., a class). One way to validate these inferences and their corresponding applications is to assess how well the model fits the dataset. Model-data fit indexes are developed to conduct the appraisal. In addition, model-data fit indexes can be used to meet other needs: model comparison and selection.

Considerable amount of attention has been paid to model fit indexes among recent studies on CDMs. There has been a demand of systematical reviews and guidances of current methods for practitioners. The second study reviews the current model fit indexes for CDMs by summarizing them into four categories according to two aspects of the indexes: (1) the level of fit analysis, i.e., global/test-level versus item-level analysis; (2) the choice of the reference model for comparison, i.e., an alternative CDM (relative/comparative fit analysis), or a saturated categorical model (absolute fit analysis). Pros and cons for each category of indexes are listed and suggestions are given, on the basis of results from current literature. A publicly available dataset is included at the end of this article to demonstrate the feasibility of some selected model fit indexes in practice.

1.3 The Third Study: The Standardized S-X^2 for Item Fit Analysis

Item fit index S-X^2 (Orlando & Thissen, 2000) is arguably one of the most popular statistics for assessing item fit of item response theory models. Sinharay (2006a) used the theoretical arguments from Chernoff and Lehmann (1954), as well as, simulations to prove that S-X^2 would not follow its theorized large-sample distribution under the null hypothesis. Therefore the inaccurate approximation of the large-sample distribution would lead to slightly inflated Type I error rates. But an adjusted (essentially, standardized) version of S-X^2 has remained elusive.

Utilizing the modification procedure of Rao and Robson (1974) the third study introduces a standardized version of S-X^2 that is proven to have a known large-sample distribution under the null hypothesis. Simulation results show the Type I error rate of the standardized version is smaller than, or equal to the nominal level, when compared to the original S-X^2. An application of the proposed statistic to a real-world dataset is analyzed to illustrate its utility.

Chapter 2

A Wald Test to Detect Mean-shift in a Group Ability Distribution

2.1 Introduction

Current studies related to IRT have paid substantial amount of attention to analyzing model-data fit on individual item-score patterns. One potential culprit of the misfit is fraudulent test behaviors. Detecting and quantifying misfit assists in identifying aberrant examinees. Fraudulent test behaviors could also happen at the group level. For example, over hundred teachers from more than 40 Atlanta public schools were allege, and many were found guilty, of cheating on state-administrated standardized tests by altering students' answers Vogell and Perry (2009). Cheating at the group level hinders the validity of tests and raised fairness concerns. It is absolutely imperative to develop appropriate statistical approaches for detecting aberrant group responses, to further facilitate identifying aberrant groups, for instance, who commit cheating behaviors, who benefit from the preknowledge on some items Kyle (2002); Hornby (2011), or whose responses might be compromised by test administrators Jacob and Levitt (2003). The purpose of this study is to provide a IRT-based statistic to detect the aberrance in group responses.

In the literature of IRT, various person-fit indexes, also referred to as appropriateness measures, have been developed to measure the difference between an observed individual response vector and its model-implied counterpart, which can be used to detect aberrant responses at the individual level. Difference can be assessed in different aspects, resulting in a variety of person-fit indexes. For instance, the oft-cited l_z statistic Drasgow, Levine, and Williams (1985) is based on the individual log-likelihood and U Wright and Stone (1979) looks at the squared residuals. A comprehensive overview of person-fit indexes can be found in the methodology review by Meijer and Sijtsma (2001). Subsequent studies Snijders (2001); Magis, Raîche, and Béland (2011); Sinharay (2016a) are focused on the technical details of the existing person-fit indexes. Among them, Snijders (2001) corrected the asymptotic null distribution of the l_z by introducing a modified statistic l_z^* taking into account the uncertainty of the estimated person parameter. Furthermore, Sinharay (2016a) extended l_z^* to polytomous and mixed-form (with both dichotomous and polytomous responses) test responses.

Another class of methods for individual responses assumes that investigators have the knowledge of which items are suspected. One example is the methods of detecting fraudulent erasures van der Linden and Jeon (2011); Wollack (1997); Wollack, Cohen, and Eckerly (2015). The rationale behind these methods is that: first, person parameters can be estimated on the basis of a set of "clean" (unsuspected) items; in subsequent, the expected responses to the suspected items (e.g., unusual erasures) can be obtained using the estimated person parameters and then compared against the observed responses. The methods of item preknowledge detection are another example of the class. Belov (2013) suggested using the approximate *Kullback-Leibler divergence* (KLD) between the two posterior distributions of the person parameter computed using the suspected and unsuspected sets of items. Sinharay (2016b) noted that the approximate KLD does not have a known null distributiona and summarized the limitations of using the empirical approach for deciding the critical value of approximate KLD. Instead, he suggested employing the Likelihood Ratio Test (LRT) and the score test to detect item preknowledge.

A limited number of approaches have been developed for detecting the aberrance in group responses. Skorupski, Fitzpatrick, and Egan (2017) examined unusual longitudinal gains of group-level abilities estimated from tests administrated at multiple time points. Sinharay (2018) aggregated the erasure detection index Wollack et al. (2015) and came up with a group-level erasure detection index.

Like some of the aforementioned approaches, the present study separates items of a test into two disjoint sets—the suspected and the unsuspected. In practice the separation can be informed using various external sources. For instance, the abnormal erasure rate of a paper-pencil test can be used to indicate items that were potentially compromised. As another example, the overexposed linking items, at a high chance of being leaked, are those that the examinees are more likely to have preknowledge of and, therefore, can be treated as a natural set of suspected items. Conditional on the known item parameters that have been already calibrated, two ability distributions of a group of examinees can be computed separately using the unsuspected and the suspected sets of items. A Wald-type statistic is employed by the current study to test the difference between the means of the two estimated latent distributions while taking into account the correlations induced by the within-subject effect.

2.2 Theory

2.2.1 Preliminaries

In the context of IRT, the probability of a correct item response is parametrized as a function of the subject's latent ability and item parameters. The function is commonly referred to as the item characteristic curve (ICC). For example, the three-parameter logistic (3PL) model assumes

$$P(Y_{ij} = 1|\theta_i) = c_j + (1 - c_j)\,\frac{\exp\left(a_j(\theta_i - b_j)\right)}{1 + \exp\left(a_j(\theta_i - b_j)\right)}, \tag{2.1}$$

where $Y_{ij} = 1$ indicates that i^{th} subject answers j^{th} item correctly. Here θ_i represents the latent ability that is a unidimensional parameter assumed to follow a standard normal distribution. The parameters a_j, b_j and c_j are discrimination, difficulty, and guessing parameters respectively; Setting $c_j = 0$ leads to an 2PL model, whereas forcing $a_j = 1$ and $c_j = 0$ results in an Rasch model. Although IRT models can be generalized to describe polytomous responses, the scope of the current study is restricted to dichotomous responses. Comprehensive discussions and reviews on IRT models can be found in the references Hambleton and Swaminathan (1985); van der Linden and Hambleton (2013).

Furthermore, IRT models assume the *local independence*, that is, responses of a test-taker are independent of each other conditional on her latent ability θ_i. That being said, the likelihood of the random response vector of the i^{th} test-taker, $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{ij}, ..., Y_{iJ})^\top$, can be written as

$$P(\boldsymbol{Y}_i|\theta_i) = \prod_{j=1}^{J} P_j(\theta_i)^{Y_{ij}} Q_j(\theta_i)^{1-Y_{ij}}.$$

$P_j(\theta_i) \equiv P(Y_{ij} = 1|\theta_i)$ and $Q_j(\theta_i) = 1 - P_j(\theta_i)$, indicating the probability of answering the item incorrectly. J denotes the number of items. By the same token, $P\left(\boldsymbol{Y}_{i|\mathcal{S}}|\theta_i\right)$ and $P\left(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\theta_i\right)$ are obtained for $\boldsymbol{Y}_{i|\mathcal{S}} = \{Y_{ij} \mid j \in \mathcal{S}\}$ and $\boldsymbol{Y}_{i|\bar{\mathcal{S}}} = \{Y_{ij} \mid j \in \bar{\mathcal{S}}\}$, where $\mathcal{S}$ and $\bar{\mathcal{S}}$ stand for the sets of integers indexing the suspected and unsuspected items.

The marginal likelihood is obtained by integrating the likelihood with respect to θ, namely,

$$P(\boldsymbol{Y}_i|\mu, \sigma) = \int P(\boldsymbol{Y}_i|\theta)\phi(\theta|\mu, \sigma)d\theta,$$

where $\phi(\theta_i|\mu, \sigma)$ is the probability density function of the normal distribution with mean μ and standard deviation σ. Quadrature methods are implemented to calculate this integration in practice. Following the same logic, marginal likelihoods $P(\boldsymbol{Y}_{i|\mathcal{S}}|\mu, \sigma)$ and $P(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\mu, \sigma)$ are computed for the suspected and the unsuspected items. As a result, the maximum likelihood estimates (MLEs) of the latent ability distribution parameters are obtained by maximizing the logarithm of the marginal likelihoods—that is,

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \sum_{i}^{N} \ell(\boldsymbol{y}_i|\mu, \sigma) \tag{2.2}$$

$$\hat{\mu}_{\mathcal{S}}, \hat{\sigma}_{\mathcal{S}} = \underset{\mu,\sigma}{\operatorname{argmax}} \sum_{i}^{N} \ell(\boldsymbol{y}_{i|\mathcal{S}}|\mu, \sigma) \tag{2.3}$$

$$\hat{\mu}_{\bar{\mathcal{S}}}, \hat{\sigma}_{\bar{\mathcal{S}}} = \underset{\mu,\sigma}{\operatorname{argmax}} \sum_{i}^{N} \ell(\boldsymbol{y}_{i|\bar{\mathcal{S}}}|\mu, \sigma) \tag{2.4}$$

where $\boldsymbol{y}_i$ represents a realization of the random vector $\boldsymbol{Y}_i$ and N denotes the number of test-takers in the group.

Notice that $\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_N$ are identically and independently distributed ($i.i.d.$) random variables following a J-dimensional multivariate Bernoulli (MVB) distribution Dai, Ding, and Wahba (2013); Teugels (1990) with a probability mass function $(\boldsymbol{Y}_i|\mu, \sigma)$. This assumption will be utilized later in the derivation that follows.

2.2.2 The suggested Wald test

Under the null hypothesis that there is no aberrance in the group responses,

$$\mu_{\mathcal{S}} = \mu_{\bar{\mathcal{S}}} = \mu_0$$

or

$$\mu_{\mathcal{S}} - \mu_{\bar{\mathcal{S}}} = 0$$

equivalently, where μ_0 denotes the mean of the latent ability distribution for the group of test-takers when there is no aberrance. The Wald statistic can be written as

$$Z_g = \frac{(\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}}) - 0}{\sigma\left(\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}}\right)}.$$

Under the null hypothesis the large-sample distribution of the Wald statistic is claimed to be accurately approximated by the standard normal distribution. The primary objective of the following sections is to derive the computation of $\sigma\left(\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}}\right)$.

Before proceeding to this derivation, it is worthwhile to note that $\sigma_{\mathcal{S}} = \sigma_{\bar{\mathcal{S}}}$ is assumed for the test. Put differently, the test is only focused on the difference between the two mean parameters. To estimate the two mean parameters given such the constraint of $\sigma_{\mathcal{S}} = \sigma_{\bar{\mathcal{S}}}$, first, $\hat{\sigma}$ is computed via (2.2); in subsequent, $\hat{\mu}_{\mathcal{S}}$ and $\hat{\mu}_{\bar{\mathcal{S}}}$ are estimated via (2.3) and (2.3) conditional on $\hat{\sigma}_{\mathcal{S}} = \hat{\sigma}_{\bar{\mathcal{S}}} = \hat{\sigma}$.

The restrictive assumption is reasoned on several arguments that follow. First, it is not uncommon in practice that certain test statistics are often used even when the equal-variance assumption is moderately violated. For example, previous studies Cochran (1947); Bradley (1978); Ramsey (1980) showed that the type-I error of the two-sample t-test is close to the nominal level regardless of the variances being equal or not when the sample size is equal or larger than 15. As presented later in this study, a sensitivity analysis is conducted in this study to demonstrate that the type-I error of the suggested Wald test is still close to the nominal level when the assumption is violated to a moderate extent. Second, a more sophisticated test statistic is needed if the constraint of $\sigma_{\mathcal{S}} = \sigma_{\bar{\mathcal{S}}}$ is relaxed as a more complexed correlation structure of the mean and standard deviation parameters must be be taken into account. In this case, the suggested Wald test just focusing on the "mean-shift" can be regarded as a starting point for the more sophisticated methods.

As mentioned above, the within-subject correlations needs to be taken into account for the computation of $\sigma\left(\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}}\right)$. Here we provide a brief explanation on the derivation of $\sigma\left(\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}}\right)$. Readers interested in the more detailed derivation are referred to in Appendix A.

Let us consider the first derivatives of the marginal log-likelihoods based on the two sets of items, $\ell'_{\mathcal{S}}(\hat{\mu}_{\mathcal{S}}) = \frac{\partial \ell(\boldsymbol{y}_{\mathcal{S}}|\mu)}{\partial \mu}|_{\mu=\hat{\mu}_{\mathcal{S}}}$ and $\ell'_{\bar{\mathcal{S}}}(\hat{\mu}_{\bar{\mathcal{S}}}) = \frac{\partial \ell(\boldsymbol{y}_{\bar{\mathcal{S}}}|\mu)}{\partial \mu}|_{\mu=\hat{\mu}_{\bar{\mathcal{S}}}}$, where $\ell(\boldsymbol{y}_{\mathcal{S}}|\mu) = \sum_i^N \ell(\boldsymbol{y}_{i|\mathcal{S}}|\mu)$ and $\ell(\boldsymbol{y}_{\bar{\mathcal{S}}}|\mu) = \sum_i^N \ell(\boldsymbol{y}_{i|\bar{\mathcal{S}}}|\mu)$. The two derivatives are equal to zero since they are evaluated at the MLEs. A first-order Taylor series can be expanded around the μ_0 for each of the two derivatives. For instance:

$$\ell'_{\mathcal{S}}(\hat{\mu}_{\mathcal{S}}) \approx \ell'_{\mathcal{S}}(\mu_0) + \ell''_{\mathcal{S}}(\mu_0)(\hat{\mu}_{\mathcal{S}} - \mu_0)$$

is the expanded derivative of the log-likelihood for the suspected set of items. With the two first-order expansions, the approximation for $\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}}$ can be obtained as

$$\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}} \approx -\frac{\ell'_{\mathcal{S}}(\mu_0)}{\ell''_{\mathcal{S}}(\mu_0)} + \frac{\ell'_{\bar{\mathcal{S}}}(\mu_0)}{\ell''_{\bar{\mathcal{S}}}(\mu_0)}. \tag{2.5}$$

Essentially, the numerators on the right-hand side of (2.5) are two random variables with a covariance written as

$$\mathrm{Cov}\left[\ell'_{\mathcal{S}}(\mu_0), \ell'_{\bar{\mathcal{S}}}(\mu_0)\right] = N\mathrm{Cov}\left[\frac{\partial \ell(\boldsymbol{Y}_{i|\mathcal{S}}|\mu)}{\partial \mu}|_{\mu=\mu_0}, \frac{\partial \ell(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\mu)}{\partial \mu}|_{\mu=\mu_0}\right].$$

The above equation holds true because $\boldsymbol{Y}_{i|\mathcal{S}}$ and $\boldsymbol{Y}_{i_\prime|\bar{\mathcal{S}}}$ are independent when $i \neq i_\prime$. Furthermore, $\{\boldsymbol{Y}_{i|\mathcal{S}}|i \in 1, ..., N\}$ and $\{\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|i \in 1, ..., N\}$ both are *i.i.d.* as mentioned previously, suggesting $\{\frac{\partial \ell(\boldsymbol{Y}_{i|\mathcal{S}}|\mu)}{\partial \mu}|_{\mu=\mu_0}|i \in 1, ..., N\}$ and $\{\frac{\partial \ell(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\mu)}{\partial \mu}|_{\mu=\mu_0}|i \in 1, ..., N\}$ are also *i.i.d.*. Therefore, the two numerators on the right-hand side of (2.5), $\ell'_{\mathcal{S}}(\mu_0)$ and $\ell'_{\bar{\mathcal{S}}}(\mu_0)$, in asymptotic will converge in distribution to the normal distributions $\mathcal{N}\left(0, N\mathcal{I}_{\mathcal{S}}(\mu_0)\right)$ and $\mathcal{N}\left(0, N\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0)\right)$ respectively, according to the Central Limited Theorem. The means of the normal distributions are 0; the variances are $N\mathcal{I}_{\mathcal{S}}(\mu_0)$ and $N\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0)$. $\mathcal{I}_{\mathcal{S}}(\mu_0)$ and $\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0)$ are the Fischer information, where $\mathcal{I}_{\mathcal{S}}(\mu_0) = \mathrm{Var}\left[\frac{\partial \ell(\boldsymbol{Y}_{i|\mathcal{S}}|\mu)}{\partial \mu}|_{\mu=\mu_0}\right]$ and $\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0) = \mathrm{Var}\left[\frac{\partial \ell(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\mu)}{\partial \mu}|_{\mu=\mu_0}\right]$.

The denominators $\ell''_{\mathcal{S}}(\mu_0)$ and $\ell''_{\bar{\mathcal{S}}}(\mu_0)$, by the Law of Large Number (LLN), will converge in probability to constants $N\mathcal{I}_{\mathcal{S}}(\mu_0)$ and $N\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0)$ when the sample size is large. Using Slutsky's theorem Casella and Berger (2001), the right-hand side of (2.5) will converge to a sum of two correlated normal variables whose standard deviation can be written as

$$\left\{\frac{1}{N\mathcal{I}_{\mathcal{S}}(\hat{\mu}_{\bar{\mathcal{S}}})} + \frac{1}{N\mathcal{I}_{\bar{\mathcal{S}}}(\hat{\mu}_{\bar{\mathcal{S}}})} - \frac{2\mathrm{Cov}[\ell'(\boldsymbol{Y}_{i|\mathcal{S}}|\hat{\mu}_{\bar{\mathcal{S}}}), \ell'(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\hat{\mu}_{\bar{\mathcal{S}}})]}{N\mathcal{I}_{\mathcal{S}}(\hat{\mu}_{\bar{\mathcal{S}}})\mathcal{I}_{\bar{\mathcal{S}}}(\hat{\mu}_{\bar{\mathcal{S}}})}\right\}^{1/2} \tag{2.6}$$

given that μ_0 is evaluated at $\hat{\mu}_{\bar{\mathcal{S}}}$. This standard deviation can be used to approximate the $\sigma(\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}})$.

2.3 Simulation Studies

2.3.1 Type-I error

Throughout the current and the following subsections, dichotomous group responses are generated in the context of the 2PL model. Person parameters are randomly simulated from $\mathcal{N}(\mu_0, \sigma_0^2)$ to generate group responses.

Two group sizes, $I = 50$ and $I = 100$, are considered for the Type-I error. Test length is set as $J = 40$ and the first 20 items are regarded as compromised. $\mu_0 \in \{-1, 0, 1\}$ and $\sigma_0 \in \{0.5, 1\}$ for the *true* latent ability distribution are examined. The two-sided test based on Z_g is conducted at the nominal level 0.05. $10,000$ replications are performed across the 12 $(2 \times 2 \times 3)$ simulation conditions. In each replication item discrimination parameter a_j and difficulty parameter b_j are randomly generated, i.e., $a_j \sim U(0.5, 2.0)$ and $b_j \sim U(-2.0, 2.0)$.

Table 2.1: The Type-I error of Z_g (Nrep = 10,000)

	$\sigma_0 = 0.5$		$\sigma_0 = 1.0$	
	$N = 50$	$N = 100$	$N = 50$	$N = 100$
$\mu_0 = -1$	0.055	0.054	0.057	0.052
$\mu_0 = 0$	0.054	0.047	0.058	0.055
$\mu_0 = 1$	0.060	0.055	0.053	0.049

Table 2.1 shows the Type-I error of Z_g is close to the nominal level and the values are stable across different combinations of μ_0 and σ_0. The Type-I error becomes even closer to the nominal level as the group size increases.

2.3.2 Sensitivity analysis of the assumption $\sigma_{\mathcal{S}} = \sigma_{\bar{\mathcal{S}}}$

Responses for a group of test-takers are simulated through the following steps. N individual ability parameters ($\theta_{i|\bar{\mathcal{S}}}$) are simulated from $\mathcal{N}(\mu_0, \sigma^2_{\bar{\mathcal{S}}})$, where $\mu_0 = 0$ and $\sigma_{\bar{\mathcal{S}}} = 1$. Responses of the unsuspected items are generated using the simulated $\theta_{i|\bar{\mathcal{S}}}$ and the item parameters generated as the last section; $\theta_{i|\mathcal{S}} = \frac{(\theta_{i|\bar{\mathcal{S}}} - \mu_0)}{\sigma_{\bar{\mathcal{S}}}}\sigma_{\mathcal{S}} + \mu_0$ is used to generate responses of the suspected items, where $\sigma_{\mathcal{S}} \in \{1.25, 1.1, 0.9, 0.75\}$. By doing so, responses violating the equal-variance assumption are generated. Different numbers of suspected items are also examined, that is, $n_{\mathcal{S}} \in \{5, 10, 20\}$. The total number of items J is fixed at 40. The two-tailed test using the Z_g on the simulated group responses is conducted at the 0.05 nominal level. If the test statistic is robust to the assumption violation, then the rate of cases being rejected should be close around the nominal level.

Table 2.2: Sensitivity analysis of Z_g under $\sigma_{\bar{\mathcal{S}}} \neq \sigma_{\mathcal{S}}$ (Nrep = 10,000)

	$n_{\mathcal{S}}$	$\frac{\sigma_{\bar{\mathcal{S}}}}{\sigma_{\mathcal{S}}} = 1:1.25$	$\frac{\sigma_{\bar{\mathcal{S}}}}{\sigma_{\mathcal{S}}} = 1:1.1$	$\frac{\sigma_{\bar{\mathcal{S}}}}{\sigma_{\mathcal{S}}} = 1:1$	$\frac{\sigma_{\bar{\mathcal{S}}}}{\sigma_{\mathcal{S}}} = 1:0.9$	$\frac{\sigma_{\bar{\mathcal{S}}}}{\sigma_{\mathcal{S}}} = 1:0.75$
	5	0.113	0.069	0.047	0.041	0.037
$N = 100$	10	0.099	0.067	0.049	0.043	0.032
	20	0.056	0.055	0.051	0.053	0.053
	5	0.153	0.071	0.051	0.048	0.054
$N = 300$	10	0.119	0.073	0.050	0.043	0.042
	20	0.063	0.059	0.052	0.054	0.060

Results reported in Table 2.2 indicate that the Wald statistic becomes less sensitive to the violation of equal variance assumption, as the increase of the group size and the decrease of the number of suspected items. Cases examined here are relatively extremer than practical ones because all of test-takers are "rescaled" to have higher values of ability parameters when answering suspected items. A relatively more practical case is that only a portion of test-takers unfairly "benefit" from the suspected items. The analysis in this case were performed but is not reported in the current study whose results show that the Wald statistic is even less sensitive to the violated assumption than the analysis presented.

2.3.3 Power

Two analyses of the power for Z_g are studied in this section. Effect sizes used in the first analysis are defined by adding an increase or a positive "shift" ($\Delta\theta$) to $\theta_{i|\bar{\mathcal{S}}}$, that is,

$$\theta_{i|\mathcal{S}} = \theta_{i|\bar{\mathcal{S}}} + \frac{\Delta\theta}{\sigma_{\bar{\mathcal{S}}}}, \tag{2.7}$$

where $\theta_{i|\bar{\mathcal{S}}} \sim \mathcal{N}(\mu_{\bar{\mathcal{S}}}, \sigma^2_{\bar{\mathcal{S}}})$. Here $\mu_{\bar{\mathcal{S}}} \in \{-1, 0, 1\}$ and $\sigma_{\bar{\mathcal{S}}} \in \{0.5, 1.0\}$. The "shift" $\Delta\theta$ is standardized by σ_0 before added to $\theta_{i|\bar{\mathcal{S}}}$. $\theta_{i|\bar{\mathcal{S}}}$ is used to generate the responses to the unsuspected items; $\theta_{i|\mathcal{S}}$ is for the suspected. By this token, simulated group responses preserve the feature that $\boldsymbol{Y}_{i|\mathcal{S}}$ is correlated with $\boldsymbol{Y}_{i|\bar{\mathcal{S}}}$ for the same test-taker i, whereas $\boldsymbol{Y}_{i|\mathcal{S}}$ and $\boldsymbol{Y}_{i'|\bar{\mathcal{S}}}$ are independent for two different test-takers.

In this simulation analysis $N = 100$ and $J = 40$. Among all the items, twenty of them are assumed as the suspected items. In addition, $a_j \sim U(0.5, 1.25)$ and $a_j \sim U(1.25, 2)$ investigate the effect of discrimination parameters, that is, item quality. Item difficulty parameters b_j are simulated from $U(-2, 2)$. Effect sizes, $\Delta\theta/\sigma_0 \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ are examined. In total there are 60 ($2 \times 2 \times 3 \times 5$) simulation conditions studied. The two-sided test using Z_g is performed with respect to the 0.05 nominal level.

Table 2.3 reports the Monte-Carlo approximated power of Z_g. As expected, the power grows as the increase of $\Delta\theta/\sigma_{\bar{\mathcal{S}}}$. The power for cases with better item quality, i.e., $\boldsymbol{a} \sim U(1.25, 2.0)$, is on average higher than the others. Power is significantly higher under the

Table 2.3: The power of the Z_g obtained using effect sizes $\Delta\theta/\sigma_{\bar{S}}$ (Nrep = 3,000)

		$\sigma_{\bar{S}} = 0.5$			$\sigma_{\bar{S}} = 1.0$		
	$\Delta\theta/\sigma_{\bar{S}}$	$\mu_{\bar{S}} = -1$	$\mu_{\bar{S}} = 0$	$\mu_{\bar{S}} = 1$	$\mu_{\bar{S}} = -1$	$\mu_{\bar{S}} = 0$	$\mu_{\bar{S}} = 1$
	0.05	0.437	0.486	0.417	0.154	0.166	0.130
	0.10	0.926	0.942	0.902	0.392	0.448	0.362
$\boldsymbol{a} \sim U(1.25, 2)$	0.15	0.990	0.994	0.990	0.698	0.769	0.654
	0.20	0.998	1.000	0.998	0.882	0.950	0.862
	0.25	1.000	1.000	1.000	0.968	0.988	0.955
	0.05	0.236	0.224	0.234	0.097	0.097	0.100
	0.10	0.666	0.660	0.640	0.208	0.217	0.191
$\boldsymbol{a} \sim U(0.5, 1.25)$	0.15	0.924	0.922	0.908	0.376	0.449	0.382
	0.20	0.988	0.989	0.981	0.622	0.666	0.566
	0.25	0.997	0.998	0.996	0.778	0.836	0.762

conditions with smaller $\sigma_{\bar{S}}$. The power for the cases with $\mu_{\bar{S}} = 0$ is slightly larger than the others; the power for the cases with $\mu_{\bar{S}} = -1$ is slightly higher than those with $\mu_{\bar{S}} = 1$, likely due to the fact that all $\Delta\theta/\sigma_{\bar{S}}$ in this simulation analysis are assumed to be positive.

Effect sizes used in the second analysis are defined in a more practical way. First, responses are simulated by means of the same mechanism used in the section of the Type-I error. θ_i used to simulate responses is sampled from the normal distributions with $\mu_0 \in \{-1, 0\}$ and $\sigma_0 = 1$. $N = 100$ and $J = 40$ are used for this analysis. Item parameters are randomly sampled as $a_j \sim U(0.5, 2.0)$ and $b_j \sim U(-2.0, 2.0)$. Second, a certain number of items are selected from the suspected set of items based on a predetermined proportion p_1; meanwhile, some test-takers are selected from the group, using a predetermined proportion p_2. Last, for the selected items and test-takers, the corresponding simulated responses are forced to be positive if they are not positive. Effect sizes are defined by the combinations of two proportions, where $p_1 \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ and $p_2 \in \{0.5, 0.7, 0.9\}$.

The group-level *erasure detection index* (EDI_g) introduced by Wollack and Eckerly

(2017) is performed as a comparison to the suggested Wald test. EDI_g is written as

$$EDI_g = \frac{\sum_{i=1}^{N}(X_i - \hat{\mu}_i) - 0.5}{\sqrt{\sum_{i=1}^{N} \hat{\sigma}_i^2}},$$

where

$$\hat{\sigma}_i = \sqrt{\sum_{j \in \mathcal{S}} P_i(\hat{\theta}_j) \left[1 - P_i(\hat{\theta}_j)\right]}$$

X_i is defined as the raw score of the test-taker i on the items in the suspected set $\mathcal{S}$ that, in the context of erasure detection, is referred to as the set of items on which erasure are found. μ_i and σ_i, respectively, are denoted as the expected value and the standard deviation of X_i. $\hat{\sigma}_i$ is computed through $\sum_{j \in \mathcal{S}} P_i(\hat{\theta}_j)$, where $\hat{\theta}_j$ is estimated using the responses to the items in the unsuspected set $\bar{\mathcal{S}}$. The constant 0.5 in the nominator of the right-hand side of the above expression of EDI_g is used for the continuity correction. Wollack and Eckerly (2017) assumed that EDI_g in large-sample follows a standard normal distribution. In this study, EDI_g is performed, as well as Z_g, using the one-tailed test with the alternative hypothesis that the raw score on the items of the suspected set is abnormally larger than the expected score.

It's noteworthy that the conditions defined by p_1 and p_2 create a particular scenario that a large number of test-takers in a group compromise or unusually benefit from a small amount of suspected items. Z_g exhibits a higher power than EDI_g for the cases with $p_2 < 0.15$ (the number of suspected items is small), whereas EDI_g becomes more powerful as p_2 raises. Overall, Z_g possesses a comparable power with respect to EDI_g; power of both tests becomes greater as the effect size (namely, the value of p_1 and p_2) increases. It's reasonable to observe that both tests have a more prevalent power in detecting the cases

Table 2.4: The power of the Z_g and the EDI_g (Nrep = 3,000)

		Z_g		EDI_g	
p_1	p_2	$\mu_0 = -1$	$\mu_0 = 0$	$\mu_0 = -1$	$\mu_0 = 0$
	0.5	0.336	0.215	0.224	0.147
0.05	0.7	0.488	0.322	0.430	0.311
	0.9	0.609	0.401	0.597	0.437
	0.5	0.695	0.459	0.628	0.424
0.10	0.7	0.861	0.649	0.795	0.517
	0.9	0.938	0.764	0.925	0.686
	0.5	0.903	0.705	0.858	0.702
0.15	0.7	0.971	0.871	0.913	0.856
	0.9	0.992	0.937	0.981	0.919
	0.5	0.973	0.879	1.000	0.991
0.20	0.7	0.997	0.968	1.000	0.998
	0.9	1.000	0.988	1.000	0.999
	0.5	0.994	0.951	1.000	0.999
0.25	0.7	0.999	0.993	1.000	1.000
	0.9	1.000	0.997	1.000	1.000

with $\mu_0 = -1$ than $\mu_0 = 0$ because, to create suspected responses, those initially generated as incorrect might be randomly forced to be correct, rather than the other way around.

2.4 Real-data Applications

Three public available datasets are analyzed here to demonstrate the utility of Z_g. The first two datasets are used as common examples in the handbook of cheating detection methods edited by Cizek and Wollack (2017). The first dataset includes item responses to the two test forms of a computer-based nonadaptive credentialing exam for a certain population of examinees. The second dataset contains the item responses to a state-administrated paper-pencil based math assessment taken by a population of fifth grade students. The third dataset collects the responses to items of a self-report assessment on verbal aggression. The dataset was first introduced by Smits, De Boeck, and Vansteelandt (2004) and included as an illustrative example in the R package “difR” Magis, Béland, Tuerlinckx, and De Boeck (2010) developed for the differential item functioning (DIF) analysis.

2.4.1 Nonadaptive credentialing assessment

The two test forms both contain 170 operational items scored dichotomously. There are 1,636 examinees taking the Form 1 and 1,644 examinees taking the Form 2. Among the 170 items, there are 63 and 61 items in the Form 1 and the Form 2 respectively, suspected as compromised items by the credentialing organization who provides the dataset Cizek and Wollack (2017). The examinees are separated into two groups based on the forms they took. Tests using the Bonferroi-adjusted L_s and the Z_g are conducts based on the responses of the two groups. The tests using the Bonferroi-adjusted L_s tests significant, indicating the two groups of examinees benefit from the suspected set of items. As expected, the tests of Z_g have results, namely, $Z_g = 3.439(p < .001)$ for the group taking the Form 1 and $Z_g = 9.826(p < .000)$ for the group taking the Form 2.

2.4.2 K-12 paper-based math assessment

The second dataset collects students' item responses over two academic years Cizek and Wollack (2017). The dataset used in this analysis only involves the responses from the fifth graders at Year 2. Specifically, the reduced dataset includes 72,686 students from 3,213 classes nested in 1,187 schools. There are five equated forms of tests and each includes 53 multiple-choice questions. Erasure information (wrong-to-right/WTR, wrong-to-wrong/WTW, right-to-wrong/RTW) have been recorded.

Item parameters are estimated by means of the Rasch model. An item is classified into the suspected set if the *total erasure rate* (combining WTR, WTW and RTW) is larger than the threshold 0.05. The *total erasure rate* is calculated for each school, implying that the suspected set of items varies across different schools.

Z_g is applied to conduct the one-tailed test with the alternative hypothesis that $\mu_{\mathcal{S}} > \mu_{\bar{\mathcal{S}}}$. Only the schools with more than 50 students are analyzed (602 in total) for the purpose of having better approximation to the limiting distribution of Z_g. Seventy-two schools are significant in the tests at the level 0.05; among them, fifteen schools are significant at the level 0.001. The results are presented in Table 2.5. N and $J_{\mathcal{S}}$ are the group size and the number of suspected items. $AER_{\mathcal{S}}$ stands for the average of the *total erasure rates* of the suspected items. The school ID numbers provided in the table are the same as those used in Cizek and Wollack (2017).

Table 2.5 reports the 15 schools (groups) flagged by the tests of Z_g significant at the nominal level 0.001, suggesting these groups of students could have unusual gains in terms of the means of their ability distributions induced by the fraudulent erasure behaviors. The results also suggest using the suggested Wald test as an omnibus test, followed by investigating the suspected individuals in detail with the person-level methods.

Table 2.5: The application of Z_g on the paper-pencil based math assessment

School ID	Z_g	N	$J_{\mathcal{S}}$	$AER_{\mathcal{S}}$
15790	3.158	61	5	0.082
245982	3.961	67	23	0.075
15517	3.382	55	28	0.073
145442	3.676	64	11	0.070
201035	3.197	71	17	0.075
243667	5.118	148	11	0.061
297195	4.439	120	13	0.064
232059	3.821	54	25	0.070
359790	3.146	91	4	0.060
296356	3.184	58	14	0.058
315235	3.333	62	10	0.084
391308	4.762	69	10	0.067
403410	3.164	52	12	0.064
214595	3.645	99	16	0.068
267082	4.519	83	5	0.070

2.4.3 Verbal aggression

There are 24 items for the test scored dichotomously and answered by 316 participants (243 females and 73 males). Table 2.6 shows the four basic situations implemented to construct the content of items. The four situations are fully crossed with two action modes ("want" or "do") and three verbal behaviors ("cursing", "shouting", or"'scolding"), resulting in 24 items in total.

Table 2.6: Four types of situation used to create verbal aggression items

S1: A bus fails to stop for me.
S2: I miss a train because a clerk gave me faulty information.
S3: The grocery store closes just as I am about to enter.
S4: The operator disconnects me when I had used up my last 10 cents for a call

According to Magis et al. (2010), there were 9 items (item 6, 8, 14, 16, 17, 19, 20, 22, and 23) identified as DIF items with respect to the *focal* group (male) using five distinct

DIF detection methods. The rationale behind the current analysis is to mimic the suspected items with the DIF items and conduct the suggested test on the responses of the male group to see if the male group "benefit" from the DIF items. To obtain the test statistic, item parameters are estimated based on the responses of the whole sample population (including both *focal*/male group (male) and *reference*/female group) and treated as known. As an outcome, $Z_g = 5.162$ with $p\text{-}value < 0.001$ for the male group; $Z_g = -0.405$, $p\text{-}value = 0.657$ for the female group.

2.5 Discussion

In this study a Wald test statistic is developed to detect abnormal responses for a group of test-takers, whereas traditionally the aberrance is assessed at the person-level using methods such as the person fit indexes. Essentially, the Wald-type statistic is the standardized difference between the mean parameters ($\mu_{\mathcal{S}}$ and $\mu_{\bar{\mathcal{S}}}$) of two ability distributions estimated from the responses to two disjointed sets of items, namely, the suspected and the unsuspected sets. The generic correlation between $\mu_{\mathcal{S}}$ and $\mu_{\bar{\mathcal{S}}}$, induced by the within-subject effect, is taken into account by the suggested approximation (2.6) to the standard deviation of the difference (i.e., $\sigma(\mu_{\mathcal{S}} - \mu_{\bar{\mathcal{S}}})$).

The type-I error rate of the suggested test is close to the nominal level across various conditions, indicating the validity of using the normal distribution to approximate the large-sample distribution of the test statistic. Results of the power analysis reveal the effectiveness of the suggested test. The feasibility of the test in practice is illustrated by applying it to three real-world datasets. The analysis using real data also suggests an useful implication—

that is, the test can be used as an omnibus test to flag the suspected groups, followed by further investigating each individual of the flagged groups by means of the person-fit analysis.

Several limitations need to be considered when the suggested test is used. First, more discretion ought to be exerted when the equal-variance assumption is violated, even though the sensitivity analysis suggests its robustness to this violation. A check on the equal-variance before conducting the Wald test is highly recommended. Second, separating test items into the suspected and the unsuspected using external information is not a very systematic way, compared to integrating indicators of the suspected items into the measurement model. For example, C. Wang, Xu, Shang, and Kuncel (2018) proposed a mixture hierarchical model on responses and response time (i.e., two measurement models for responses and response time respectively at the first level, and a covariance structure of the latent variables at the second level), wherein an augmented latent indicator, δij (indicating whether item j is compromised by test-taker i), is assumed to follow the Bernoulli distribution with $\pi_j = P(\delta ij)$. Note that δij is dependent on an item-level parameter π_j which can be easily generalized as a group-specific parameter π_{jg} in accordance with the purpose of the current study. Third, item parameters are assumed as known in this study, indicating that the sampling error carried over from the use of estimated item parameters is overlooked. It will yield overstated accuracy of the estimation of the mean and the standard deviation parameters. Although this effect becomes negligible when the size of the sample used for calibration is sufficiently large, it should be taken into account in future studies. One should notice the sizes of groups examined in the simulations of this study are not the size of the calibration sample. Last, the group sizes used in the simulation studies might overestimate those in practice. The performance of the Wald test, compared with other comparable methods, can be investigated in future research.

The utility of the Wald test with respect to being easily adapted to the conventional IRT-based response modeling often geared towards frequentist should not be voided by the limitations mentioned above. Plus, the Wald test is not restricted in detecting the aberrant group responses. The Wald statistic provides an alternative approach for measuring the group-level gain (e.g., the change between the pre-test and the post-test) in the scale provided by IRT models. The technical details of the suggested Wald test bear a close resemblance to the methods developed by the "ability-gain" studies Embretson (1991); Fischer (2003); W.-C. Wang and Chen (2004). Such studies advocated measuring the gain in the latent scale instead of the raw score used under the Classic Testing Theory (CTT) due to its superiority in terms of reliability.

Appendix A

Throughout the section, functions are assumed to be as regular as needed. In other words, when we write a derivation or an integral, we assume that they exist. Estimators for unknown parameters are assumed to be interior points lying in the corresponding parameter space. Notations are the same with those used in the main sections.

By taking the derivative of the marginal log-likelihood function with respect to the mean parameter, we have

$$\ell'_{\mathcal{S}}(\hat{\mu}_{\mathcal{S}}) = \frac{\partial \ell_{\mathcal{S}}(\mu)}{\partial \mu}|_{\mu=\hat{\mu}_{\mathcal{S}}} = 0.$$

Under the null that $\mu_{\mathcal{S}} = \mu_{\bar{\mathcal{S}}} = \mu_0$ and the regularity conditions, the MLE ($\hat{\mu}_{\mathcal{S}}$ and $\hat{\mu}_{\bar{\mathcal{S}}}$) must be consistent with μ_0. A first-order Taylor series approximation of $\ell'_{\mathcal{S}}(\hat{\mu}_{\mathcal{S}})$ about μ_0 is developed, i.e.,

$$\ell'_{\mathcal{S}}(\hat{\mu}_{\mathcal{S}}) \approx \ell'_{\mathcal{S}}(\mu_0) + \ell''_{\mathcal{S}}(\mu_0)(\hat{\mu}_{\mathcal{S}} - \mu_0).$$

Given $\ell'_{\mathcal{S}}(\hat{\mu}_{\mathcal{S}}) = 0$,

$$\hat{\mu}_{\mathcal{S}} \approx \mu_0 - \frac{\ell'_{\mathcal{S}}(\mu_0)}{\ell''_{\mathcal{S}}(\mu_0)}.$$

Similarly,

$$\hat{\mu}_{\bar{\mathcal{S}}} \approx \mu_0 - \frac{\ell'_{\bar{\mathcal{S}}}(\mu_0)}{\ell''_{\bar{\mathcal{S}}}(\mu_0)}.$$

The last two approximations lead to

$$\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}} \approx -\frac{\ell'_{\mathcal{S}}(\mu_0)}{\ell''_{\mathcal{S}}(\mu_0)} + \frac{\ell'_{\bar{\mathcal{S}}}(\mu_0)}{\ell''_{\bar{\mathcal{S}}}(\mu_0)}.$$

Notice that

$$\ell'_{\mathcal{S}}(\mu_0) = \sum_{i=1}^{I} \ell'(\mathbf{Y}_{i|\mathcal{S}}|\mu_0) = \sum_{i=1}^{I} \frac{\partial \ell(\mathbf{Y}_{i|\mathcal{S}}|\mu)}{\partial \mu}|_{\mu=\mu_0}. \tag{2.8}$$

$\{\mathbf{Y}_{i|\mathcal{S}}|i \in 1, ..., N\}$ are $i.i.d$ and follow the MVB distribution as mentioned in the main sections, suggesting that $\{\frac{\partial \ell(\mathbf{Y}_{i|\mathcal{S}}|\mu)}{\partial \mu}|_{\mu=\mu_0}|i \in 1, ..., N\}$ are $i.i.d$ as well. According to the Central Limited Theorem, $\ell'_{\mathcal{S}}(\mu_0)$ will converge in distribution to a normal distribution as I increases, that is,

$$\ell'_{\mathcal{S}}(\mu_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, N\mathcal{I}_{\mathcal{S}}(\mu_0)\right), \tag{2.9}$$

where $\mathcal{I}_{\mathcal{S}}(\mu_0)$ is the Fisher information about μ_0 based on an individual response vector of the suspected items, i.e.,

$$\mathcal{I}_{\mathcal{S}}(\mu_0) = \mathrm{Var}[\ell'_{\mathcal{S}}(\mathbf{Y}_{i|\mathcal{S}}|\mu_0)]. \tag{2.10}$$

Similarly,

$$\ell'_{\bar{\mathcal{S}}}(\mu_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, N\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0)\right). \tag{2.11}$$

Under the regularity conditions, the Law of Large Number gives that

$$\ell''_{\mathcal{S}}(\mu_0) \xrightarrow{P} N\mathcal{I}_{\mathcal{S}}(\mu_0) \qquad\qquad \ell''_{\bar{\mathcal{S}}}(\mu_0) \xrightarrow{P} N\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0). \tag{2.12}$$

Detailed proofs of (2.9) - (2.12) can be found in statistical texts such as Hogg, Mckean, and Craig (2013). Together (2.9) - (2.12) with the Slutsky's theorem Casella and Berger (2001), we have

$$-\frac{\ell'_{\mathcal{S}}(\mu_0)}{\ell''_{\mathcal{S}}(\mu_0)} + \frac{\ell'_{\bar{\mathcal{S}}}(\mu_0)}{\ell''_{\bar{\mathcal{S}}}(\mu_0)} \xrightarrow{\mathcal{D}} -\frac{X}{c_1} + \frac{Y}{c_2},$$

where X and Y follow $\mathcal{N}(0, N\mathcal{I}_{\mathcal{S}}(\mu_0))$ and $\mathcal{N}(0, N\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0))$ respectively; c_1 and c_2 are constants, where $c_1 = N\mathcal{I}_{\mathcal{S}}(\mu_0)$ and $c_2 = N\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0)$. As a result, we obtain

$$\begin{aligned} \operatorname{Var}(\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}}) &\approx \operatorname{Var}\left[-\frac{\ell'_{\mathcal{S}}(\mu_0)}{\ell''_{\mathcal{S}}(\mu_0)} + \frac{\ell'_{\bar{\mathcal{S}}}(\mu_0)}{\ell''_{\bar{\mathcal{S}}}(\mu_0)}\right] \\ &\approx \frac{1}{N\mathcal{I}_{\mathcal{S}}(\mu_0)} + \frac{1}{N\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0)} - \frac{2\mathrm{Cov}[\ell'_{\mathcal{S}}(\mu_0), \ell'_{\bar{\mathcal{S}}}(\mu_0)]}{I^2\mathcal{I}_{\mathcal{S}}(\mu_0)\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0)}. \end{aligned} \tag{2.13}$$

$\mathcal{I}_{\mathcal{S}}(\mu_0)$ and $\mathcal{I}_{\bar{\mathcal{S}}}(\mu_0)$ in practice can be computed using the observed variance of $\ell'(\boldsymbol{Y}_{i|\mathcal{S}}|\mu_0)$ and $\ell'(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\mu_0)$. $\mathrm{Cov}\left[\ell'_{\mathcal{S}}(\mu_0), \ell'_{\bar{\mathcal{S}}}(\mu_0)\right] = \mathrm{Cov}\left[\sum_{i=1}^{I} \ell'(\boldsymbol{Y}_{i|\mathcal{S}}|\mu_0), \sum_{i'=1}^{I} \ell'(\boldsymbol{Y}_{i'|\bar{\mathcal{S}}}|\mu_0)\right]$ according to (2.8). $\boldsymbol{Y}_{i|\mathcal{S}}$ and $\boldsymbol{Y}_{i'|\bar{\mathcal{S}}}$ are independent with each other when $i \neq i'$. The independence does not hold true for the two sets of responses from a same test-taker, i.e., when $i = i'$. Therefore,

$$\mathrm{Cov}\left[\ell'_{\mathcal{S}}(\mu_0), \ell'_{\bar{\mathcal{S}}}(\mu_0)\right] = \sum_{i=1}^{I} \mathrm{Cov}\left[\ell'(\boldsymbol{Y}_{i|\mathcal{S}}|\mu_0), \ell'(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\mu_0)\right] = I\mathrm{Cov}\left[\ell'(\boldsymbol{Y}_{i|\mathcal{S}}|\mu_0), \ell'(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\mu_0)\right].$$

$\mathrm{Cov}[\ell'(\boldsymbol{Y}_{i|\mathcal{S}}|\mu_0), \ell'(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\mu_0)]$ is empirically computed by the sample covariance between $\ell'(\boldsymbol{Y}_{i|\mathcal{S}}|\mu_0)$ and $\ell'(\boldsymbol{Y}_{i|\bar{\mathcal{S}}}|\mu_0)$. In practice, the unknown μ_0 is valued at $\hat{\mu}_{\bar{\mathcal{S}}}$ obtained from responses to the unsuspected items. Together these arguments with (2.13), we have (2.6) to approximate the $\sigma(\hat{\mu}_{\mathcal{S}} - \hat{\mu}_{\bar{\mathcal{S}}})$.

Chapter 3

Global- and Item-level Fit Indexes for Cognitive Diagnostic Models

3.1 Introduction

One of the primary goals in cognitive diagnosis is to use the item responses from a cognitive diagnostic assessment to make inferences about what skills a test-taker has. Much of the research to date has focused on the parametric inference made under cognitive diagnosis models (CDMs), which requires that the parametric model does an adequate job of describing the item response distribution of the population of examinees being studied. Given the importance of model-data fit, it is necessary to have methods for investigating the ability of a model to fit observed data from an assessment.

Misfit for CDMs stems from a variety of sources. First, incorrectly specifying the model parameterization (e.g., DINA v.s. DINO) is a major source of misfit. Second misfit might be prompted by violating the assumptions of CDMs. For example, the *local independence* assumption presumes that items on the assessment are conditionally independent given the skills being measured, that is, given a specific latent attribute class. Yet it could be too strong to fit the actual data. Third, there are some certain types of misfit for CDMs. For instance, the item-attribute/item-skill (e.g., Q-matrix) and the structure of latent attribute pattern (e.g., the number of attributes and the hierarchy among skills) are another source of misfit for CDMs. Given these potential misspecification and misfit, users of CDMs need tools to investigate model-data misfit from a variety of angles.

In this chapter we separate model fit indexes into four categories defined by two aspects of the indexes: (1) the level of the fit analysis, i.e, global/test-level versus item-level; and (2) the choice of the alternative model for comparison, i.e., an alternative CDM (relative fit), or a saturated categorical model (absolute fit).

Global model fit has been a major focus for recent research (de la Torre & Douglas, 2008; Sinharay & Almond, 2007). In this category, global relative fit utilizes conventional information-based indexes to conduct model selection. In contrast, global absolute fit attempts to assess how exact the model reproduces the observed data by examining squared-residual based statistics (e.g. model-level χ^2, G^2 and root mean square error of approximation, RMSEA) or non-inferential Indexes (e.g. mean absolute difference, MAD). Typically, these measures can serve as general-purpose statistics to test the model assumptions such as specification of the model parametric form, the *local independence*, specification of the Q-matrix and the dimensionality.

Additional attention should be drawn on the issue of Q-matrix specification. Q-matrix is often subjectively constructed by domain experts and could be misspecified, sometime resulting in model misfit. Q-matrix refinement and validation methods have shown promising empirical performance in addressing this concern (de la Torre & Chiu, 2016; Chiu, 2013). However, the problem of Q-matrix misspecification and refinement should not be isolated from the issues on Q-matrix learning and identification. An integrated view of these problems is helpful to the understanding of the model-data fit analysis for CDMs.

Item-level fit analysis, often referred as to item fit analysis, focuses on "local" misfit caused by the misspecification of the parametric form of an individual or subsets of items. Item fit analysis allows practitioners to identify aberrant items and provides guidance about how to refine the measurement instrument. This use of item fit analysis has been supported by recent empirical studies showing that the assessment with items assumed to follow different models (e.g., including both DINA and DINO items), instead of uniformly having a single form, might better fit the real data (de la Torre & Lee, 2013; de la Torre, van der Ark, & Rossi, 2018). To achieve the refinement, item-level relative fit indexes offer a way to compare nested models such as Likelihood Ratio (LR), Wald (W), and Lagrange multiplier (LM) tests. Absolute fit indexes can be adapted to the item-level as well. For example, item-level goodness-of-fit statistics (Orlando & Thissen, 2000; C. Wang, Shu, Shang, & Xu, 2015) are constructed on the basis of the squared residual of observed and expected proportion of correctness that are obtained by grouping respondents. Different grouping strategies lead to various types of fit statistics, which has been a focus in recent studies. Item-level absolute fit statistics can also be extended to detect misfit for item pairs or triplets. It is particularly useful if one is interested in locating the source of misfit and taking remedial action when the global model test identifies the existence of overall misfit and *local dependence* is the potential culprit.

It's also worth mentioning the person-fit analysis that is not discussed in this chapter, offering another perspective to investigate model-data misfit. Person-fit methods are concerned with identifying misfit in individual response vectors that present atypical test-taking behaviors such as cheating and speeding. Several person-fit Indexes and tests have been proposed particularly for CDMs such as the hierarchy consistency index (Cui & Leighton, 2009) and the generalized LR test (Liu, Douglas, & Henson, 2009). Person-fit analysis developed for other latent variable models such as the item response theory (IRT) can also be employed for CDMs (Meijer & Sijtsma, 2001).

This chapter restricts its focus on four categories of indexes. After a review of indexes, the use of several selected indexes in practice is illustrated by analyzing on a real data. For each category of indexes, pros and cons are summarized based on results from current simulation studies. General guidance about which fit indexes should be used under what circumstances is provided as well.

3.2 The Model Framework

This chapter employs the generalized DINA (G-DINA) model (de la Torre, 2011) as the basic framework to discuss model fit methods. As other general frameworks of CDMs such as the general diagnostic model (von Davier, 2008) and the log-linear CDM (LCDM) (Henson, Templin, & Willse, 2009), the G-DINA model relates several CDMs by its flexible parameterization.

The G-DINA model requires a $K \times D$ Q-matrix (with binary elements $\{q_{kd}\}$), where K indicates the number of items and D represents the number of attributes. The required number of attributes for item k can be denoted D_k^*, where $D_k^* = \sum_{d=1}^{D} q_{kd}$. Such a representation efficiently reduces the attribute vector of item k from $\boldsymbol{a}_l = (a_{l1}, a_{l2}..., a_{lD})$ to $\boldsymbol{a}^\star_{lk} = (a^\star_{l1}, a^\star_{l2}, \ldots, a^\star_{lD^*_i})$, where the number of classes partitioned by item k is reduced from 2^D to $2^{D^*_k}$. For example, if $D = 3$ and the k^{th} has q-vector $\boldsymbol{q}_k = (1,1,0)^\top$, then the full attributes vectors $\boldsymbol{a}_l = (0,1,0)$ and $\boldsymbol{a}_{l'} = (1,1,0)$ are simplified as reduced vectors $\boldsymbol{a}^\star_{lk} = (0,1)$ and $\boldsymbol{a}^\star_{l'k} = (1,1)$. The probability of respondents with latent profile $\boldsymbol{a}^\star_{lk}$ answering item k correctly is denoted by $P(X_k = 1|\boldsymbol{a}^\star_{lk}) = P(\boldsymbol{a}^\star_{lk})$, more specifically,

$$P(\boldsymbol{a}^*_{lk}) = \delta_{k0} + \sum_{d=1}^{D^*_i} \delta_{kd} a^\star_{ld} + \sum_{d=1}^{D^*_k} \sum_{d'=d+1}^{D^*_k} \delta_{kdd'} a^\star_{ld} a^\star_{ld'} + \cdots + \delta_{k12\ldots D^*_k} \prod_{d=1}^{D^*_k} a^\star_{ld}, \tag{3.1}$$

where δ_{k0} is the intercept for item i; δ_{kd} is the main effect due to a_d; $\delta_{kdd'}$ and $\delta_{k12\ldots D^*_k}$ are interactions for the two-way and other higher orders among $a_1, \ldots, a_{D^*_k}$. Conventionally, the monotonicity constraints are imposed on item parameters to make sure that subjects owning more skills have a higher probability of answering an item correctly than those who own fewer skills. Notice that (3.1) uses the *identity link* function that can be modified through the use of other transform functions such as the *logistic link* and the *log link*.

It is not hard to tell the flexibility of such a formulation. For example, the DINA model can be obtained by using identity-link function and setting all parameters to 0 except for δ_{k0} and $\delta_{i12\ldots D^*_k}$; in which case the guessing parameter follows $g_k = \delta_{k0}$ and the slipping parameter satisfies $s_k = 1 - \delta_{k0} + \delta_{k12\ldots D^*_k}$. Notice that the flexibility enables us to summarize and estimate the parameters of multiple CDMs by a single parametric framework. The G-DINA model provides a convenient basis for comparing nested models and allows us to examine one item at a time.

3.3 Relative Fit Indexes

Relative fit Indexes evaluate the fit of a model compared to some competing models. In the following two subsections, we first review the Indexes working for the global-level fit and then look at how some of them can be used at the item-level.

3.3.1 Global-level

One way to evaluate the comparative fit of a model relative to a competing model, when it is a nested model, is the likelihood ratio test (LRT). A nested model is one that can be defined by enforcing some constraints on some of the model parameters. For example, within the G-DINA framework, the DINA model is nested within the G-DINA model because it can be obtained by setting all coefficients other than the intercept and the highest-order interaction term equal to zero. The LRT compares the fit of the two models by comparing the log-likelihoods ℓ_r and ℓ_f evaluated at the maximum likelihood estimates (MLEs) for the reduced and full models respectively, where the log-likelihood is defined as

$$\ell(\boldsymbol{X}|\boldsymbol{\delta},\boldsymbol{\gamma}) = \sum_{n=1}^{N} \log \sum_{l=1}^{L} p(\boldsymbol{a}_l|\boldsymbol{\gamma}) \prod_{k=1}^{K} P(\boldsymbol{a}_{lk}^{\star})^{X_{nk}} \left[1 - P(\boldsymbol{a}_{lk}^{\star})\right]^{(1-X_{nk})}, \tag{3.2}$$

where N is the number of participants and $L = 2^D$; $p(\boldsymbol{a}_l|\boldsymbol{\gamma})$ is the prior probability of $\boldsymbol{a}_l$. The item response probability $P(\boldsymbol{a}_{lk}^{\star})$ is obtained by compressing $\boldsymbol{a}_{lk}$ as what we show in previous. The maximum likelihood estimates of the item parameter vector, $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_K)$, and the latent class proportion parameters $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{L'})$ ($L' = L$ and $p(\boldsymbol{a}_l|\boldsymbol{\gamma}) = \gamma_l$ if an unrestricted attribute space is assumed) can be estimated using an expectation-maximization (EM) algorithm (de la Torre, 2011; George, Ünlü, Kiefer, Robitzsch, & Groß, 2016).

The likelihood ratio test statistic that is typically used is two times the difference between the log-likelihoods,

$$\lambda = 2\left(\ell_f(\boldsymbol{X}|\boldsymbol{\delta}_f, \boldsymbol{\gamma}_f) - \ell_c(\boldsymbol{X}|\boldsymbol{\delta}_c, \boldsymbol{\gamma}_c)\right),$$

, in the case where observations have been randomly sampled, the statistic λ is approximately chi-squared distributed when the reduced model is the correct model; the degrees of freedom of the distribution is equal to the difference in the number of parameters in the two models. For example, if the full model is the G-DINA model and the reduced model is the DINA, the number of parameters are $p_f = \sum_{k=1}^{K} 2^{D_k^*} + L - 1$ and $p_r = 2K + L - 1$ respectively.

The likelihood ratio test has a couple of limitations. First, according to the old adage, 'all models are wrong', the LRT tends to find evidence against simpler models when the sample size N is large. Second, the likelihood ratio test requires the reduced model to be nested within the full model framework.

Two information-based criteria attempting to address these issues are Akaike's information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz, 1978), which are defined as

$$AIC = -2\ell(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}}) + 2p$$

$$BIC = -2\ell(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}}) + p\ln(N),$$

To use AIC and/or BIC for model evaluation, the user should estimate multiple competing models. In both cases, the model that should be selected is the one that minimizes the criterion. So, if one is interested in whether the DINA model fit a specific data set, the researchers would fit the DINA model and other candidates from the G-DINA framework and then check if the AIC and/or BIC for the DINA model is the smallest.

The difference between the penalty terms makes the BIC penalize the model with a larger number of parameters more than the AIC does. This is partially due to the purposes of each; AIC attempts to find the model that best predicts future observations, whereas BIC attempts to quantify evidence for a model in model-selection problems. Kunina-Habenicht, Rupp, and Wilhelm (2012) found that AIC and BIC are effective in selecting the model with a correctly specified Q-matrix against those with misspecified Q-matrices within the framework of the log-linear CDM. J. Chen, de la Torre, and Zhang (2013) showed that AIC and BIC perform well in selecting among nested models within the G-DINA framework.

Another way to compare non-nested models is the log-penalty index (Gilula & Haberman, 1994) which is obtained by dividing the AIC by the number of observations in the sample. It is more like the BIC penalizing the number of parameters while accounting for the sample size. The index has been used in comparing models within the framework of GDM (von Davier, 2008).

The likelihood ratio test, AIC, BIC and log-penalty index all require MLEs for the model parameters, and thus are used in frequentist applications. The deviance information criterion (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) and the Bayes factor (Kass & Raftery, 1995), in contrast, are applicable for global relative fit within the Bayesian modeling framework. The DIC is defined as

$$DIC = \bar{D} + p_D,$$

where $\bar{D}$ is the expectation of $-2\ell(\boldsymbol{\delta}, \boldsymbol{\gamma})$ over the joint posterior distribution of $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ given the observed assessment data. The quantity $p_D = \bar{D} - 2\ell(\bar{\boldsymbol{\delta}}, \bar{\boldsymbol{\gamma}})$, where $(\bar{\boldsymbol{\delta}}, \bar{\boldsymbol{\gamma}})$ are the posterior mean vectors is a measure of the complexity of the Bayesian model.

The Bayes factor is the Bayesian analog to the frequentist likelihood ratio test.

$$BF_{12} = \frac{P(\boldsymbol{X}|\mathcal{M}_1)}{P(\boldsymbol{X}|\mathcal{M}_2)}$$

where

$$P(\boldsymbol{X}|\mathcal{M}_m) = \int \exp\left[\ell(\boldsymbol{\delta}_m, \boldsymbol{\gamma}_m)\right] p(\boldsymbol{\delta}_m, \boldsymbol{\gamma}_m|\mathcal{M}_m) d\boldsymbol{\delta}_m d\boldsymbol{\gamma}_m$$

and $p(\boldsymbol{\delta}_m, \boldsymbol{\gamma}_m|\mathcal{M}_m)$ is the joint prior density of parameters from the m^{th} model. In most applications exact calculation of the Bayes factor is difficult or impossible. A possible approach for approximating the marginal likelihoods needed to calculate the Bayes factor is with the Laplace-Metropolis estimator as proposed by Raftery (1996).

In psychometrics, DIC and Bayes factors have been suggested and used in the model comparison for CDMs de la Torre and Douglas (2008, 2004); Sinharay and Almond (2007). For example, de la Torre and Douglas (2004, 2008) implemented the Bayes factor to compare the Higher-order DINA and multiple-strategy DINA models against the traditional DINA model.

3.3.2 Item-level

The G-DINA framework allows us to evaluate the parametric form of an assumed CDM used at the item-level by performing specific hypothesis tests. In these hypothesis tests, the null hypothesis (H_0) assumes the reduced model (e.g., DINA) is correct and the alternative (H_1) states that the general (or full) model (e.g., G-DINA) is correct. The size of parameter space for the full model is determined by the number of skills required by the item. Let's say, for instance, the Q-matrix specifies up to 3 skills but the item only requires 2 skills.

The full model of the item can have up to 4 parameters according to the equation (3.1): an "intercept", two "main effect", and an "interaction".

The likelihood ratio (LR) introduced earlier for model-level fit evaluation could be applied to item-level fit by fitting the assumed model as the reduced model, and a second model that assumes a G-DINA structure for that item. To check the fit of all K items, it would require estimating $K+1$ models—namely, a reduced model for each item, and a separate "full" model; this somewhat limits the use of the likelihood ratio statistic for item-level evaluation when K is large.

Unlike the LR statistic and testing procedure, the Lagrange multiplier (LM), or score test only requires estimation of the reduced model, which makes it particularly useful for evaluating item-level fit of a model. The general idea of the score test is that if the null hypothesis is correct, then the first derivative of the full model likelihood evaluated at the reduced model maximum likelihood estimates should be close to zero. If $\hat{\boldsymbol{\delta}}_k^0$ denotes the maximum likelihood estimator of the item parameters for item k under the reduced model, then the LM statistic is

$$LM = \left[\frac{\partial \ell_f(\boldsymbol{\delta}_k)}{\partial \boldsymbol{\delta}_k} \mid_{\boldsymbol{\delta}_k = \hat{\boldsymbol{\delta}}_k^0}\right]^T \boldsymbol{I}^{-1}(\boldsymbol{\delta}_k) \left[\frac{\partial \ell_f(\boldsymbol{\delta}_k)}{\partial \boldsymbol{\delta}_k} \mid_{\boldsymbol{\delta}_k = \hat{\boldsymbol{\delta}}_k^0}\right], \tag{3.3}$$

where $\boldsymbol{I}(\boldsymbol{\delta}_k) = V\left[\frac{\partial \ell_f(\boldsymbol{\delta}_k)}{\partial \boldsymbol{\delta}_k} \mid_{\boldsymbol{\delta}_k = \hat{\boldsymbol{\delta}}_k^0}\right]$ is the information matrix (from the full model) for the item parameter vector $\boldsymbol{\delta}_k$ evaluating at $\hat{\boldsymbol{\delta}}_k^0$; in practice the information matrix is approximated with the observed information matrix $\boldsymbol{I}(\hat{\boldsymbol{\delta}}_k^0)$. Under the null hypothesis the distribution of the LM approach, the chi-squared distribution with $p_f - p_r$ degrees of freedom (df), where p_f and p_r, by an abuse of the notation, denote the number of item parameters for the item k under the full and reduced models.

The likelihood ratio test and the Lagrange multiplier test are asymptotically equivalent to one another, so the results tend to be similar for large sample sizes. A third asymptotically equivalent test statistic is the Wald test statistic. The Wald test for item-level model fit assessment requires fitting the full model (e.g., G-DINA) in order to evaluate the fit of the reduced model (e.g., DINA). As discussed earlier, the DINA model can be obtained from the G-DINA model by assuming all parameters other than the intercept and the highest-order interaction term are equal to zero. For example, suppose we have an item measuring two skills. Then the full model parameter vector is $\boldsymbol{\delta}_k = (\delta_{k0}, \delta_{k1}, \delta_{k2}, \delta_{k12})^\top$; the test to evaluate fit of the DINA model assumes a null hypothesis of the form $H_0 : \boldsymbol{\delta}_k = (\delta_{k0}, 0, 0, \delta_{k12})^\top$, or equivalently $H_0 : \boldsymbol{R}_k\boldsymbol{\delta}_k = (0, 0)^\top$, where $\boldsymbol{R}_k$ is the restriction matrix

$$\boldsymbol{R}_k = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

For general models R_k is a $(p_f - p_r) \times p_f$ matrix describing the null model restrictions; see de la Torre (2011) for examples. The Wald test is then defined

$$W = \left[\boldsymbol{R}_k\hat{\boldsymbol{\delta}}_k^1\right]^\top \left[\boldsymbol{R}_k V(\hat{\boldsymbol{\delta}}_k^1)\boldsymbol{R}_k^\top\right]^{-1} \left[\boldsymbol{R}_k\hat{\boldsymbol{\delta}}_k^1\right], \tag{3.4}$$

where $\hat{\boldsymbol{\delta}}_k^1$ is the maximum likelihood estimator under the full model (H_1) and $V(\hat{\boldsymbol{\delta}}_k^1)$. It should be noted that $V(\hat{\boldsymbol{\delta}}_k^1)$ is the sub-matrix of the covariance matrix of the MLEs for all item parameters and latent attribute distribution parameters. The covariance matrix is usually approximated with the inverse of the observed information matrix. The asymptotic distribution under the null hypothesis is also $\chi^2_{(p_f - p_r)}$.

Simulation studies by de la Torre and Lee (2013) and Sorrel, Abad, Olea, de la Torre, and Barrada (2017) showed the statistics have accurate Type I error rates and high power with large N and small D for typical significance levels. Sorrel et al. (2017) found that the likelihood ratio and Wald tests perform better than the Lagrange multiplier test in terms of the Type I error and power across cases with $N \leq 1000$, $K \leq 36$ and $D = 4$. However, all statistics were found to be highly affected when items have low discrimination (Sorrel et al., 2017; Ma, Iaconangelo, & de la Torre, 2016) .

3.4 Absolute Fit Indexes

This section begins with a review of the global-level statistics, which is followed by introducing item-level statistics. A review of posterior predictive methods that assess model-data misfit using the Bayesian approach is included as the end.

3.4.1 Global-level

Classical goodness-of-fit (GOF) statistics such as Pearson's χ^2 and the likelihood ratio G^2 are fundamental overall fit Indexes in categorical data analysis. For a test with K dichotomous items,

$$\chi^2 = N \sum_{c=1}^{2^K} \frac{(p_c - \hat{\pi}_c)^2}{\hat{\pi}_c} \quad \text{and} \quad G^2 = 2N \sum_{c=1}^{2^K} p_c \ln(\frac{p_c}{\hat{\pi}_c})$$

where p_c and $\hat{\pi}_c$ are the observed and model-based expected proportions for one cell c in the 2^K contingency table (for all possible response patterns). The model-based proportions, $\hat{\pi}_c$, is calculated by the marginal likelihood in the right-hand side of (3.2) with estimated parameters. For small K and under the null hypothesis that the assumed CDM is the correct model, the statistics follow the chi-square distribution with $2^K - p - 1$ *df*, where p is the total number of model parameters.

These full-information statistics suffer from the problem of sparsity when K is large and N is small, which can create unknown asymptotic distributions of the statistics. One could use the resampling and bootstrapping techniques to obtain empirical p-values, yet prohibited by the computational overhead. Maydeu-Olivares and Joe (2005) introduced the limited-information family of statistics to address the issues for IRT models. Hansen, Cai, Monroe, and Li (2016) and Liu, Tian, and Xin (2016) implemented statistics in this family to evaluate global fit for CDMs.

The idea is to utilize the up-to-r^{th}-order moments, $\boldsymbol{\pi}_r$, rather than the proportions of all possible response patterns (or referred as all cells in the contingency table, $\boldsymbol{\pi}$, to formulate the fit statistic. For instance,

$$\boldsymbol{\pi_2} = \begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \dot{\pi}_{12} \\ \dot{\pi}_{13} \\ \dot{\pi}_{23} \end{pmatrix} = \boldsymbol{T}_2\boldsymbol{\pi} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \pi_{000} \\ \pi_{100} \\ \pi_{010} \\ \pi_{001} \\ \pi_{110} \\ \pi_{101} \\ \pi_{011} \\ \pi_{111} \end{pmatrix} \tag{3.5}$$

for the case of $K = 3$; $\boldsymbol{T}_2$ is the matrix transforming $\boldsymbol{\pi}$ to $\boldsymbol{\pi}_2$. The limited-information statistic M_r is written as

$$M_r = N(\boldsymbol{p}_r - \hat{\boldsymbol{\pi}}_r)^\top \hat{\boldsymbol{C}}_r(\boldsymbol{p}_r - \hat{\boldsymbol{\pi}}_r)$$

on the basis of the up-to-r^{th} moments. Given a specified CDM model being the null model, M_r follows the chi-square distribution with $df = s_r - p$, where $s_r = \sum_{i=1}^{r} \binom{K}{i}$ is the number of elements in $\boldsymbol{\pi}_r$. The detailed derivation of $\hat{\boldsymbol{C}}_r$ is described in Maydeu-Olivares and Joe (2005).

Hansen et al. (2016) and Liu et al. (2016) examined the limited information statistic for the evaluation of CDMs. Simulations in both studies show that M_2 has more stable performance in detecting misfit simulated from Q-matrix misspecification than χ^2 and G^2 for moderate sample sizes. Hansen et al. (2016) also found that M_2 is sensitive to misfit from item-level model misspecification and to violations of local independence, but insensitive to the misspecification of the higher-order structure of the attributes.

One of the shortcomings of GOF statistics is that they treat the model under the null hypothesis as the desired model, and the model under the alternative hypothesis as the saturated model. The true model in practice is likely to be more complex than any assumed model, and therefore will be rejected with a sufficiently large sample size. To deal with this issue, Browne and Cudeck (1992), introduced the root mean squared error of approximation (RMSEA), which attempts to measure the discrepancy between the population ($\boldsymbol{\pi}_T$) and the null model ($\boldsymbol{\pi}_0$) probability vectors.

$$\text{RMSEA} = \sqrt{\max\left(\frac{\hat{\chi}^2 - df}{N \times df},\ 0\right)}$$

where $\hat{\chi}^2$ is the observed χ^2 statistic for the data set. Maydeu-Olivares and Joe (2014) gives the limited-formation version that is

$$\text{RMSEA}_r = \sqrt{\max\left(\frac{\hat{M}_r - df_r}{N \times df_r},\ 0\right)}.$$

The 90% of confidence interval of RMSEA_r is derived from the non-central chi-square distribution $F_{\chi^2}(\hat{M}_r; df_r)$. Maydeu-Olivares and Joe (2014) shown that RMSEA_r ($r \leq 3$) has more accurate confidence intervals than RMSEA when $2^K > 300$ for simulations generated under dichotomous IRT models.

In practice, the cut-off values of RMSEA are suggested to determine the degree of fit. For example, Oliveri and von Davier (2011) suggested using $\text{RMSEA}_1 > 0.1$ as poor fit when they measure the item-level misfit for the PISA (Programme for international Student Assessment) data with the GDM; Liu et al. (2016) recommended the cut-off values (less than) 0.030 and 0.045 for RMSEA_2 as an "excellent" and a "good" fit under the LCDM.

Item-level and item-pairwise fit indexes were also used to assess the overall misfit in the current literature. For example:

$$
\begin{aligned}
\text{MAD}_k &= |\dot{p}_k - \dot{\hat{\pi}}_k|, \\
\chi^2_{kk'} &= N \sum_{x_k=0}^{1} \sum_{x_{j'}=0}^{1} \frac{(p_{x_k x_{k'}} - \hat{\pi}_{x_k x_{k'}})^2}{\hat{\pi}_{x_k x_{k'}}},
\end{aligned}
$$

where $\dot{\hat{\pi}}_k$ is the model-implied proportion of answering the item k correctly; $\hat{\pi}_{x_k x_{k'}}$ is the expected probability of cell in the bivariate table for item k and k'; $\dot{p}_k$ and $p_{x_k x_{k'}}$ are observed probabilities. In addition, implementing the Fisher transformation of item-pair correlations and the item-pairwise log-odds ratio to assess model-data fit was studied by J. Chen et al. (2013). J. Chen et al. (2013); Lei and Li (2016) recommended to apply the aforementioned single-item or pair-wise fit indexes to assess the overall model-data fit in practice by simply averaging the results of multiple tests or conducting multiple tests with a Bonferroni-adjustment. Both studies showed that the pairwise fit indexes perform with better power in detecting the overall misfit than the single-item fit indexes.

3.4.2 Item-level

Squared-residual based statistics play a vital role in item-level fit analysis. To collect the squared residuals, we partition the test-takers into groups by certain schemes. Once the groups are given, we can calculate o_{ks} and e_{ks} denoting the observed and expected proportion of answering the item k right for the test-takers in group s. It's easy to see that different grouping schemes lead to different statistics.

Yen (1981) proposed Q_1 by grouping the test-takers according to their latent abilities. In the context of CDMs, the examinees are grouped by their attribute patterns. In practice the assignment of a subject to her latent attribute class is given by the posterior $P(\hat{\boldsymbol{a}}_l|\boldsymbol{x}_n)$ where $\hat{\boldsymbol{a}}_l$ and $\boldsymbol{x}_n$ are the attribute pattern l and response vector for subject n. Yen (1981) approximated the limiting distribution of Q_1 by the chi-square distribution with df $2^D - p_k - 1$, where p_k is the number of parameters for item k. The statistic is criticized for two points. First, some latent attribute classes are extremely rare, especially when D is large, which means that almost no test-taker will be assigned in these classes. Some researchers suggested binning the race classes to reduce the effect of sparsity. But how to bin them appropriately is still a complex question. Second, the uncertainty of the class assignment is not considered in the approximation of Q_1's limiting distribution.

$S-\chi^2_k$ and $S-G^2_k$ proposed by Orlando and Thissen (2000) address these problems. The statistics are defined as

$$S-\chi^2_k = \sum_{s=1}^{S-1} N_s \frac{(o_{ks}-e_{ks})^2}{e_{ks}(1-e_{ks})}$$

$$S-G^2_k = 2\sum_{s=1}^{S-1} N_s \left[o_{ks}\log\left(\frac{o_{ks}}{e_{ks}}\right) + (1-o_{ks})\log\left(\frac{1-o_{ks}}{1-e_{ks}}\right)\right]$$

where s indicates the group of test-takers who score s; N_s is the number of examinees in group s; o_{ks} and e_{ks} are what we define before; e_{ks} is calculated as

$$e_{ks} = \frac{\sum_{l=1}^{2^D} P(X_{ik} = 1|\boldsymbol{a}_l) P\left(S^{(-k)} = s - 1|\boldsymbol{a}_l\right) p(\boldsymbol{a}_l)}{\sum_{l=1}^{2^D} P\left(S = s|\boldsymbol{a}_l\right) p(\boldsymbol{a}_l)}.$$

$P\left(S^{(-k)} = s - 1|\boldsymbol{a}_l\right)$ is recursively computed using the algorithm developed by Lord and Wingersky (1984), as described in Orlando and Thissen (2000) in detail.

Orlando and Thissen (2000) approximated the distribution of $S - \chi_k^2$ and $S - G_k^2$ by the chi-square distribution with $df = K - 1 - p_k$, where p_k is the number of item parameters for the item k. Notice that the squared residuals are grouped by raw scores rather than by estimated latent ability groups. Simulation studies conducted by Orlando and Thissen (2000) showed that these two statistics have more sensible Type-I error than Q_1 does. However, Sorrel et al. (2017) noted that although the use of $S - \chi_k^2$ avoids the inflated Type I error, the power of $S - \chi_k^2$ is quite unacceptable in many cases when it is used to detect the item-level misfit for the G-DINA model.

To take the uncertainty of $\hat{\boldsymbol{a}}_l$ into account, C. Wang et al. (2015) suggested applying Stone's method (Stone, 2000) to Q_1. Instead of using observed counts grouped by point estimated $\hat{\boldsymbol{a}}_l$ to create squared residuals, Stone (2000) computed

$$O_{kl}^* = \sum_{n=1}^{N} x_{nk} p(\hat{\boldsymbol{a}}_l|\boldsymbol{x}_n)$$

using the posterior distribution of $\hat{\boldsymbol{a}}_l$. In this setting, the chi-square distribution is no longer a good approximation of the limiting distribution of the new statistic given the dependence among examinees introduced from $p(\hat{\boldsymbol{a}}_l|\boldsymbol{x}_n)$. A Monte Carlo resampling technique is suggested to obtain the empirical distribution of the statistic. This is the idea behind Stone's method.

Simulation studies in C. Wang et al. (2015) showed that Stone's Q_1 has more promising power and Type I error than its original counterpart to detect Q-matrix and model-type misspecification under the DINA model. One drawback of Stone's method is that it is computationally expensive.

3.4.3 Posterior predictive assessment

The posterior predictive model-checking (PPMC) method (Rubin, 1984) is one of the popular approaches within the Bayesian paradigm, not because of its intuitive appeal and ease of implementation, but more importantly, due to its strong theoretical basis.

Sinharay (2006a) argued that $S-\chi_k^2$ and $S-G_k^2$ do not have the assumed limiting distribution due to the use of item parameters estimated from ungrouped observations. Sinharay (2006a) suggested using the PPCM method, working along with Markov Chain Monte Carlo (MCMC) sampling technique, to simply sample the empirical distributions for $S-\chi_k^2$ and $S-G_k^2$ that approximate their actual posterior distributions.

Specifically, the idea behind the PPMC is to compare the observed data $\boldsymbol{x}$ against the *replicated data* $\boldsymbol{x}^{rep}$ generated from the *posterior predictive distribution*

$$p\left(\boldsymbol{x}^{rep}|\boldsymbol{x}\right) = \int p\left(\boldsymbol{x}^{rep}|\boldsymbol{\theta}\right) p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta}. \tag{3.6}$$

$\boldsymbol{\theta}$ contains $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$, or hyper-parameters according to the assumed prior(s); $p\left(\boldsymbol{x}^{rep}|\boldsymbol{\theta}\right)$ is the joint likelihood function and $p(\boldsymbol{\theta}|\boldsymbol{x})$ is the posterior distribution given the observed data.

Test quantities, sometimes referred to as *discrepancy measures*, $D(\boldsymbol{x},\ \boldsymbol{\theta})$, are defined (Gelman, Meng, & Stern, 1996) to evaluate the adequacy of a model; the lack-of-fit can be summarized by the *posterior predictive p-value* (*ppp*)

$$ppp = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{x}^{rep}} I_{[D(\boldsymbol{x},\boldsymbol{\theta}) \leq D(\boldsymbol{x}^{rep},\boldsymbol{\theta})]} p\left(\boldsymbol{x}^{rep}|\boldsymbol{\theta}\right) p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{x}^{rep} d\boldsymbol{\theta}, \tag{3.7}$$

where $I[\cdot]$ is the indicator function. The analytical difficulty in (3.6) and (3.7) can be reduced by numerically carrying out along with the MCMC steps. Model parameters $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, ..., $\boldsymbol{\theta}^{(M)}$ are simulated from the (approximate) posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ within the converged MCMC algorithm. The replicated data, $\boldsymbol{x}^{rep(m)}$, is generated from the likelihood $p\left(\boldsymbol{x}^{rep}|\boldsymbol{\theta}^{(m)}\right)$ for $m = 1, ..., M$. This process leads to M draws from the joint distribution $p(\boldsymbol{x}^{rep}, \boldsymbol{\theta}|\boldsymbol{x})$, which can then be used to approximate the *ppp* by calculating the proportion of replicated datasets having a larger value of discrepancy measure than the value computed from the observed dataset.

The choice of $D(\boldsymbol{x}, \boldsymbol{\theta})$ is vital but also flexible for the PPMC method. Sinharay and Almond (2007) suggested examining the item-fit by Q_1. C. Wang et al. (2015) employed the power-divergence (PD; a more general statistic family including Q_1) and Stone-type PD to check item-level fit. Sinharay and Almond (2007) assessed the overall fit by looking at the residual between individual raw score and expected score. GOF statistics and RMSEA mentioned above could be chosen as the discrepancy measure for detecting overall misfit.

Robins, van der Vaart, and Ventura (2000) showed that the *ppp* tends to be conservative for some choices of discrepancy measure. Similar issues have been found in C. Wang et al. (2015), indicating that the *ppp* is more conservative than its classic GOF counterparts. However, as argued by many, a conservative diagnostic with reasonable power is better than tests with unknown properties or poor Type I error rates.

Other posterior predictive based methods, such as the direct display (for overall fit) and the odds ratios (for item association/pairs fit), are not covered in this chapter. We refer readers to Sinharay (2006a) for more details about these methods which have been used in model diagnostics for Bayesian networks.

3.5 Empirical Illustration

A publicly available dataset for the 28-item Examination for the Certificate of Proficiency in English (ECPE) is analyzed in this section as an example. ECPE was developed and scored by the English Language Institute at the University of Michigan. The data has been used to investigate multidimensional cognitive attributes (Buck & Tatsuoka, 1998; Templin & Hoffman, 2013) and to examine attribute hierarchy (Templin & Bradshaw, 2014).

Previous discussions on the attribute hierarchy are noteworthy. von Davier and Haberman (2014) pointed out that the hierarchical diagnostic classification models (HDCMs; Templin and Bradshaw, 2013) are equivalent to an ordered latent class model. Additionally, Templin and Hoffman (2013) found that the HDCMs and the G-DINA models do not perform substantially better than the unidimensional two-parameter IRT model. von Davier and Haberman (2014) suggested staring with the simplest possible model rather than with a potentially overly complex model.

In this illustrative example, the hierarchy among attributes is not considered. Several common CDMs are compared using information criterion and the absolute overall fit is examined. Item-level fit is checked when the DINA framework is assumed to fit the data well.

Specifically, for the ECPE dataset, three attributes are intended to be measured: morphosyntactic rules, cohesive rules, and lexical rules (Buck & Tatsuoka, 1998). The dataset includes the responses from 2,922 test-takers and the Q-matrix of the items, which has been used in R packages G-DINA (Ma & de la Torre, 2016) and CDM (Robitzsch, Kiefer, George, & Uenlue, 2016) for an illustrative purpose.

3.5.1 Results of global fit results

3.5.1.1 Relative fit

Table 3.1 presents the performance of AIC, BIC and sample-size adjusted BIC across the saturated G-DINA, the Additive-CDM (ACDM) and a mixed form (MIX) of G-DINA and ACDM. ACDM only contains terms in (3.1) up-to main effects. For the mixed form, Item 3, 11, 12, 17 and 21 are set as the ACDM since their estimated second-order interaction coefficients are not significantly different from 0 under the G-DINA model. Non-constrained G-DINA (NC-GDINA) denotes the saturated G-DINA without monotonicity constraints.

The information criterion in Table 3.1 picks out the ACDM. It also shows that G-DINA and NC-GDINA are different models, which should be noted when choosing a model. Notice that the NC-GDINA model is probably not identified. The general discussions of the identification issue related to monotonicity constraints can be found in von Davier (2014). The NC-GDINA model is used to emphasize that the monotonicity constraints should not be ignored in model fitting and selection.

Table 3.1: Relative overall fit indexes for CDMs on the ECPE dataset

	p	AIC	BIC	sBIC
DINA	63	85813.98	86190.72	86191.24
G-DINA	81	85642.67	86127.05	86127.71
NC-GDINA	81	85639.19	86123.57	86124.24
ACDM	72	85639.01	86069.57	86070.16
MIX	76	85642.17	86096.65	86097.27

3.5.1.2 Absolute fit

Table 3.2 provides the absolute fit of ACDM, MIX, and DINA. The statistics M_2 and RMSEA_2 are limited-information based statistics as mentioned previously. The p-values for the test statistics and the 95-percent confidence intervals for the RMSEA are given in parentheses following the various statistics. The final column, $\max(\chi^2_{kk'})$, is the largest $\chi^2_{kk'}$ among all pairs of items; the p-value for the statistic is obtained by the Holm-Bonferroni procedure.

Both limited-information and item-pairwise test statistics suggest that none of the three models provide adequate fit to the data. A possible reason is the misspecification (under-specification) of the Q-matrix, which would lead to local dependence among the items. In contrast, the RMSEA suggests that all these three models adequately fit the dataset. The difference between the results from RMSEA and the results from other absolute fit analyses supports the aforementioned: absolute fit statistics, such as limited-information M_2, tend to reject the null model when sample size is large, whereas RMSEA takes the effect of sample size into consideration.

Table 3.2: Absolute overall fit indexes for CDMs on the ECPE dataset

	M_2	df	RMSEA_2	$\max(\chi^2_{kk'})$
ACDM	474.557 (.000)	325	.013 (.010, .015)	38.712 (.000)
MIX	500.841 (.000)	330	.013 (.010, .016)	39.639 (.000)
DINA	515.707 (.000)	343	.013 (.011, .015)	26.608 (.000)

3.5.2 Results of item-level fit

3.5.2.1 Relative Fit

Table 3.3 lists the chi-square statistics based on the Wald test. The Wald test, as in the first column of the table, examines the null that the item is DINA against its alternative that is the G-DINA. The second column is for the ACDM case.

The table lists the items rejected under the DINA null. Among them, Item 3, item 7 and item 21 are not rejected under the ACDM null. The df is 2 for the DINA null and 1 for the ACDM null since there are only 2 attributes required by these items.

Table 3.3: Item-level relative fit indexes for CDMs on the ECPE dataset

	DINA χ^2_{Wald}	ACDM χ^2_{Wald}
Item 1	39.823 (.000)	26.342 (.000)
Item 3	23.871 (.000)	0.102 (.750)
Item 7	213.444 (.000)	36.029 (.000)
Item 11	98.963 (.000)	1.173 (0.279)
Item 12	201.990 (.000)	201.607 (.000)
Item 16	106.427 (.000)	5.966 (.015)
Item 17	27.508 (.000)	4.194 (.041)
Item 20	76.782 (.000)	37.586 (.000)
Item 21	130.965 (.000)	2.399 (.121)

3.5.2.2 Absolute fit

Table 3.4 shows the absolute fit results. RMSEA$_k$ (Oliveri & von Davier, 2011) is the item-level RMSEA based on RMSEA$_1$. $S-\chi^2$ is the raw-score based Pearson's chi-square statistic from Orlando and Thissen (2000). $S-RR-\chi^2$ and $S-DN-\chi^2$ are Rao-Robson (RR) and Dzhaparidze-Nikulin (DN) adjusted versions for $S-\chi^2$, which will be discussed in detail momentarily.

Table 3.4: Item-level absolute fit indexes for CDMs on the ECPE dataset

	RMSEA	$S-\chi^2$	$S-RR-\chi^2$	$S-DN-\chi^2$
Item 2	.012	46.723 (.000)	46.727 (.000)	39.465 (.002)
Item 10	.032	54.763 (.000)	54.791 (.000)	29.236 (.032)
Item 15	.026	49.838 (.000)	49.854 (.000)	33.857 (.009)
Item 19	.033	51.656 (.000)	51.689 (.000)	28.647 (.038)
Item 22	.042	61.712 (.000)	61.754 (.000)	27.957 (.045)
Item 23	.016	59.212 (.000)	59.225 (.000)	38.331 (.002)
Item 24	.029	75.482 (.000)	75.521 (.000)	45.462 (.000)

Chernoff and Lehmann (1954) have shown that a χ^2 statistic computed from the cells of probabilities (e.g., e_{ks} in $S-\chi^2$) based on grouped individual observations, while its estimates (e.g., item parameters $\hat{\boldsymbol{\delta}}_k$) are from ungrouped observations, does not have the expected limiting distribution.

To address the issue, Rao and Robson (1974) modified the squared-residual based statistics, in the item-level case $\boldsymbol{v}_k = (v_{k,1}, v_{k,2}, ..., v_{k,K-1})^T$, as

$$RR-\chi^2 = \boldsymbol{v}_k^T \left(\boldsymbol{I}_{K-1} - \boldsymbol{BJ^{-1}B^T}\right)^{-1} \boldsymbol{v}_k,$$

where

$$v_{k,s}(\hat{\boldsymbol{\delta}}_k) = \frac{\sqrt{N_k}(o_{ks} - e_{ks}(\hat{\boldsymbol{\delta}}_k))}{\sqrt{e_{ks}(\hat{\boldsymbol{\delta}}_k)(1-e_{ks}(\hat{\boldsymbol{\delta}}_k))}};$$

$\boldsymbol{J}$ is the information matrix w.r.t the k^{th} item parameters $\hat{\boldsymbol{\delta}}_k$ and $\boldsymbol{B}$ is the Jacobian matrix of $\boldsymbol{e}_k = (e_{k,1}, ..., e_{k,K-1})^T$ w.r.t $\hat{\boldsymbol{\delta}}_k$. The statistic is essentially

$$\boldsymbol{v}_k^T Cov\left(\boldsymbol{v}_k\right)^{-1} \boldsymbol{v}_k$$

which follows χ^2_{K-1} instead of $\chi^2_{K-1-p_k}$. Dzaparidze and Nikulin (1975) proposed a similar statistic

$$DN - \chi^2 = \boldsymbol{v}_k^T \left(\boldsymbol{I}_{K-1} - \boldsymbol{B}(\boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{B})^{-\mathbf{1}}\boldsymbol{B}^{\boldsymbol{T}}\right)^{-1} \boldsymbol{v}_k,$$

which follows $\chi^2_{K-1-p_k}$. The connection between the statistics has been discussed by McCulloch (1985). Simply put, the idea is to approximate the actual covariance matrix for residual $\boldsymbol{v}_k$ based on the MLEs calculated from the ungrouped data.

Table 3.4 presents the significant items across the three statistics under the saturated G-DINA model. The dfs are $20 - 1 - 2 = 17$, $20 - 1 = 19$, and $20 - 1 - 2 = 17$ for each column respectively. Notice that $K = 20$ since the cells are merge if the observed counts of the cell is less than 5. For the item-level fit detection, parameters for the other items and the size of latent classes are assumed to be invariant; plus, all flagged items are DINA items. Therefore, $p_k = 2$. The results suggest that a more flexible parametric form or a more sophisticated Q-matrix should be considered.

3.6 Discussion

While this chapter attempts to review some of the most commonly used measures and approaches for evaluating the model-data fit of CDMs, it is by no means complete. New methods are appearing quite regularly. For example, Chalmers and Ng (2017) modified the squared-residual based statistics by using plausible value imputation to generate and account for the uncertainty because of the use of latent trait estimates. The idea is rather similar to the resampling-based and the PPMC approaches.

Residual-based display techniques, used to assess the item-level absolute fit in the Bayesian approach, are not covered in this chapter. The graphical model diagnosis implemented for the Bayesian networks (Sinharay, 2006b) can be borrowed to examine item absolute fit for CDMs, providing a potential topic for future research.

The current methods that mainly focus on dichotomous responses can be generalized to ploytomous or mixed-form responses, which is certainly a promising topic for future research, as is the evaluation of those methods. Simulation-based and empirical studies on the performance of the comparable methods are needed to provide practitioners with useful guidance on how to choose amongst the methods.

It is also necessary to consider and assess the practical significance of model-data fit assessment and the consequence of model misfit, as no model is perfect. This issue has been stressed in the context of IRT framework by Hambleton and Han (2005) and Sinharay and Haberman (2014). Whereas Sinharay and Haberman (2014) discussed the significance of assessing item fit for the high-stack tests, van Rijn, Sinharay, Haberman, and Johnson (2016) investigated it for the low-stack assessments. The findings in the two studies reveal that model misfit hardly impacts test outcomes. To the best of our knowledge, such topics have not been studied thoroughly for CDM model-data fit methods, suggesting a promising direction for future research.

Chapter 4

The Standardized S-X^2 for Item Fit Analysis

4.1 Introduction

Standard 4.10 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014) recommends that evidence of model fit should be documented when an item response theory (IRT) model is used to make inferences from test data. Analysis of fit of IRT models in operational testing typically consists of the examination of item fit using residual plots and χ^2-type statistics (Hambleton & Han, 2005). Among the χ^2-type statistics that are used to assess item fit, S-X^2 (Orlando & Thissen, 2000) is one of the most popular in the IRT literature, presumably because: (i) the construction of S-X^2 is based on the grouping of examinees with respect to their observed total scores rather than their unobserved ability estimates; (ii) S-X^2 has been found to perform respectably in terms of Type I error rate and power in recent comparison studies (C. A. Glas & Falcón, 2003; Sinharay, 2006a; Sinharay & Lu, 2008; Stone & Zhang, 2003); (iii) the simple and intuitive nature of S-X^2 enables it to be easily generalized to cases beyond dichotomous responses and beyond the unidimensional latent trait (Kang & Chen, 2008, 2010; Roberts, 2008; Zhang & Stone, 2007).

Notwithstanding these appealing features, S-X^2 should not be used without considering its limitations. As noted by researchers such as Sinharay (2006a), the S-X^2 statistic, grounded on the Pearson's χ^2 statistic (Pearson, 1992), would not have a chi-square asymptotic distribution in typical IRT applications if the maximum likelihood estimates (MLEs) of item parameters are used to compute the statistic. Instead, the values of S-X^2 on average are slightly larger than the theorized χ^2 distribution. As an outcome, C. A. Glas and Falcón (2003); Sinharay (2006a); Sinharay and Lu (2008) found the Type I error rate of S-X^2 to be slightly larger than the nominal level. The goal of this paper is to suggest a modified S-X^2 statistic that has a known chi-square distribution asymptotically.

The study starts with an introduction to the Pearson's χ^2 and the S-X^2, which is followed by a discussion on the issue of using MLE-based Pearson's χ^2 statistic (Chernoff & Lehmann, 1954). The Background section ends with a brief discussion of a solution to the Chernoff-Lehmann issue suggested by Rao and Robson (1974). Subsequently, the Method section gives a review of the solution of Rao and Robson (1974) in detail, followed by the derivation of the modification of the S-X^2 so that the modified statistic has a known chi-square large-sample distribution. The Simulation section provides a comparison for S-X^2 and its origin in terms of the type-I error and the power. The Real Data section includes the applications of the two statistics to several real-world datasets. Conclusions are drawn and recommendations are provided in the last section.

4.2 Background: Pearson's χ^2 and S-X^2, the Chernoff-Lehmann Problem and a Solution

4.2.1 Pearson's χ^2

In statistical inference, it is typically assumed that a sample of observations with size N, belonging to a certain population, follow a probability distribution characterized by the parameter vector $\boldsymbol{\eta} = (\eta_1, ..., \eta_L)$, where L is the dimension of the vector. The construction of the Pearson's χ^2 begins with separating the sample into K groups (sometimes referred to as cells in statistical literature), from which the proportion (p_k) of observations in each group to the total is obtained. Next, the expected proportion (π_k) based on the assumed null distribution is calculated. As a result, the Pearson's χ^2 statistic (Pearson, 1992) is written as

$$\text{P-}X^2 = \sum_{k=1}^{K} \frac{n(p_k - \pi_k)^2}{\pi_k} = \boldsymbol{u}^\top \boldsymbol{u}, \tag{4.1}$$

where $\boldsymbol{u} = (\frac{\sqrt{n}(p_1-\pi_1)}{\sqrt{\pi_1}}, \frac{\sqrt{n}(p_2-\pi_2)}{\sqrt{\pi_2}}, \ldots, \frac{\sqrt{n}(p_K-\pi_K)}{\sqrt{\pi_K}})^\top$; π_k is short for $\pi_k(\boldsymbol{\eta})$. Intuitively, the above statistic is the sum of squared standardized residuals. Typically, P-X^2 that is computed using estimated parameters is claimed to follow a χ^2 distribution with the degree of freedom (df) of $K-1$ when the sample size is large Fisher (1924). The one df is reduced from the K to account for the constrain of $\sum_{k=1}^{K} \pi_k = 1$.

4.2.2 Orlando and Thissen's S-X^2

Orlando and Thissen (2000) proposed S-X^2 by adopting the idea behind P-X^2 to assess the item fit of IRT models for dichotomous responses. For item j, S-X^2 is defined as

$$\text{S-}X^2 = \sum_{k=1}^{K} \frac{n_k(o_k - e_k)^2}{e_k(1-e_k)} = \boldsymbol{v}^\top \boldsymbol{v}, \tag{4.2}$$

where

$$\boldsymbol{v} = \left(\frac{\sqrt{n_1}(o_1 - e_1)}{\sqrt{e_1(1-e_1)}}, \frac{\sqrt{n_2}(o_2 - e_2)}{\sqrt{e_2(1-e_2)}}, \ldots, \frac{\sqrt{n_K}(o_K - e_K)}{\sqrt{e_K(1-e_K)}} \right)$$

In the above expressions, K is the number of groups and n_k is the number of test-takers in the k^{th} group ; o_k and e_k are the observed and the expected proportion of test-takers in the k^{th} group who answer item j correctly.

In the setting of test-takers being grouped using their total (raw) scores, $K = J - 1$ because $k = 0$ and $k = J$ are trivial cases in which e_k equals to 0 and 1 for certain. Merging those groups having few test-takers would reduce the effect of sparseness. As suggested by Orlando and Thissen (2000), groups having less than 5 test-takers are merged in the present study. For notational convenience, merging is not applied to the introduction and derivation that follow.

In (4.2) $e_k \equiv e_k(\boldsymbol{\eta})$, specifically,

$$e_k(\boldsymbol{\eta}) = \frac{\int P(Y_j = 1|\theta, \boldsymbol{\eta}_j) S\left(T_{(-j)} = k - 1|\theta, \boldsymbol{\eta}_{(-j)}\right) \psi(\theta)\ d\theta}{\int S\left(T = k|\theta, \boldsymbol{\eta}\right) \psi(\theta)\ d\theta}. \tag{4.3}$$

Y_j denotes the response to item j and $P(Y_j = 1|\theta)$ represents the probability of answering item j correctly given the ability θ and the item parameter $\boldsymbol{\eta}_j$. T signifies the total score and $T_{(-j)}$ represents the total score from which the score of item j is excluded. $\boldsymbol{\eta} = \{\boldsymbol{\eta}_1, .., \boldsymbol{\eta}_J\}$, denoting a set of vectors including item parameters for a test; $\boldsymbol{\eta}_{(-j)} =$

$\{\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_{j-1}, \boldsymbol{\eta}_{j+1}, ..., \boldsymbol{\eta}_J\}$ the set of vectors of item parameters except for those of item j. The probability $P(Y_j = 1|\theta)$ is determined by the IRT model being used; for example, if the two-parameter logistic (2PL) model is selected, $P(Y_j = 1|\theta) = \frac{\exp a_j(\theta - b_j)}{1+expa_j(\theta - b_j)}$. $S\left(T = k|\theta, \boldsymbol{\eta}\right)$ denotes the probability of a test-taker answering k items correctly given her person parameter at θ. For instance, if $J = 2$,

$$S\left(T = 1|\theta, \boldsymbol{\eta}\right) = P\left(Y_1 = 1|\theta, \boldsymbol{\eta}_1\right) Q\left(Y_2 = 1|\theta, \boldsymbol{\eta}_2\right) + Q\left(Y_1 = 1|\theta, \boldsymbol{\eta}_1\right) P\left(Y_2 = 1|\theta, \boldsymbol{\eta}_2\right),$$

where $Q\left(Y_j = 1|\theta, \boldsymbol{\eta}_j\right) = 1 - P\left(Y_j = 1|\theta, \boldsymbol{\eta}_j\right)$. Similarly, $S\left(T_{(-j)} = k - 1|\theta, \boldsymbol{\eta}_{(-j)}\right)$ represents the probability of a test-taker answering $k - 1$ of the items, excluding the item j, correctly given the person parameter at θ. $\psi(\theta)$ is the probability density function (*pdf*) of θ.

When sample size is large, S-X^2 is claimed to follow a chi-square distribution with the *df* of $K - L_j$, namely, $J - L_j - 1$, where the L_j is the number of item parameters for item j (Orlando & Thissen, 2000). There are a few noteworthy differences between S-X^2 and P-X^2: (i) e_k of the former is a conditional proportion, while π_k of the latter is a marginal proportion, indicating that the constraint $\sum_k \pi_k = 1$ is not applicable to the e_k's; (ii) S-X^2 has n_k in its numerator and $e_k(1 - e_k)$ in its denominator. As discussed below, these nuanced differences are not trivial in the derivation of the modified S-X^2.

4.2.3 The Chernoff-Lehmann problem with Pearson's χ^2

Chernoff and Lehmann (1954) pointed out, in order to have a known (χ^2) distribution for cases with unknown $\boldsymbol{\eta}$, P-X^2 should be calculated using $\pi_k(\tilde{\boldsymbol{\eta}})$ that is estimated by

$$\underset{\boldsymbol{\eta}}{\operatorname{argmax}} \; n \sum_k p_k \log \pi_k(\boldsymbol{\eta}).$$

The quantity $n \sum_k p_k \log \pi_k(\boldsymbol{\eta})$ is the log-likelihood of observations belonging to the groups, forming the basis of the computation of P-X^2. They described this approach as obtaining estimates from grouped data. This type of parameter estimation is also referred to as the minimum χ^2 estimation in the statistical literature (Harris & Kanji, 1983) since the above maximization (of the likelihood) is equivalent to the minimization of P-X^2 with respect to $\boldsymbol{\eta}$. In this grouped-data approach, $\pi_k(\tilde{\boldsymbol{\eta}})$ is short for $\tilde{\pi}_k$ and $P - X^2$ computed with estimated parameters is written as $\tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{u}}$.

However, MLE is more often employed in practice due to its computational simplicity. The MLE $\hat{\boldsymbol{\eta}}$ is obtained by maximizing the log-likelihood function constructed on the ungrouped data, namely,

$$\underset{\boldsymbol{\eta}}{\operatorname{argmax}} \sum_{i}^{n} \log f(\boldsymbol{y_i}; \boldsymbol{\eta})$$

where $f(\boldsymbol{y_i}|\boldsymbol{\eta})$ is the likelihood of a realization $\boldsymbol{y_i}$ sampled from the random variable $\boldsymbol{Y}$. We denote $\pi_k(\hat{\boldsymbol{\eta}})$ as $\hat{\pi}_k$ and P-$X^2 = \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}$ in this ungrouped-data (MLE-based) approach.

Chernoff and Lehmann (1954) showed

$$\tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{u}} \ \sim \ \chi^2_{K-L-1}, \tag{4.4}$$

yet

$$\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}} \ \sim \ \chi^2_{K-L-1} + \sum_{l=1}^{L} \lambda_l(\boldsymbol{\eta}) \chi^2_1, \tag{4.5}$$

where $0 < \lambda_l < 1$. As a result, the null hypothesis will be rejected more often than is appropriate (and the Type I error rate of P-X^2 will be larger than the nominal level) if the limiting distribution of P-X^2 ($\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}$) is approximated by the χ^2_{K-L-1} distribution.

As mentioned above, a similar issue of an inflated Type I error rate has been found to occur with S-X^2 (C. A. Glas & Falcón, 2003; Sinharay, 2006a; Sinharay & Lu, 2008). To address the issue, Sinharay (2006a) suggested implementing S-X^2 as a discrepancy measure of posterior predicative model checking (PPMC) to conduct an item fit analysis in the Bayesian approach. The resampling-based approach of Stone (2000) and Stone and Zhang (2003) offers another solution. Although both approaches successfully avoid making decisions based on an inaccurate asymptotic distribution, their applications in practice require intensive computation.

4.2.4 Modified statistics

A class of approaches for addressing the Chernoff-Lehmann problem involve adjusting the P-X^2 statistic so as to have a known (χ^2) asymptotic distribution. Rao and Robson (1974) suggested modifying the P-X^2 statistic as

$$\text{P-}X^2_{RR} = \hat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}}^{-1} \hat{\boldsymbol{u}}, \tag{4.6}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}}$ is the covariance matrix of $\hat{\boldsymbol{u}}$. Essentially, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}}^{-1/2} \hat{\boldsymbol{u}}$ is standardized so that it follows $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ where $\boldsymbol{I}_K$ is an K-dimensional identity matrix. As a result,

$$\text{P-}X^2_{RR} \sim \chi^2_{K-1}.$$

in asymptotic.

The article applies such an adjustment to S-X^2 and derives a modified statistic

$$\text{S-}X^2_{RR} = \hat{\boldsymbol{v}}^\top \boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}}^{-1} \hat{\boldsymbol{v}} \sim \chi^2_K, \tag{4.7}$$

where $K = J - 1$ as mentioned above. The key aspect of the derivation is the computation of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}}$. In the section that follows, we begin with a review of the computation of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}}$ introduced by Rao and Robson (1974) and then proceed to derive $\boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}}$.

4.3 Method

4.3.1 $\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}}$ for P-X^2

Recall that $\hat{\boldsymbol{u}}$, the vector of standardized residuals for P$-X^2$, is calculated with the MLE $\hat{\boldsymbol{\eta}}$. Let us expand $\hat{\boldsymbol{u}}$ around the true unknown parameter vector $\boldsymbol{\eta_0}$ by a first-order Taylor series, that is,

$$\hat{\boldsymbol{u}} \approx \boldsymbol{u_0} - \boldsymbol{B_u}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}), \tag{4.8}$$

where

$$\boldsymbol{B_u} = -\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\pi_0}} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta_0}} = \left\{ \sqrt{n} \left(\frac{1}{\dot{\pi}_k^{1/2}} + \frac{p_k - \dot{\pi}_k}{\dot{\pi}_k^{3/2}} \right) \frac{\partial \pi_k}{\partial \dot{\eta}_l} \right\}_{K \times L} \approx \left\{ \sqrt{\frac{n}{\dot{\pi}_k}} \frac{\partial \pi_k}{\partial \dot{\eta}_l} \right\}_{K \times L}. \tag{4.9}$$

In the above expressions, the symbols $\dot{\eta}_l$ and $\dot{\pi}_k$ are employed to represent the elements of $\boldsymbol{\eta_0}$ and $\boldsymbol{\pi_0}$. $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\pi_0}}$ and $\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta_0}}$ are short for $\frac{\partial \boldsymbol{u}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}|_{\boldsymbol{\pi}=\boldsymbol{\pi_0}}$ and $\frac{\partial \boldsymbol{\pi}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}|_{\boldsymbol{\eta}=\boldsymbol{\eta_0}}$ respectively. This style of notations are applied throughout the following sections for notational convenience. The approximation in (4.9) holds true because p_k converges to $\dot{\pi}_k$ in probability as the increase of sample size.

The approximation (4.8) suggests

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}} \approx \boldsymbol{\Sigma}_{\boldsymbol{u_0}} - 2\,\mathrm{Cov}\left[\boldsymbol{u_0}, \boldsymbol{B_u}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0})\right] + \boldsymbol{B_u}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{B_u}^{\top}.$$

The covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}$ is computed by $\frac{\boldsymbol{J}^{-1}}{n}$ where $\boldsymbol{J}$ is the Fischer information calculated from the log-likelihood based on ungrouped data. $\boldsymbol{\Sigma}_{\boldsymbol{u_0}} = \boldsymbol{I_K} - \boldsymbol{q}\boldsymbol{q}^{\top}$, where $\boldsymbol{q} = (\sqrt{\dot{\pi}_1}, ..., \sqrt{\dot{\pi}_K})^{\top}$. The key object of this approximation is to find out the computation of $\mathrm{Cov}\left[\boldsymbol{u_0}, \boldsymbol{B_u}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0})\right]$.

The log-likelihood function of grouped data, $\tilde{\ell}(\boldsymbol{\eta}) = n\sum_k p_k \log \pi_k(\boldsymbol{\eta})$, is maximized by the minimum χ^2 estimator $\tilde{\boldsymbol{\eta}}$. The estimator can be obtained by solving

$$\frac{\partial \tilde{\ell}}{\partial \eta_l}\Big|_{\boldsymbol{\eta}=\tilde{\boldsymbol{\eta}}} = n\sum_k \frac{p_k}{\tilde{\pi}_k}\frac{\partial \pi_k}{\partial \tilde{\eta}_l} = 0, \text{ for } l = 1, ..., L. \tag{4.10}$$

Both $\tilde{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\eta}}$ are consistent estimators of the true unknown $\boldsymbol{\eta_0}$. That being said, $\tilde{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\eta}}$ become close enough to approximate $\boldsymbol{\eta_0}$ in a large sample size. Put differently, these estimators are in the vicinity of $\boldsymbol{\eta_0}$ (i.e., $\tilde{\boldsymbol{\eta}}, \hat{\boldsymbol{\eta}} \in \mathcal{N}_{\boldsymbol{\eta_0}}$). This statement also suggests that $\tilde{\pi}_k$ in (4.10) can be approximated by $\hat{\pi}_k$. Let us subtract

$$n\sum_k \frac{\dot{\pi}_k}{\hat{\pi}_k}\frac{\partial \pi_k}{\partial \hat{\eta}_l}$$

from the derivative (4.10), which leads to

$$\sum_k \frac{n(p_k - \dot{\pi}_k)}{\hat{\pi}_k}\frac{\partial \pi_k}{\partial \hat{\eta}_l} = 0 - \sum_k \frac{n\dot{\pi}_k}{\hat{\pi}_k}\frac{\partial \pi_k}{\partial \hat{\eta}_l} = \sum_k \frac{n(\hat{\pi}_k - \dot{\pi}_k)}{\hat{\pi}_k}\frac{\partial \pi_k}{\partial \hat{\eta}_l}, \text{ for } l = 1, ..., L. \tag{4.11}$$

Note that the second equity of the above equation remains valid due to the constraint $\sum_k \hat{\pi}_k = 1$. Using the neighborhood "trick" again, we can rewrite the equation (4.11) as

$$\sum_k \frac{n(p_k - \dot{\pi}_k)}{\dot{\pi}_k}\frac{\partial \pi_k}{\partial \dot{\eta}_l} = \sum_k \frac{n(\hat{\pi}_k - \dot{\pi}_k)}{\dot{\pi}_k}\frac{\partial \pi_k}{\partial \dot{\eta}_l},$$

or in matrices,

$$\boldsymbol{B}_{\boldsymbol{u}}^{\top}\boldsymbol{u_0} = \boldsymbol{B}_{\boldsymbol{u}}^{\top}\boldsymbol{D}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi_0}), \tag{4.12}$$

where $\boldsymbol{D}$ is a diagonal matrix with $\left\{\sqrt{\frac{n}{\dot{\pi}_k}}\right\}_{K\times 1}$ on its diagonal.

Let us expand $\hat{\boldsymbol{\pi}}$ around $\hat{\boldsymbol{\eta}}$ by a first-order Taylor series to have the following approximation:

$$\hat{\boldsymbol{\pi}} - \boldsymbol{\pi_0} \approx \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta_0}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}).$$

Multiplying $\boldsymbol{D}$ to both sides of the above approximation leads to

$$\boldsymbol{D}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi_0}) \approx \boldsymbol{D}\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta_0}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}) = \boldsymbol{B_u}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}). \tag{4.13}$$

(4.13) together with (4.12) indicates

$$\boldsymbol{B_u}^\top \boldsymbol{u_0} \approx \boldsymbol{B_u}^\top \boldsymbol{B_u}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}), \tag{4.14}$$

or equivalently,

$$\boldsymbol{u_0} \approx \boldsymbol{B_u}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}).$$

Given this expression, we can show

$$\mathrm{Cov}\left[\boldsymbol{u_0}, \boldsymbol{B_u}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0})\right] \approx \mathrm{Cov}\left[\boldsymbol{B_u}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}), \boldsymbol{B_u}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0})\right] = \boldsymbol{B_u}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{B_u}^\top.$$

Finally, we have

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}} \approx \boldsymbol{I}_K - \boldsymbol{q}\boldsymbol{q}^\top - \frac{\boldsymbol{B_u}\boldsymbol{J}^{-1}\boldsymbol{B_u}^\top}{n}. \tag{4.15}$$

4.3.2 $\boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}}$ for S-X^2

Utilizing the first-order Taylor expansion, we can approximate the residual vector $\hat{\boldsymbol{v}}$ for S-X^2 as

$$\hat{\boldsymbol{v}} \approx \boldsymbol{v_0} - \boldsymbol{B_v}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}),$$

where

$$\boldsymbol{B_v} \approx \left\{ \sqrt{\frac{n_k}{\dot{e}_k(1-\dot{e}_k)}}\frac{\partial e_k}{\partial \dot{\eta}_l} \right\}_{K \times L_j}$$

and

$$\boldsymbol{v_0} = \left\{ \frac{\sqrt{n_k}(o_k - \dot{e}_k)}{\sqrt{\dot{e}_k(1-\dot{e}_k)}} \right\}_{K \times 1}.$$

The above approximation of $\hat{\boldsymbol{v}}$ indicates

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}} \approx \boldsymbol{\Sigma}_{\boldsymbol{v_0}} - 2\,\mathrm{Cov}\left[\boldsymbol{v_0}, \boldsymbol{B_v}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0})\right] + \boldsymbol{B_v}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{B_v}^\top.$$

It is worthy noting that $\boldsymbol{\Sigma}_{\boldsymbol{v_0}} = \boldsymbol{I}_K$ because the $\dot{e}_k$ is conditional proportion and there is no correlation among the elements of $\boldsymbol{v_0}$.

Similar to that of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}}$, the most critical part in the approximation of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}}$ is the computation of $\mathrm{Cov}\left[\boldsymbol{v_0}, \boldsymbol{B_v}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0})\right]$. Before proceeding further with the computation of $\mathrm{Cov}\left[\boldsymbol{v_0}, \boldsymbol{B_v}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0})\right]$, let us consider an important difference in the derivation of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}}$. From the previous section, it can be noted that (4.11) makes an important contribution to the derivation of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}}$. (4.11) holds true because of the constraint $\sum_k \hat{\pi}_k = 1$ that, however, does not apply to the e_k of S-X^2. It is this nuanced difference that results in a different approach that follows to deriving the computation of $\mathrm{Cov}\left[\boldsymbol{v_0}, \boldsymbol{B_v}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0})\right]$.

Let us first focus on the minimum χ^2 estimator $\tilde{\boldsymbol{\eta}}$ maximizing the log-likelihood function of grouped data $\tilde{\ell}(\boldsymbol{\eta})$, where

$$\tilde{\ell}(\boldsymbol{\eta}) = \log \prod_k e_k(\boldsymbol{\eta})^{n_k o_k} \left(1 - e_k(\boldsymbol{\eta})\right)^{n_k(1-o_k)}.$$

Typically, $\tilde{\boldsymbol{\eta}}$ is obtained by solving

$$\frac{\partial \tilde{\ell}}{\partial \tilde{\boldsymbol{\eta}}} = \left\{ \sum_k \frac{n_k(o_k - \tilde{e}_k)}{\tilde{e}_k(1 - \tilde{e}_k)} \frac{\partial e_k}{\partial \tilde{\eta}_l} \right\}_{L_j \times 1} = \mathbf{0}_{L_j \times 1}.$$

Using the "trick" of $\tilde{\boldsymbol{\eta}} \in \mathcal{N}_{\boldsymbol{\eta_0}}$ to approximate the $\tilde{\eta}_l$ and $\tilde{e}_k$ with $\dot{\eta}_l$ and $\dot{e}_k$, we can rewrite the above derivative as

$$\boldsymbol{B_v}^\top \boldsymbol{v_0} \approx \frac{\partial \tilde{\ell}}{\partial \tilde{\boldsymbol{\eta}}} = \mathbf{0}_{L_j \times 1},$$

or equivalently,

$$-\boldsymbol{B_v}^\top \boldsymbol{v_0} \approx \frac{\partial \tilde{\ell}}{\partial \tilde{\boldsymbol{\eta}}} = \mathbf{0}_{L_j \times 1}. \tag{4.16}$$

Second, the MLE $\hat{\boldsymbol{\eta}}$ is obtained by solving

$$\frac{\partial \ell}{\partial \hat{\boldsymbol{\eta}}} = \mathbf{0}_{L_j \times 1},$$

where ℓ is the log-likelihood of ungrouped data. Let us expand this derivative by a first-order Taylor series around $\boldsymbol{\eta_0}$, i.e.,

$$\frac{\partial \ell}{\partial \boldsymbol{\eta_0}} + \boldsymbol{A}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}) \approx \frac{\partial \ell}{\partial \hat{\boldsymbol{\eta}}} = \mathbf{0}_{L_j \times 1}, \tag{4.17}$$

where $\boldsymbol{A} = \frac{\partial^2 \ell}{\partial \boldsymbol{\eta_0} \partial \boldsymbol{\eta_0}}$. In practice $\boldsymbol{A}$ is computed using the Fischer information $\boldsymbol{J}$, that is, $\boldsymbol{A} = -n\boldsymbol{J}$.

(4.16) and (4.17) indicate

$$-\boldsymbol{B}_{\boldsymbol{v}}^{\top} \boldsymbol{v_0} \approx \frac{\partial \ell}{\partial \boldsymbol{\eta_0}} + \boldsymbol{A}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0})$$

that plays a similar role as (4.14) does in the derivation of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{u}}}$. The above approximation can be rewritten as

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0} \approx -\boldsymbol{A}^{-1} \boldsymbol{B}_{\boldsymbol{v}}^{\top} \boldsymbol{v_0} - \boldsymbol{A}^{-1} \frac{\partial \ell}{\partial \boldsymbol{\eta_0}},$$

suggesting

$$\mathrm{Cov}\left[\boldsymbol{B_v}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}), \boldsymbol{v_0}\right] \approx \mathrm{Cov}\left(-\boldsymbol{B_v} \boldsymbol{A}^{-1} \boldsymbol{B}_{\boldsymbol{v}}^{\top} \boldsymbol{v_0} - \boldsymbol{B_v} \boldsymbol{A}^{-1} \frac{\partial \ell}{\partial \boldsymbol{\eta_0}}, \boldsymbol{v_0}\right).$$

Note that $\boldsymbol{B_v} \boldsymbol{A}^{-1} \frac{\partial \ell}{\partial \boldsymbol{\eta_0}}$ converges to a constant as the sample size increases, implying

$$\mathrm{Cov}\left(\boldsymbol{v_0}, \boldsymbol{B_v} \boldsymbol{A}^{-1} \frac{\partial \ell}{\partial \boldsymbol{\eta_0}}\right) = \mathbf{0}_{K \times K}.$$

Accordingly,

$$\mathrm{Cov}\left[\boldsymbol{B_v}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}), \boldsymbol{v_0}\right] \approx \mathrm{Cov}\left(-\boldsymbol{B_v} \boldsymbol{A}^{-1} \boldsymbol{B}_{\boldsymbol{v}}^{\top} \boldsymbol{v_0}, \boldsymbol{v_0}\right) = -\boldsymbol{B_v} \boldsymbol{A}^{-1} \boldsymbol{B}_{\boldsymbol{v}}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{v_0}}.$$

Given that $\boldsymbol{A}^{-1} = -\frac{\boldsymbol{J}^{-1}}{n} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v_0}} = \boldsymbol{I_K}$, on can derive

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}} &\approx \boldsymbol{\Sigma}_{\boldsymbol{v_0}} - 2\,\mathrm{Cov}\left[\boldsymbol{B_v}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta_0}), \boldsymbol{v_0}\right] + \boldsymbol{B_v}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{B_v}^{\top} \\ &\approx \boldsymbol{\Sigma}_{\boldsymbol{v_0}} - 2\boldsymbol{B_v}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{B_v}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{v_0}} + \boldsymbol{B_v}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}}\boldsymbol{B_v}^{\top} \\ &= \boldsymbol{I}_K - \frac{\boldsymbol{B_v}\boldsymbol{J}^{-1}\boldsymbol{B_v}^{\top}}{n}.\end{aligned}$$

4.4 Simulation Studies

This section discusses the results from two simulation studies. The first study examines the type-I error of S-X^2 and S-X^2_{RR} employed to the item fit of the 2PL model. The second study investigates the power of S-X^2 and S-X^2_{RR} across the Rasch, the 2PL and the 3PL models.

4.4.1 Outlines of the simulations

m denotes the number of simulated sets of responses. For each set, responses are simulated by means of the generating model M_g (IRT) with a predefined number of item J and a sample size N; parameters of M_g are randomly generated by the usual distributions, that is,

$$\begin{aligned}&a_j \sim U(1, 2),\\ &b_j \sim U(-3, 3),\\ &c_j \sim U(0.05, 0.3), \text{ for } j = 1, 2, ..., J,\end{aligned}$$

where the a, b and c are the discrimination, difficulty and guessing parameters of the 3PL model respectively and U stands for the uniform distribution. By doing so, a different set of item parameters is used for each replication. If the 2PL model is desired, $c_j \sim U(0, 0.3)$ can be restricted to $c_j = 0$ for $j = 1, 2, \ldots, J$; furthermore, $a_j \sim U(1, 2)$ can be limited to $a_j = 1$ if the Rasch model is desired.

Next, a calibrating model M_c is chosen to fit the simulated responses; the item fit statistics are computed accordingly. The type-I error or power is computed through counting the times that the statistic is larger than the critical value at the 5% nominal level. The type-I error is obtained if M_c is the same as M_g, whereas the power is examined if M_c and M_g are different. For example, in the second simulation study, each of the three IRT models is used to generate simulated data sets; then a calibrating model M_c being simpler than the generating model is employed to fit the simulated data.

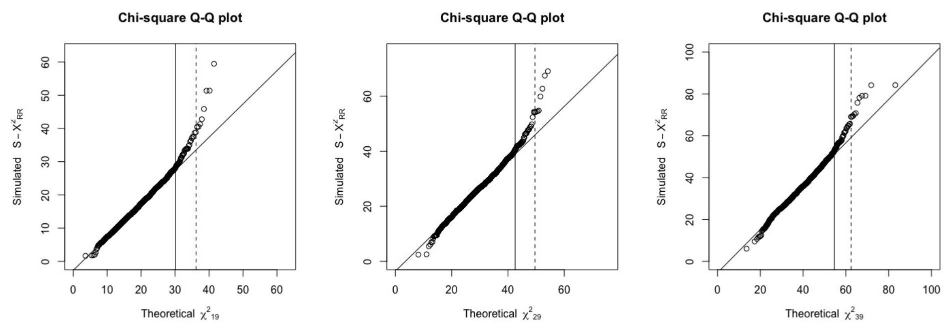
4.4.2 Results

Table 4.1: The type-I error of S-X^2 and S-X^2_{RR} for the 2PL model

J	N	S-X^2_{RR}	S-X^2
12	500	0.035	0.067
	800	0.039	0.064
20	1,000	0.049	0.073
	2,000	0.047	0.068
30	1,500	0.054	0.074
	3,000	0.050	0.073
40	2,000	0.060	0.072
	4,000	0.053	0.081
60	3,000	0.072	0.084
	6,000	0.068	0.093

In the first simulation study, $m = 3,000$ and responses are generated from the 2PL model according to a variety of combinations of sample size (N) and test length (J). Table 4.1 shows that the type-I error of S-X^2_{RR} is closer to the nominal 5% level than that of S-X^2 presented in the parentheses of the table. S-X^2 has a higher type-I error than is desired, which is in accordance with the results in existing simulation studies, as mentioned in Background section of this discussion. The modified statistic S-X^2_{RR} is slightly conservative—the type-I error of S-X^2_{RR} never exceeds that of S-X^2.

Figure 4.1: Chi-square quantile-quantile (QQ) plots of the empirical and the theoretical distributions of S-X^2_{RR}



(a) $J = 20$, $N = 2000$ (b) $J = 30$, $N = 3000$ (c) $J = 40$, $N = 4000$

The quantile-quantile (QQ) plots can be drawn using the empirical distributions of S-X^2_{RR} verse the theorized chi-square distributions. Figure 4.1 displays the QQ plots of S-X^2_{RR} under three simulation conditions: $J = 20$ and $N = 2,000$, $J = 30$ and $N = 3,000$, $J = 40$ and $N = 4,000$. The solid vertical line indicates the critical value of the chi-square distribution at the significant level of 0.05; the dashed line stands for he critical value at the level of 0.01. It can be noted that the 95 percent quantile of the empirical distribution is close to that of the theoretical distribution across the three conditions; the 95 percent quantile of the empirical distribution gradually approaches to the theoretical one as the number of items increases. It is suggested that using the 0.05 nominal level to conduct the test based on S-X^2_{RR} is relative sample compared to the 0.01 level when there is a limited number of items.

The second simulation study investigates the power of S-X^2 and S-X^2_{RR} for the Rasch, the 2PL, and the 3PL models. Responses are simulated from the generating model M_g; the calibrating model M_c having a more restricted form than the M_g is used to fit the simulated responses. Notice that the generating model M_g is only applied to one item in the test and the underlying models of the other items are assumed to be the same with the calibrating model M_c. Without loss of generality, the responses of Item 1 are generated using the generating model M_g and the item fit of Item 1 is assessed by the two statistics. For example, the responses of Item 1 are generated using the 2PL model (the generating model) and the responses of the other items are simulated using the Rasch model (the calibrating model); the Rasch model then is employed to calibrate all items and the item fit of Item 1 is assessed. In this study, multiple combinations of N and J are examined with 3,000 replications. Table 4.2 reveals that the modified statistic has conservative, but decent, power compared to its origin. Interestingly, the table shows that as J increases, the power becomes more substantive as the increase of J, namely, the number of groups (raw scores). This indicates, however, a common limitation of the χ^2-type item fit statistics.

Table 4.2: The power of S-X^2_{RR} (S-X^2) for the second simulation study

		M_g/M_c		
N	J	$2PL/1PL$	$3PL/1PL$	$3PL/2PL$
500	12	0.266(0.389)	0.834(0.767)	0.093(0.134)
1,000	20	0.690(0.795)	0.939(0.914)	0.224(0.274)
2,000	30	0.960(0.974)	0.981(0.970)	0.407(0.462)
4,000	60	0.997(0.999)	0.992(0.993)	0.607(0.670)

Tests of item fit based on the two statistics are only conducted for Item 1 in each simulation condition.

4.5 Real Data

This section applies S-X^2 and S-X^2_{RR} to analyze the item fit for a real-world dataset. The dataset was analyzed in Sinharay (2017), including dichotomous responses from 2,000 examinees, randomly selected from the full sample, to a state-administrated test with 46 item designed to measure students' achievement in mathematics.

Table 4.3: The number of items with significant values of S-X^2 and S-X^2_{RR} for the three IRT models for the real data set

Statistic	Rasch	2PL	3PL
S-X^2	33	18	6
S-X^2_{RR}	31	12	3

The Rasch, the 2PL, and the 3PL models are fit to the dataset and S-X^2_{RR} and S-X^2 are computed across all items for each of the models. A few raw-score based groups were merged, resulting in 42 groups based on raw scores. Table 4.3 reports the number of misfitting items that are identified by the item fit statistics at the 5% level of significance for the three IRT models. The table shows that for each IRT model, the use of S-X^2_{RR} leads to fewer misfitting items compared to that of S-X^2, with the difference being more prominent for the 2PL model. While both statistics are significant for a considerable number of items for the Rasch and 2PL model, they are significant for only 6 and 3 items, respectively, for the 3PL model. Although the 3PL model seems to adequately fit the Math dataset, more tests and further investigations, including tests for local independence (W.-H. Chen & Thissen, 1997), should be conducted to finalize this conclusion.

The three panels of Figure 4.2 show scatter plots of S-X^2 versus S-X^2_{RR} for the real dataset under the Rasch, the 2PL, and the 3Pl models. The range of the X-axis and Y-axis

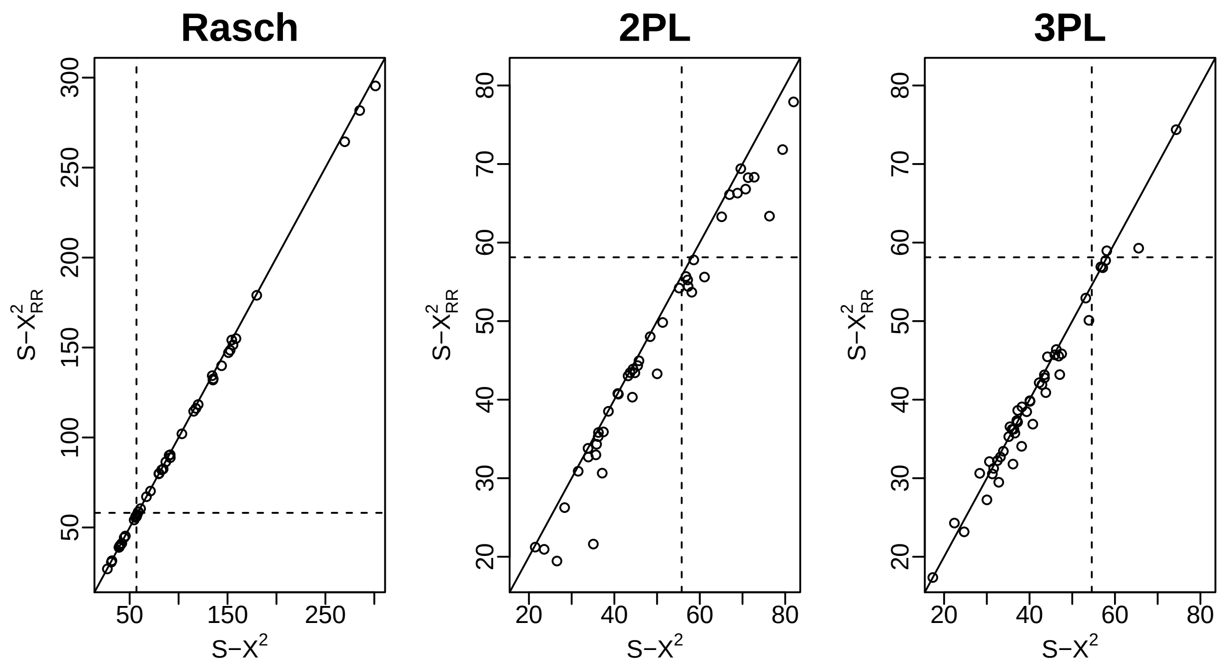


Figure 4.2: Plot of S-X^2 versus S-X^2_{RR} for the three IRT models for the real data.

are the same in the last two panels while the range is much wider for the Rasch model. The values of the two statistics are not shown for one item in the second panel; for this item, S-X^2 and S-X^2_{RR} are between 117 and 120 for the 2PL model. The panels include a diagonal line and also vertical and horizontal dashed lines indicating the critical values at 5% level of significance for the respective statistics.[1] The last two panels show that S-X^2_{RR} and also shows that for several items, S-X^2 is larger than its critical value, but S-X^2_{RR} is smaller than its critical value. Because misfitting items are often removed from the item pool (Sinharay & Haberman, 2014) and items are costly, these results indicate that the use of S-X^2_{RR}, rather than S-X^2 in operational testing, may lead to considerable saving of resources.

[1]For example, in each panel, a dashed vertical line is drawn at 58.12, which is the 95th percentile with a χ^2 distribution with 42 degrees of freedom, and is the critical value for S-X^2_{RR} for each IRT model.

4.6 Conclusions and Recommendations

The item fit statistic S-X^2, in spite of its simplicity and popularity, does not have a known chi-square limiting distribution (Sinharay, 2006a). The present study adopts the modification procedure suggested by Rao and Robson (1974) to provide a modified version of S-X^2 that has a known chi-square asymptotic distribution under the null hypothesis. The statistic S-X^2 can be written as $\hat{\boldsymbol{v}}^T\hat{\boldsymbol{v}}$. Essentially, the idea of the modification is to obtain a standardized quadratic form for $\hat{\boldsymbol{v}}$, that is, $\hat{\boldsymbol{v}}^T\Sigma_{\hat{\boldsymbol{v}}}^{-1}\hat{\boldsymbol{v}}$. One important contribution of the article is to derive the computation of the $\Sigma_{\hat{\boldsymbol{v}}}$. In sum, this paper suggests a χ^2-type statistic that (a) can be used to assess item fit for any IRT model for dichotomous items and (b) has a known asymptotic distribution under the null hypothesis.[2]

Simulation studies were conducted to show the performance of S-X^2 and S-X^2_{RR} in terms of the type-I error and power rate. Results obtained from the simulations suggest the type-I error of S-X^2_{RR} is closer to the nominal level than S-X^2 across different conditions. Meanwhile, S-X^2_{RR} was shown to have a slightly conservative power in comparison with S-X^2. Analysis of the two item fit statistics in real sets revealed that S-X^2_{RR} performs similarly as S-X^2 in terms of the number of misfitting items identified by the two statistics. In practice, S-X^2_{RR} should be used along with other methods such as informative graphics and pair-wised item fit indexes in order to gain an overarching understanding of the type of misfit.

Several limitations are noteworthy for this study, which could lead future directions. First, the study limits its scope to dichotomous response models. The statistic S-X^2_{RR} and the corresponding simulations can be extended to polytomous responses and mixed-form

[2]Item-fit statistics that have known asymptotic distribution under the null hypothesis have been suggested for the Rasch model by researchers such as C. A. W. Glas (1988).

test data. Second, the study only investigates three unidimensional IRT models assuming the latent variable follows a normal distribution. To obtain better understanding of the suggested statistic, one can look into its performance for cases with non-normal ability distributions, multidimensional latent variables, or discrete latent variables (latent classes). Third, grouping based on raw scores working well for IRT models might not be appropriated for latent class models such as cognitive diagnostic models. Future studies could develop new grouping schemes and compare them with the existing schemes. Last, the purpose of the study is by no means to replace the S-X^2 with S-X^2_{RR} but to offer an alternative to the existing methods of item fit analysis.

Chapter 5

Thoughts on Limitation and Future Research

Methods discussed in this manuscript virtually focus on the model-data fit of the latent variable models (mainly, IRT and CDMs) employed in educational tests, although several approaches for selecting models (model-model fit) have been reviewed in the second study. By studying this topic, I have also found some limitations of the use of model-data fit methods and potential future directions that are noteworthy.

No model is perfect except that some are easier to be disapproved than others, as found by Box and Draper (1987); a similar statement is as well noted in the context of psychometric models by Lord and Novick (1968). Assessment of model-data fit is necessary when a model is used to make inferences from the observed data. Subsequently, a question emerges: how wrong is the model so that it can be rendered as useless? The question has been asked by research on the use of model-data fit methods for IRT models (Hambleton & Han, 2005; I. W. Molenaar, 1997). Sinharay and Haberman (2014) suggested conducting analyses on the practical significance of misfit via evaluating the agreement between test outcomes (e.g., determination of cut-off score and selecting items in adaptive testing) from before and after using a model with better fit, excluding a few misfitting items and examines, or collapsing unpopular score categories of polytomous items. If a disagreement is observe, misfit is determined practically significant; otherwise, misfit is not significant. Note that in some contexts significance of misfit cannot be easily appraised, such as changing the phrase of some items and deleting items from the item pool implemented for adaptive testing, which makes the evaluation of significance a cost-consuming task.

Model-data fit methods themselves are not statistically perfect either. For example, the power of the chi-square type statistics, as the one suggested by the third study of the present manuscript, is sensitive to (positively correlated to) the number of groups (cells) that are predetermined to obtain the residuals between the observed values and the predictions by the model. Plus, with the increase of sample size, the statistics tend to be significant. Therefore, methods using graphical plots are suggested, for example, the residual analysis proposed by Sinharay and Haberman (2014) to assess item fit for unidimensional IRT models.

Thanks to increasingly use of computer-based tests, information beyond response patterns such as response time and action sequence has become accessible. However, the model-data fit methods purely based on response patterns would not be able to meet the needs of the updated models describing the new type of data. Discussions on the use of response time (D. Molenaar & de Boeck, 2018; van der Linden, 2007; Van Der Linden, 2009; van der Linden, Entink, & Fox, 2010; C. Wang et al., 2018) and action sequence (Bergner & von Davier, 2018) shed some lights on future directions of model-data fit. First, model-data fit methods could be developed for the approaches that directly model the new type of data beyond traditional responses. Second, methods could be adapted to integrate the information provided by the new type of data with the traditional measurement model, to assist in assessing the fit of the measurement model to responses.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactionson Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Belov, D. I. (2013). Detection of test collusion via Kullback-Leibler divergence. *Journal of Educational Measurement*, *50*(2), 141–163. doi: 10.1111/jedm.12008

Bergner, Y., & von Davier, A. A. (2018). Process Data in NAEP: Past, Present, and Future. *Journal of Educational and Behavioral Statistics*. Retrieved from `https://doi.org/10.3102/1076998618784700` doi: 10.3102/1076998618784700

Box, G. E. P., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. New York, NY: Wiley. doi: 10.2307/2982196

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8317.1978.tb00581.x` doi: 10.1111/j.2044-8317.1978.tb00581.x

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. Retrieved from `http://journals.sagepub.com/doi/10.1177/0049124192021002005` doi: 10.1177/0049124192021002005

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157. doi: 10.1177/026553229801500201

Casella, G., & Berger, R. (2001). *Statistical Inference*. Pacific Grove: Duxbury. doi: 10.1057/pt.2010.23

Chalmers, R. P., & Ng, V. (2017). Plausible-Value Imputation Statistics for Detecting Item Misfit. *Applied Psychological Measurement*, *41*(5), 372–387. Retrieved from `https://doi.org/10.1177/0146621617692079` doi: 10.1177/0146621617692079

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123–140. doi: 10.1111/j.1745-3984.2012.00185.x

Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289. doi: 10.2307/1165285

Chernoff, H., & Lehmann, E. L. (1954). The Use of Maximum Likelihood Estimates in $\chi 2$ Tests for Goodness of Fit. *The Annals of Mathematical Statistics*, *25*(3), 579–586. doi: 10.1214/aoms/1177728726

Chiu, C. Y. (2013). Statistical Refinement of the Q-Matrix in Cognitive Diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618. Retrieved from `https://doi.org/10.1177/0146621613488436` doi: 10.1177/0146621613488436

Cizek, G., & Wollack, J. A. (2017). *Handbook of Quantitative Methods for Detecting Cheating on Tests* (G. Cizek & J. A. Wollack, Eds.). New York, NY: Routledge.

Cochran, W. G. (1947). Some Consequences When the Assumptions for the Analysis of Variance are not Satisfied. *Biometrics*, *3*(1), 22–38. Retrieved from `http://www.jstor.org/stable/3001535` doi: 10.2307/3001535

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*(4), 429–449. Retrieved from `http://dx.doi.org/10.1111/j.1745-3984.2009.00091.x` doi: 10.1111/j.1745-3984.2009.00091.x

Dai, B., Ding, S., & Wahba, G. (2013). Multivariate Bernoulli distribution. *Bernoulli*, *19*(4), 1465–1483. doi: 10.3150/12-bejsp10

de la Torre, J. (2011). The Generalized DINA Model Framework. *Psychometrika*, *76*(2), 179–199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., & Chiu, C. Y. (2016, jun). A General Method of Empirical Q-matrix Validation. *Psychometrika*, *81*(2), 253–273. Retrieved from `https://doi.org/10.1007/s11336-015-9467-8` doi: 10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. a. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. doi: 10.1007/BF02295640

de la Torre, J., & Douglas, J. A. (2008, mar). Model Evaluation and Multiple Strategies in Cognitive Diagnosis: An Analysis of Fraction Subtraction Data. *Psychometrika*, *73*(4), 595. Retrieved from `https://doi.org/10.1007/s11336-008-9063-2` doi: 10.1007/s11336-008-9063-2 de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald Test for Item-Level Comparison of Saturated and Reduced Models in Cognitive Diagnosis. *Journal of Educational Measurement*, *50*(4), 355–373. Retrieved from http://dx.doi.org/10.1111/jedm.12022 doi: 10.1111/jedm.12022

de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of Clinical Data From a Cognitive Diagnosis Modeling Framework. *Measurement and Evaluation in Counseling and Development*, *51*(4), 281–296. Retrieved from https://doi.org/10.1177/0748175615569110 doi: 10.1080/07481756.2017.1327286

Drasgow, F., Levine, M. V., & Williams, E. a. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86.

Dzaparidze, K. O., & Nikulin, M. S. (1975). On a Modification of the Standard Statistics of Pearson. *Theory of Probability & Its Applications*, *19*(4), 851–853. Retrieved from https://doi.org/10.1137/1119098 doi: 10.1137/1119098

Embretson, S. E. (1991, sep). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495–515. Retrieved from https://doi.org/10.1007/BF02294487 doi: 10.1007/BF02294487

Fischer, G. H. (2003). The Precision of Gain Scores Under an Item Response Theory Perspective: A Comparison of Asymptotic and Exact Conditional Inference About Change. *Applied Psychological Measurement*, *27*(1), 3–26. Retrieved from https://doi.org/10.1177/0146621602239474 doi: 10.1177/0146621602239474

Fisher, R. A. (1924). Adelaide Research and Scholarship: The Conditions Under Which x2 Measures the Discrepancy Between Observation and Hypothesis. *Journal of the Royal Statistical Society*, *87*, 442–450. Retrieved from http://digital.library.adelaide.edu.au/dspace/handle/2440/15181{\%}5Cnpapers2://publication/uuid/F848BD41-3D1D-4FEE-AC32-90894CF0F1FD

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807. Retrieved from http://www3.stat.sinica.edu.tw/statistica/j6n4/j6n41/j6n41.htm doi: 10.1.1.142.9951

George, A. C., Ünlü, A., Kiefer, T., Robitzsch, A., & Groß, J. (2016). The R Package CDM for Cognitive Diagnosis Models . *Journal of Statistical Software*, *74*(2). Retrieved from http://www.jstatsoft.org/v74/i02/ doi: 10.18637/jss.v074.i02

Gilula, Z., & Haberman, S. J. (1994). Conditional Log-Linear Models for Analyzing Categorical Panel Data. *Journal of the American Statistical Association*, *89*(426), 645–656. Retrieved from http://www.jstor.org/stable/2290867

Glas, C. A., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106. doi: 10.1177/0146621602250530

Glas, C. A. W. (1988). The derivation of some tests for the {Rasch} model from the multinomial distribution. *Psychometrika*, *53*, 525–546. doi: 10.1007/bf02294405

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Washington, DC: Degnon Associates.

Hambleton, R. K., & Swaminathan, H. (1985). *ITEM RESPONSE THEORY Principles and Applications*. Netherlands: Springer. doi: 10.1007/978-94-017-1988-9

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 225–252. Retrieved from `http://dx.doi.org/10.1111/bmsp.12074` doi: 10.1111/bmsp.12074

Harris, R. R., & Kanji, G. K. (1983). On the Use of Minimum Chi-Square Estimation. *The Statistician*, *32*(4), 379. doi: 10.2307/2987540

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. doi: 10.1007/s11336-008-9089-5

Hogg, R., Mckean, J., & Craig, A. (2013). Maximum Likelihood Methods. In *Introduction to mathematical statistics* (7th ed., pp. 321–374). Pearson.

Hornby, L. (2011). *Gaming the GRE test in China, with a little online help.* Retrieved 2018-03-20, from `https://www.reuters.com/article/us-china-testing-cheating/gaming-the-gre-test-in-china-with-a-little-online-help-idUSTRE76Q19R20110727`

Jacob, B. A., & Levitt, S. D. (2003). Catching Cheating Teachers: The Results of an Unusual Experiment in Implementing Theory. *BrookingsWharton Papers on Urban Affairs*, *2003*(1), 185–220. doi: 10.1353/urb.2003.0010

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406. Retrieved from `http://doi.wiley.com/10.1111/j.1745-3984.2008.00071.x` doi: 10.1111/j.1745-3984.2008.00071.x

Kang, T., & Chen, T. T. (2010). Performance of the generalized SX2 item fit index for the graded response model. *Asia Pacific Education Review*, *12*(1), 89–96. doi: 10.1007/s12564-010-9082-4

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.1080/01621459.1995.10476572

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The Impact of Model Misspecification on Parameter Estimation and Item-Fit Assessment in Log-Linear Diagnostic Classification Models. *Journal of Educational Measurement*, *49*(1), 59–81. doi: 10.1111/j.1745-3984.2011.00160.x

Kyle, T. (2002). *Cheating scandal rocks GRE, ETS.* Retrieved 2018-03-20, from `http://www.thedartmouth.com/article/2002/08/cheating-scandal-rocks-gre-ets/`

Lei, P.-W., & Li, H. (2016). Performance of Fit Indices in Choosing Correct Cognitive Diagnostic Models and Q-Matrices. *Applied Psychological Measurement*, *40*(6), 405–417. Retrieved from `https://doi.org/10.1177/0146621616647954` doi: 10.1177/0146621616647954

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing Person Fit in Cognitive Diagnosis. *Applied Psychological Measurement*, *33*(8), 579–598. Retrieved from `https://doi.org/10.1177/0146621609331960` doi: 10.1177/0146621609331960

Liu, Y., Tian, W., & Xin, T. (2016). An Application of M2 Statistic to Evaluate the Fit of Cognitive Diagnostic Models. *Journal of Educational and Behavioral Statistics*, *41*(1), 3–26. Retrieved from `https://doi.org/10.3102/1076998615621293` doi: 10.3102/1076998615621293

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT True-Score and Equipercentile Observed-Score "Equatings". *Applied Psychological Measurement*, *8*(4), 453–461. doi: 10.1177/014662168400800409

Ma, W., & de la Torre, J. (2016). *GDINA: The generalized DINA model framework.* Retrieved from `http://cran.r-project.org/package=GDINA`

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model Similarity, Model Selection, and Attribute Classification. *Applied Psychological Measurement*, *40*(3), 200–217. Retrieved from `https://doi.org/10.1177/0146621615621717` doi: 10.1177/0146621615621717

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous. *Behavior Research Methods*, *42*(3), 847–862. doi: 10.3758/BRM.42.3.847

Magis, D., Raîche, G., & Béland, S. (2011, jul). A Didactic Presentation of Snijders's l z * Index of Person Fit With Emphasis on Response Model Selection and Ability Estimation. *Journal of Educational and Behavioral Statistics*, *37*(1), 57–81. doi: 10.3102/1076998610396894

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. Retrieved from `http://pubs.amstat.org/doi/abs/10.1198/016214504000002069` doi: doi:10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate Fit in Categorical Data Analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. doi: 10.1080/00273171.2014.911075

McCulloch, C. E. (1985). Relationships among some chi-square goodness of fit statistics. *Communications in Statistics - Theory and Methods*, *14*(3), 593–603. Retrieved from `http://dx.doi.org/10.1080/03610928508828936` doi: 10.1080/03610928508828936

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135.

Molenaar, D., & de Boeck, P. (2018, jun). Response Mixture Modeling: Accounting for Heterogeneity in Item Characteristics across Response Times. *Psychometrika*, *83*(2), 279–297. Retrieved from `https://doi.org/10.1007/s11336-017-9602-9` doi: 10.1007/s11336-017-9602-9

Molenaar, I. W. (1997). Lenient or strict application of IRT with an eye on practical consequences. In J. Rost & R. Langenheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 38–49). Munster, Germany: Waxmann.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333. Retrieved from `http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011{\_}20110927/04{\_}Oliveri.pdf`

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. doi: 10.1177/01466216000241003

Pearson, K. (1992). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Breakthroughs in Statistics*. doi: doi:10.1080/14786440009463897

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, *83*(2), 251–266. Retrieved from http://biomet.oxfordjournals.org/content/83/2/251.abstract doi: 10.1093/biomet/83.2.251

Ramsey, P. H. (1980). Exact Type 1 Error Rates for Robustness of Student's t Test with Unequal Variances. *Journal of Educational Statistics*, *5*(4), 337–349. Retrieved from http://journals.sagepub.com/doi/10.3102/10769986005004337 doi: 10.3102/10769986005004337

Rao, K. C., & Robson, D. S. (1974). A Chi-Square Statistic For Goodness-Of-Fit Tests Within The Expohential Family. *Communications in Statistics*, *3*(12), 1139–1153. doi: 10.1080/03610927408827216

Roberts, J. S. (2008). Modified likelihood-based item fit statistics for the generalized graded unfolding model. *Applied Psychological Measurement*, *32*(5), 407–423. doi: 10.1177/0146621607301278

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic Distribution of P Values in Composite Null Models. *Journal of the American Statistical Association*, *95*(452), 1143–1156. doi: 10.1080/01621459.2000.10474310

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2016). *CDM: Cognitive Diagnosis Modeling.* Retrieved from https://cran.r-project.org/package=CDM

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician. *The Annals of Statistics*, *12*(4), 1151–1172. doi: 10.1214/aos/1176346785

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. Retrieved from http://projecteuclid.org/euclid.aos/1176344136 doi: 10.1214/aos/1176344136

Sinharay, S. (2006a). Bayesian item fit analysis for unidimensional item response theory models. *The British journal of mathematical and statistical psychology*, *59*(2), 429–49. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17067420 doi: 10.1348/000711005X66888

Sinharay, S. (2006b). Model Diagnostics for Bayesian Networks. *Journal of Educational and Behavioral Statistics*. doi: 10.3102/10769986031001001

Sinharay, S. (2016a). Asymptotically Correct Standardization of Person-Fit Statistics Beyond Dichotomous Items. *Psychometrika*, *81*(4), 992–1013. doi: 10.1007/s11336-015-9465-x

Sinharay, S. (2016b). Detection of Item Preknowledge Using Likelihood Ratio Test and Score Test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46–68. doi: 10.3102/1076998616673872

Sinharay, S. (2017). How to Compare Parametric and Nonparametric Person-Fit Statistics Using Real Data. *Journal of Educational Measurement*, *54*(4), 420–439. doi: 10.1111/jedm.12155

Sinharay, S. (2018). Detecting Fraudulent Erasures at an Aggregate Level. *Journal of Educational and Behavioral Statistics*, *43*(3), 286–315. Retrieved from `https://doi.org/10.3102/1076998617739626` doi: 10.3102/1076998617739626

Sinharay, S., & Almond, R. G. (2007). Assessing Fit of Cognitive Diagnostic Models A Case Study. *Educational and Psychological Measurement*, *67*(2), 239–257. Retrieved from `https://doi.org/10.1177/0013164406292025` doi: 10.1177/0013164406292025

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*(1), 23–35. doi: 10.1111/emip.12024

Sinharay, S., & Lu, Y. (2008). A further look at the correlation between item parameters and item fit statistics. *Journal of Educational Measurement*, *45*(1), 1–15. doi: 10.1111/j.1745-3984.2007.00049.x

Skorupski, W., Fitzpatrick, J., & Egan, K. (2017). A Bayesian hierarchical linear modeling approach for detecting cheating and aberrance. In G. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 232–244). New York, NY: Routledge.

Smits, D. J., De Boeck, P., & Vansteelandt, K. (2004). The inhibition of verbally aggressive behaviour. *European Journal of Personality*, *18*(7), 537–555. doi: 10.1002/per.529

Snijders, T. a. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential Item-Fit Evaluation in Cognitive Diagnosis Modeling. *Applied Psychological Measurement*, *41*(8), 614–631. Retrieved from `https://doi.org/10.1177/0146621617707510` doi: 10.1177/0146621617707510

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series*

B: Statistical Methodology, *64*(4), 583–616. doi: 10.1111/1467-9868.00353

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*(1), 58–75. doi: 10.1111/j.1745-3984.2000.tb01076.x

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331–352.

Templin, J., & Bradshaw, L. (2014). Hierarchical Diagnostic Classification Models: A Family of Models for Estimating and Testing Attribute Hierarchies. *Psychometrika*, *79*(2), 317–339. doi: 10.1007/s11336-013-9362-0

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using mplus. *Educational Measurement: Issues and Practice*. doi: 10.1111/emip.12010

Teugels, J. L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *Journal of Multivariate Analysis*, *32*(2), 256–268. doi: 10.1016/0047-259X(90)90084-U

Van Der Linden, W. J. (2009). Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, *46*(3), 247–272. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3984.2009.00080.x` doi: 10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J. (2007, aug). A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, *72*(3), 287. Retrieved from `https://doi.org/10.1007/s11336-006-1478-z` doi: 10.1007/s11336-006-1478-z

van der Linden, W. J., Entink, R. H. K., & Fox, J.-P. (2010). IRT Parameter Estimation With Response Times as Collateral Information. *Applied Psychological Measurement*, *34*(5), 327–347. Retrieved from `https://doi.org/10.1177/0146621609349800` doi: 10.1177/0146621609349800

van der Linden, W. J., & Hambleton, R. K. (2013). Item Response Theory: Brief History, Common Models, and Extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1–28). New York, NY: Springer. doi: 10.1007/978-1-4757-2691-6_1

van der Linden, W. J., & Jeon, M. (2011). Modeling Answer Changes on Test Items. *Journal of Educational and Behavioral Statistics*, *37*(1), 180–199. doi: 10.3102/1076998610396899

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016, jul). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assessments in Education*, *4*(1), 10. Retrieved from https://doi.org/10.1186/s40536-016-0025-3 doi: 10.1186/s40536-016-0025-3

Vogell, H., & Perry, J. (2009). *Are drastic swings in CRCT scores valid?* Retrieved from https://www.ajc.com/news/local/are-drastic-swings-crct-scores-valid/1uNxbbiLUZjvYQx6gMkyyN/

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307. doi: 10.1348/000711007X193957

von Davier, M. (2014). The Log-Linear Cognitive Diagnostic Model (LCDM) as a Special Case of the General Diagnostic Model (GDM). *ETS Research Report Series*, *2014*(2), 1–13. doi: 10.1002/ets2.12043

von Davier, M., & Haberman, S. J. (2014). Hierarchical Diagnostic Classification Models Morphing into Unidimensional 'Diagnostic' Classification Models-A Commentary. *Psychometrika*, *79*(2), 340–346. doi: 10.1007/s11336-013-9363-z

Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing Item-Level Fit for the DINA Model. *Applied Psychological Measurement*, *39*(7), 525–538. Retrieved from https://doi.org/10.1177/0146621615583050 doi: 10.1177/0146621615583050

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting Aberrant Behavior and Item Preknowledge: A Comparison of Mixture Modeling Method and Residual Method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469–501. doi: 10.3102/1076998618767123

Wang, W.-C., & Chen, H.-C. (2004). The Standardized Mean Difference within the Framework of Item Response Theory. *Educational and Psychological Measurement*, *64*(2), 201–223. Retrieved from https://doi.org/10.1177/0013164403261049 doi: 10.1177/0013164403261049

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, *21*(4), 307–320. doi: 10.1177/01466216970214002

Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting Test Tampering Using Item Response Theory. *Educational and Psychological Measurement*, *75*(6), 931–953. doi: 10.1177/0013164414568716

Wollack, J. A., & Eckerly, C. A. (2017). Detecting test tampering at the group level. In G. Cizek & J. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 214–231). Washington, DC: Routledge.

Wright, B. D., & Stone, M. H. (1979). *Best test design, Rasch Measurement.* Chicago: The University of Chicago: Mesa Press.

Yen, W. M. (1981). Using Simulation Results to Choose a Latent Trait Model. *Applied Psychological Measurement*, *5*(2), 245–262. doi: 10.1177/014662168100500212

Zhang, B., & Stone, C. A. (2007). Evaluating Item Fit for Multidimensional Item Response Models. *Educational and Psychological Measurement*, *68*, 181–196. doi: 10.1177/0013164407301547