Evaluating Person Fit for Cognitive Diagnostic Assessment

Applied Psychological Measurement

2015, Vol. 39(3) 223–238

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621614557272

apm.sagepub.com

Ying Cui¹ and Johnson Li²

Abstract

Methods evaluating person fit for cognitive diagnostic assessment are an important area of research because failing to detect misfitting responses can lead to the misinterpretation of students' attribute profiles, which may result in faulty remediation decisions. This article aims to examine ways of detecting person misfit for cognitive diagnostic assessments. The authors first investigated whether the well-known I_z statistic, developed under the framework of item response theory, can be extended for use in the context of cognitive diagnostic models. The authors also introduce a new person fit statistic, response conformity index (RCI), developed for detecting misfitting response patterns for cognitive diagnostic assessments. The authors conduct both simulation and real data studies to compare the detection rates of I_z and our new statistic.

Keywords

person misfit, aberrant responding behavior, cognitive diagnostic model, misfitting response patterns, person fit statistics

Introduction

In educational and psychological measurement, many person fit statistics have been proposed to help identify observed response patterns that are incongruent with the measurement model used in test development and analysis (see Meijer & Sijtsma, 2001, for a review of existing person fit statistics; also see Rupp, 2013, for a review of the methodology for conducting simulation studies in the area of person fit statistics). One of the most studied person fit statistics in the literature is probably the I_z statistic developed by Drasgow, Levine, and Williams (1985), which is a standardized version of the I_0 index (Levine & Rubin, 1979). I_0 is the log likelihood of an observed item response pattern calculated based on an item response theory (IRT) model, which is given by

¹University of Alberta, Edmonton, Canada

²University of Manitoba, Winnipeg, Canada

Corresponding Author:

Ying Cui, Associate Professor, Department of Educational Psychology, University of Alberta, 6-110D Education North, Edmonton, Alberta, Canada T6G 2G5.

Email: yc@ualberta.ca

$$l_{0i} = \ln \left\{ \prod_{j=1}^J P_j(\theta_i)^{X_{ij}} [1 - P_j(\theta_i)]^{1-X_{ij}} \right\}, \quad (1)$$

where X_{ij} is the binary (0, 1) response to item j ($j = 1, 2, \dots, J$) by student i , θ_i is the latent trait for student i , and $P_j(\theta_i)$ is the probability of correctly answering item j given θ_i computed based on an IRT model. A low value of l_{0i} suggests that the probability of obtaining the response pattern produced by student i is small given the hypothesized IRT model, and hence, the response pattern can be considered as a misfit of the IRT model. As pointed out by Drasgow et al. (1985), an undesirable property of l_0 lies in the fact that l_0 is conditional on the latent trait, θ , which suggests that the classification of an observed response vector as fitting or misfitting is influenced by θ . Therefore, the standardized normally distributed statistic l_z is written as

$$l_z = \frac{l_0 - E(l_0)}{[\text{Var}(l_0)]^{1/2}}, \quad (2)$$

where

$$E(l_0) = \sum_{j=1}^J \{P_j(\theta) \ln[P_j(\theta)] + [1 - P_j(\theta)] \ln[1 - P_j(\theta)]\} \quad (3)$$

and

$$\text{Var}(l_0) = \sum_{j=1}^J P_j(\theta) [1 - P_j(\theta)] \left[\ln \frac{P_j(\theta)}{1 - P_j(\theta)} \right]^2. \quad (4)$$

Because large negative values of l_z suggests potential misfit, the cut score for detecting misfit can be found from the left tail of the standardized normal distribution with a chosen significance level (e.g., $\alpha = .5$). Molenaar and Hoijtink (1990) proposed a modified version of l_z by calculating the likelihood function using the observed sum score of items instead of the latent trait θ for the Rasch model. Furthermore, von Davier and Molenaar (2003) extended Molenaar and Hoijtink's index under the framework of latent class models. Thus far, there is no empirical evidence in the literature showing that the modified l_z is significantly better the original l_z . Therefore, this study focuses on the original l_z only.

Although l_z was initially developed under IRT, one may ask whether it could be extended as a person fit statistic to detect misfitting response patterns relative to a *cognitive diagnostic model* (CDM). Rupp and Templin (2008) and Rupp, Templin, and Henson (2010) provided a review of many currently existing CDMs. CDMs are relatively new measurement models aimed at classifying students into a set of partially ordered latent classes or states, often called attribute patterns, which are defined in term of the mastery or non-mastery of a set of attributes (e.g., knowledge and skills) being measured by the test. These results aim to provide students and teachers with more instructionally relevant information that supports classroom teaching and learning (cf. Leighton & Gierl, 2007).

Computationally, l_z can be straightforwardly adopted into the CDM framework by calculating the log likelihood of an observed response vector given a CDM (instead of an IRT model). That is, the item response function of a CDM is used to calculate the probability of an observed response, which serves as the basis of computing the likelihood of the response vector. More specifically, to compute l_z under the CDM framework, $P_j(\theta)$ in Equations 1 to 4 needs to be

calculated using the item response function of a CDM model (e.g., Equation 6 to be discussed later). However, two questions arise as follows:

1. Does the I_z based on a CDM still follow the standardized normal distribution?
2. How well does the I_z detect misfitting response patterns under the CDM framework?

Therefore, the first purpose of the present study is to conduct simulation studies designed to answer these two questions. The second purpose is to propose a new person fit statistic for detecting misfitting response patterns in CDMs. The authors conduct both simulation and real data studies to evaluate and compare the detection rates of I_z and our new statistic.

Person Fit for Cognitive Diagnostic Model

During the past few decades, researchers and practitioners have become increasingly interested in investigating ways of improving the diagnostic value of educational and psychological tests, so that test results yield more detailed information indicating the strengths and weaknesses of each student. The new diagnostic tests are designed to measure a set of attributes, with a so-called Q matrix (K. Tatsuoka, 2009) to specify the attributes measured by each item. Many CDMs have been proposed to estimate student's attribute patterns, denoted by $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK})'$, where $\alpha_{ik} = 1$ or 0 indicating whether or not attribute k is mastered or possessed by student i . These CDMs differ primarily in their model complexities as indicated by the number of item parameters and the ways that attribute mastery and random noise are conceptually formulized and statistically estimated. For a review of existing CDMs, the reader is referred to Rupp and Templin (2008), Rupp et al. (2010), and DiBello, Roussos, and Stout (2007).

The development of cognitive diagnostic tests based on CDMs holds great promise for improving the educational benefits of testing by making inferences about student strengths and weaknesses to inform instructional practices (e.g., Rupp et al., 2010). Thus far, tremendous efforts and progress have been made in this new area of research. For example, Gierl, Alves, and Majeau (2010) conducted a pilot study to examine the technical and practical feasibility of obtaining more instructional results from provincial achievement tests in Canada. A total of three diagnostic mathematics tests, which aimed to provide teachers with diagnostic information regarding student mathematical skills, were developed by seven content specialists. These tests measured a total number of 48 mathematical skills in four content areas: (a) Number, (b) Patterns and Relations, (c) Shape and Space, and (d) Statistics and Probability at Grade 3 and 6 levels. For each test, the number of items ranged from 16 to 24. The results from the cognitive diagnostic tests have proven to be valuable for diagnosing students' strengths and weaknesses in learning mathematics.

Person fit analysis is useful in cognitive diagnostic tests as it can help validate individual student diagnostic results. Invalid skill diagnosis can lead to the misinterpretation of students' skill profiles, which can lead to faulty remediation decisions. This may not only result in a waste of students' and teachers' time and effort but also adversely affect students' educational and future employment opportunities. Although diagnostic results are important for students and teachers, relatively few person fit statistics have been developed under the framework of CDMs. Cui and Leighton (2009) proposed a person fit statistics called the *hierarchical consistency index* (HCI). The HCI was developed under the framework of the Attribute Hierarchy Method (AHM; Leighton, Gierl, & Hunka, 2004) in which items are assumed to measure a set of hierarchically related attributes. It examines whether students' actual item response patterns match the expected response patterns based on the hierarchical relationship among attributes

measured by a test. Thus, the HCI can achieve the maximum power when the hierarchical structure of the attributes is well defined. When the attributes being measured by a test are not hierarchically related, the usefulness of the HCI is limited.

To examine person fit for CDMs, Liu, Douglas, and Henson (2009) developed a likelihood ratio test to identify students who have a tendency to respond aberrantly. The likelihood ratio test compares the likelihood of the response pattern due to a hypothesized misfitting behavior against the likelihood as a result of the normal behavior. To obtain the likelihood of the misfitting response pattern, the item response function in CDMs must be modified according to the hypothesized misfitting behavior. As pointed out by the authors, "various specific causes of misfitting can be proposed and tested as alternative hypotheses" (Liu et al., 2009, p. 19). The authors exemplified two types of misfit: spuriously high scorers and spuriously low scorers. A new estimation algorithm is needed to reliably calibrate the parameters of the modified item response function.

In this study, the authors proposed a new person fit index for CDMs. A unique feature of a CDM is its use of the Q matrix for specifying the knowledge and skills required to solve items on a cognitive diagnostic test. A challenge of developing/using such a matrix is that it may not accurately describe the knowledge and skills *every* student uses to solve items. This could potentially lead to person misfit on diagnostic tests. Therefore, it is important to examine the fit of student responses relative to expectations from the Q matrix. Our new index is designed to explicitly evaluate the consistency between student observed responses and the Q-matrix expectations. We call this new index the response conformity index (RCI), which is given by

$$RCI_i = \sum_{j=1}^J |RCI_{ij}| = \sum_{j=1}^J \left| \ln \left[-\frac{X_{ij} - P_j(\alpha_i)}{I_j(\alpha_i) - P_j(\alpha_i)} \right]^{X_{ij} + I_j(\alpha_i)} \right|, \quad (5)$$

where α_i is the attribute pattern for student i , $P_j(\alpha_i)$ is the probability of a correct response to item j given α_i , and $I_j(\alpha_i)$ is the ideal response to item j given α_i . $I_j(\alpha_i)$ is equal to 1 or 0 depending on whether or not all the attributes required by item j , as indicated in the Q matrix, are present in α_i . In practice, α_i , the true attribute pattern of student i is never available. As a result, the estimated attribute pattern, $\hat{\alpha}_i$ must be used instead. It should be noted that the derivation of the ideal responses in calculating RCI assumes that the mastery of all the required attributes is necessary for successful performance on a given test item. That is, the ideal response to an item will be 1 only if all the attributes required by the item are present in the attribute pattern; if one or more attributes are missing, the ideal response to the item will be 0. The derivation of the ideal response can be considered as the "endpoint" of the conceptual continuum under any CDM in terms of attribute masteries required for a correct response.

For each test item, RCI_j measures the degree to which the observed item response X_j deviates from the ideal response $I_j(\alpha_i)$, $j = 1, 2, \dots, J$. Given that both X_j and $I_j(\alpha_i)$ are binary variables (0 or 1), a total four possible combinations of X_j and $I_j(\alpha_i)$ exist. For item j , when $X_j = I_j(\alpha_i)$, RCI_j will have a value of 0. However, when $X_j \neq I_j(\alpha_i)$, the value of RCI_j depends on the degree of the discrepancy between $X_j - P_j(\alpha_i)$ and $I_j(\alpha_i) - P_j(\alpha_i)$. These two quantities concern how far the observed and ideal item responses deviate from the probability of a correct response estimated by the CDM. If the probability of a correct response is around .5, we expect the discrepancy between X_j and $I_j(\alpha_i)$ is due to chance only and therefore the ratio of $X_j - P_j(\alpha_i)$ to $I_j(\alpha_i) - P_j(\alpha_i)$ is close to -1 and RCI_j has a value close to zero. As the probability of a correct response gets closer to 1 or 0, the mismatch of X_j and $I_j(\alpha_i)$ becomes more unexpected and the ratio of $X_j - P_j(\alpha_i)$ to $I_j(\alpha_i) - P_j(\alpha_i)$ deviates further from -1 , which results in a larger value of RCI_j .

Specifically, when $X_j - P_j(\alpha_i)$ is relatively large in comparison with $I_j(\alpha_i) - P_j(\alpha_i)$, it indicates the student responds to the item unexpectedly and therefore aberrant response behavior, such as cheating or creative responding, may occur. In this case, RCI_j will have a large positive value. In contrast, when the absolute value of $I_j(\alpha_i) - P_j(\alpha_i)$ is comparatively large relative to $X_j - P_j(\alpha_i)$, one possible reason is the poor quality of the item or the use of alternative strategies in solving the item other than that specified in the Q matrix. And RCI_j will have a large negative value. Therefore, the inspection of the direction of RCI_j could help identify the reason of misfit at the individual item level. The RCI, by taking the sum of the absolute values of all RCI_j , $j = 1, 2, \dots, J$, measures the overall fit of a student's observed item responses relative to the expectations of the CDM used in the test design and analysis. It should be noted that we have also considered to calculate the RCI by using the sum of squared values of the RCI_j . In our simulation study, we compared the use of sum of square values and the sum of absolute values of the RCI_j in calculating the overall RCI across different simulation conditions. Results indicate that the use of the sum of absolute values is consistently associated with stronger power rates. Due to limited space, detailed results were not presented in the article but will be available upon request from the authors.

To set the critical values for RCI so as to classify an observed response pattern as normal or misfitting, the simulation approach commonly used in the literature of person fit for finding critical values with simulated data could be used (e.g., Seo & Weiss, 2013; van Krimpen-Stoop & Meijer, 2002). That is, data are first simulated based on estimated student and item parameters to approximate the empirical distributions of RCIs under the null hypothesis of fitting responses and therefore empirical cut points could be identified. For each attribute pattern, data consisting of a large number of *normal* student item responses (e.g., 2,000 or 10,000) are generated. The value of RCI is then calculated for each of the simulated item response patterns and ordered from the lowest to the highest value. In this way, the critical value of RCI could be identified at any desirable level of significance. For example, using $\alpha = .05$, the 95th percentile value of RCI can be chosen as the cut point because a higher value of the index indicates a higher degree of misfit. Observed item response patterns will be considered as misfitting when the observed value is larger than the chosen cut point. This process can be repeated across all attribute patterns.

Simulation Study I: The Null Distributions of I_z and the RCI

A simulation study was first conducted to examine the null distributions of I_z and the RCI under the CDM framework to (a) investigate the distributional properties of the two statistics and (b) find empirical critical values that were later used to detect misfits in our second simulation study in which we evaluated the Type I errors and power of the two statistics. Item responses to hypothetical diagnostic tests were generated to find the simulated distributions of I_z and the RCI. A total of 12 tests were created by manipulating three factors, including the number of attributes (3 and 6), test length (20 and 40), and the discrimination power of test items (high, low, and mixed). All three factors have been found to be key factors for the measurement accuracy of diagnostic tests (e.g., Cui, Gierl, & Chang, 2012) and for the detection of person misfit relative to CDMs (Cui & Leighton, 2009). The number of attributes and test length were chosen to reflect the common settings in recent simulation studies and real data applications of cognitive diagnosis reported in the literature. For example, de la Torre and Hong (2010) considered six attributes and 15 items in their simulation study. And Choi, Templin, Cohen, and Atwood's (2010) simulation study considered 40 items and four attributes and their real data application used 25 items and three attributes. As suggested by Meijer and Tendeiro (2012), model fit should be examined prior to person fit analysis to eliminate or reduce the possibility that the

misfit of a student response pattern is due to the general misfit of the measurement model. With adequate model and item fit, results of person fit are expected to be less confounding. As a result, we did not manipulate the overall model misfit in the simulation. Rather, we assumed the true measurement model was used to analyze student responses.

To generate the Q matrix for tests with 3 attributes, we assume each item, on average, measuring 1.5 attributes; for tests with 6 attributes, we assume each item, on average, measuring 3 attributes. The Q matrices for the four tests are provided in the appendix. Although there are many CDMs available in the literature, for the purpose of illustration, the compensatory–reparameterized unified model (C-RUM; Hartz, 2002) was used as the major framework for the simulation of person and item parameters. The C-RUM model allows the item parameters to be tailored to each item, and the attribute parameters to be tailored to each attribute (Rupp & Templin, 2008). In C-RUM, the item response probability is written as

$$P(X_{ij} = 1 | \alpha_i) = \left\{ 1 + e^{-\left[\beta_j + \sum_{k=1}^K \gamma_{jk} a_{ik} q_{jk} \right]} \right\}^{-1}, \quad (6)$$

where X_{ij} is the response of student i to item j , $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK})'$ is the student attribute pattern, β_j is an item parameter related to the likelihood of correctly answering item j when student i has mastered none of the attributes required by the item, γ_{jk} is related to the increased likelihood of responding to item j correctly when student i has mastered attribute k , $q_{ij} = 0$ or 1 is the element in the Q matrix showing whether attribute k is required for getting a correct response to item j , and $a_{ik} = 0$ or 1 is the k th element of α_i indicating whether or not student i has mastered attribute k .

One way to reflect the discriminating power is to use two probabilities: the probability of answering item j correctly by a student who has mastered none of the attributes required by the item, P_{g_j} , and the probability by a student who has mastered all of attributes required by the item, P_{h_j} . With C-RUM,

$$P_{g_j} = \{1 + e^{-\beta_j}\}^{-1} \quad (7)$$

and

$$P_{h_j} = \left\{ 1 + e^{-\left(\beta_j + \sum_{k \in R_j} \gamma_{jk} \right)} \right\}^{-1}, \quad (8)$$

where R_j is the index set for attributes required by item j . To generate high discriminating items, P_{g_j} and P_{h_j} were generated from the normal distribution with mean = .10 and $SD = .02$ and the normal distribution with mean = .90 and $SD = .02$, respectively. The item parameter β_j can be found based on its direct relationship with P_{g_j} as shown in Equation 7. Then γ_{jk} can be solved based on Equation 8 by assuming that the increased likelihood due to attribute mastery is the same across individual attributes required by each item. That is, all values of γ_{jk} are equal for $k \in R_j$.

For the items with low discriminating power, we followed the same procedure, except that P_{g_j} and P_{h_j} were generated from the normal distribution with mean = .25 and $SD = .05$ and the normal distribution with mean = .75 and $SD = .25$, respectively. To generate tests with mixed

discriminating items, we assumed that half of items were high discriminating and the other half were low discriminating.

For each hypothetical test, the null distributions of I_z and the RCI were investigated for each attribute pattern. That is, item responses of 10,000 simulatees were generated based on each attribute pattern. For each simulated response pattern, the corresponding attribute pattern $\hat{\alpha}$ was estimated using maximum likelihood estimation, and then I_z and the RCI were computed with $\hat{\alpha}$ through Equations 4 and 5, respectively. The 10,000 values of I_z and the RCI were used to obtain the corresponding null distributions for each attribute pattern of the hypothetical test. Each simulation condition was replicated 100 times.

Table 1 presents, for each attribute pattern of tests with three attributes across 100 replications, the average first four moments of the simulated distribution of I_z and the average fifth percentile value of I_z , which can serve as the empirical critical value from the simulated distribution at significance level of .05. Standard deviations are presented in parentheses. Results showed that, for tests with 20 items, when test items were highly discriminating, the means of the simulated distribution of I_z were very close to zero (from .02 to .06) and the standard deviations were very close to one (from .92 to .97) across different attribute patterns. In addition, the distribution of I_z was slightly negatively skewed, with the values of skewness ranging from $-.61$ to $-.41$. Furthermore, the kurtosis tended to be negative but fairly close to zero, with the value of kurtosis from $-.21$ to $-.02$. This suggested that when test items are high discriminating, the shapes of the simulated distributions of I_z are very close to standard normal. Inspection of empirical critical values across different attribute patterns revealed that empirical critical values, ranging from -1.55 to -1.78 , were slightly different from the theoretical value, -1.64 , from the standardized normal distribution. For tests with mixed and low discriminating items, the means and standard deviations were not as close to 0 and 1 anymore compared with those for the test with high discriminating items. However, the shapes of the distributions were very close to normal, with all values of skewness and kurtosis near 0. Inspection of empirical critical values indicates that they were consistently larger than the theoretical critical value, -1.64 . This finding suggests the use of theoretical critical value may result in more conservative classification of misfitting response patterns.

For tests with 40 items, results indicate that when the number of items increased, the distributions of I_z were closer to standard normal, with means closer to 0 and standard deviations closer to 1. For the test with low discriminating items, empirical critical values were all larger than the theoretical value -1.64 , indicating the use of theoretical values could lead to more conservative detection of misfitting responses. However, when items were mixed or high discriminating, the empirical critical values tended to be smaller than the theoretical value. That is, the use of -1.64 as the critical value could result in more liberal detection of misfitting responses. For tests with 6 attributes, similar patterns were found for the empirical distributions of I_z . Given that there were a total of 64 attribute patterns for each test, detailed results are not presented here due to limited space.

Table 2 shows the distributional properties of the RCI for three-attribute tests with 20 and 40 items. The distributions of the RCI appeared to be fairly normal, with values of skewness and kurtosis all close to zero. The comparison of distributions for tests with 20 and 40 items revealed that the means, standard deviations, and empirical critical values at $\alpha = .05$ of the distribution of the RCI depended highly on the number of items, with these values being increased as the number of items increased to 40. However, the distributions of the RCI were more comparable across different attribute patterns for the same test. The distributional properties of the RCI for tests with six attributes followed similar trend, and, due to limited space, they are not presented and discussed here but could be obtained from authors upon request.

Table 1. The Null Distribution of I_z for Three-Attribute Tests.

Discriminating	20-item test					40-item test				
	M	SD	Skewness	Kurtosis	Fifth percentile	M	SD	Skewness	Kurtosis	Fifth percentile
Low										
{0, 0, 0}	0.14 (0.02)	0.87 (0.02)	0.06 (0.05)	-0.26 (0.09)	-1.28 (0.05)	0.32 (0.03)	0.77 (0.02)	0.04 (0.05)	-0.31 (0.08)	-0.94 (0.06)
{0, 0, 1}	0.06 (0.02)	0.92 (0.02)	-0.03 (0.05)	-0.3 (0.07)	-1.46 (0.05)	0.22 (0.03)	0.79 (0.02)	0.03 (0.06)	-0.41 (0.07)	-1.08 (0.06)
{0, 1, 0}	0.06 (0.02)	0.92 (0.02)	-0.02 (0.06)	-0.31 (0.07)	-1.45 (0.05)	0.22 (0.03)	0.8 (0.02)	0.02 (0.06)	-0.39 (0.07)	-1.1 (0.05)
{0, 1, 1}	0.05 (0.02)	0.92 (0.01)	-0.04 (0.05)	-0.32 (0.07)	-1.48 (0.04)	0.2 (0.03)	0.8 (0.02)	0.01 (0.06)	-0.44 (0.07)	-1.12 (0.06)
{1, 0, 0}	0.06 (0.02)	0.92 (0.01)	-0.04 (0.06)	-0.33 (0.08)	-1.48 (0.04)	0.21 (0.03)	0.8 (0.02)	0.02 (0.06)	-0.43 (0.07)	-1.11 (0.06)
{1, 0, 1}	0.06 (0.02)	0.92 (0.02)	-0.04 (0.05)	-0.29 (0.06)	-1.47 (0.05)	0.22 (0.03)	0.8 (0.02)	0.02 (0.06)	-0.41 (0.06)	-1.1 (0.05)
{1, 1, 0}	0.06 (0.02)	0.92 (0.02)	-0.03 (0.05)	-0.3 (0.06)	-1.47 (0.06)	0.22 (0.03)	0.8 (0.02)	0.02 (0.06)	-0.4 (0.07)	-1.09 (0.06)
{1, 1, 1}	0.14 (0.02)	0.87 (0.02)	0.04 (0.05)	-0.26 (0.08)	-1.29 (0.05)	0.32 (0.03)	0.77 (0.02)	0.04 (0.05)	-0.31 (0.07)	-0.94 (0.06)
Mixed										
{0, 0, 0}	0.03 (0.02)	0.96 (0.02)	-0.26 (0.04)	-0.17 (0.1)	-1.64 (0.05)	0.24 (0.03)	0.79 (0.02)	-0.12 (0.05)	-0.48 (0.07)	-1.08 (0.05)
{0, 0, 1}	0 (0.02)	0.99 (0.01)	-0.39 (0.06)	-0.02 (0.12)	-1.74 (0.05)	0.09 (0.03)	0.87 (0.02)	-0.1 (0.06)	-0.44 (0.06)	-1.38 (0.05)
{0, 1, 0}	0 (0.02)	0.99 (0.02)	-0.38 (0.05)	-0.03 (0.13)	-1.74 (0.05)	0.1 (0.03)	0.85 (0.02)	-0.13 (0.05)	-0.44 (0.06)	-1.35 (0.05)
{0, 1, 1}	0.01 (0.02)	0.99 (0.01)	-0.42 (0.05)	0 (0.14)	-1.76 (0.05)	0.09 (0.02)	0.87 (0.02)	-0.14 (0.06)	-0.45 (0.07)	-1.39 (0.06)
{1, 0, 0}	0 (0.02)	0.99 (0.02)	-0.41 (0.05)	0 (0.12)	-1.75 (0.06)	0.09 (0.03)	0.87 (0.02)	-0.14 (0.05)	-0.45 (0.07)	-1.39 (0.06)
{1, 0, 1}	0 (0.02)	0.99 (0.01)	-0.39 (0.06)	0 (0.13)	-1.75 (0.06)	0.1 (0.03)	0.86 (0.02)	-0.13 (0.06)	-0.45 (0.06)	-1.36 (0.05)
{1, 1, 0}	0.01 (0.02)	0.99 (0.01)	-0.39 (0.05)	-0.04 (0.12)	-1.74 (0.05)	0.08 (0.03)	0.87 (0.02)	-0.11 (0.05)	-0.44 (0.07)	-1.39 (0.05)
{1, 1, 1}	0.03 (0.02)	0.96 (0.01)	-0.25 (0.05)	-0.15 (0.1)	-1.63 (0.05)	0.24 (0.02)	0.79 (0.02)	-0.12 (0.05)	-0.48 (0.06)	-1.08 (0.06)
High										
{0, 0, 0}	0.01 (0.02)	0.99 (0.01)	-0.39 (0.05)	0.01 (0.13)	-1.72 (0.06)	0.06 (0.02)	0.92 (0.01)	-0.41 (0.05)	-0.21 (0.12)	-1.56 (0.06)
{0, 0, 1}	0 (0.02)	1 (0.02)	-0.49 (0.05)	0.13 (0.16)	-1.78 (0.06)	0.02 (0.02)	0.96 (0.01)	-0.57 (0.06)	-0.06 (0.14)	-1.73 (0.06)
{0, 1, 0}	0 (0.02)	1 (0.02)	-0.5 (0.06)	0.17 (0.15)	-1.78 (0.06)	0.02 (0.02)	0.96 (0.02)	-0.58 (0.06)	-0.05 (0.13)	-1.74 (0.05)
{0, 1, 1}	0 (0.02)	1 (0.02)	-0.51 (0.06)	0.18 (0.19)	-1.79 (0.06)	0.02 (0.02)	0.97 (0.02)	-0.6 (0.06)	-0.02 (0.15)	-1.78 (0.06)
{1, 0, 0}	0 (0.02)	1 (0.02)	-0.5 (0.06)	0.15 (0.18)	-1.79 (0.06)	0.02 (0.02)	0.97 (0.02)	-0.61 (0.06)	-0.02 (0.14)	-1.78 (0.05)
{1, 0, 1}	0 (0.02)	1 (0.02)	-0.49 (0.06)	0.13 (0.16)	-1.78 (0.06)	0.02 (0.02)	0.96 (0.02)	-0.58 (0.05)	-0.04 (0.13)	-1.74 (0.06)
{1, 1, 0}	0 (0.02)	1 (0.02)	-0.49 (0.05)	0.14 (0.16)	-1.78 (0.06)	0.02 (0.02)	0.96 (0.02)	-0.57 (0.05)	-0.06 (0.13)	-1.74 (0.06)
{1, 1, 1}	0 (0.02)	0.99 (0.02)	-0.38 (0.05)	0 (0.12)	-1.72 (0.07)	0.06 (0.02)	0.92 (0.01)	-0.42 (0.05)	-0.21 (0.1)	-1.55 (0.05)

Note. The values in the parentheses represent the standard deviations across replications.

Table 2. The Null Distribution of RCI for Three-Attribute Tests.

Discriminating	20-item test					40-item test				
	M	SD	Skewness	Kurtosis	Fifth percentile	M	SD	Skewness	Kurtosis	Fifth percentile
Low										
{0, 0, 0}	9.8 (0.16)	2.61 (0.09)	-0.01 (0.05)	-0.29 (0.09)	14.09 (0.28)	4.22 (0.11)	1.57 (0.07)	0.02 (0.05)	-0.31 (0.07)	6.82 (0.23)
{0, 0, 1}	7.96 (0.2)	2.4 (0.08)	0.34 (0.08)	0.24 (0.19)	12.09 (0.26)	3.72 (0.15)	1.5 (0.07)	0.22 (0.07)	-0.16 (0.12)	6.29 (0.18)
{0, 1, 0}	7.99 (0.22)	2.41 (0.08)	0.34 (0.07)	0.23 (0.14)	12.13 (0.26)	3.76 (0.14)	1.51 (0.06)	0.22 (0.06)	-0.18 (0.12)	6.35 (0.2)
{0, 1, 1}	8.21 (0.37)	2.36 (0.07)	0.36 (0.08)	0.19 (0.18)	12.31 (0.36)	3.9 (0.23)	1.49 (0.07)	0.22 (0.07)	-0.24 (0.11)	6.46 (0.24)
{1, 0, 0}	7.68 (0.22)	2.44 (0.07)	0.45 (0.09)	0.35 (0.18)	11.97 (0.23)	3.69 (0.13)	1.53 (0.08)	0.27 (0.07)	-0.21 (0.11)	6.34 (0.19)
{1, 0, 1}	8.44 (0.39)	2.36 (0.08)	0.28 (0.08)	0.12 (0.15)	12.48 (0.38)	3.96 (0.21)	1.47 (0.06)	0.19 (0.06)	-0.22 (0.1)	6.47 (0.22)
{1, 1, 0}	8.47 (0.37)	2.37 (0.08)	0.27 (0.07)	0.09 (0.14)	12.51 (0.4)	4.01 (0.23)	1.48 (0.07)	0.16 (0.05)	-0.26 (0.1)	6.51 (0.26)
{1, 1, 1}	9.94 (0.15)	2.53 (0.09)	-0.02 (0.05)	-0.23 (0.08)	14.11 (0.29)	4.39 (0.14)	1.51 (0.08)	-0.04 (0.07)	-0.24 (0.08)	6.86 (0.25)
Mixed										
{0, 0, 0}	9.48 (0.15)	3.45 (0.08)	0.25 (0.04)	-0.17 (0.1)	15.45 (0.23)	4.08 (0.1)	1.94 (0.07)	0.16 (0.04)	-0.43 (0.06)	7.39 (0.19)
{0, 0, 1}	7.66 (0.22)	2.91 (0.07)	0.47 (0.07)	0.25 (0.2)	12.78 (0.29)	4.02 (0.13)	1.74 (0.06)	0.31 (0.07)	-0.16 (0.12)	7.04 (0.18)
{0, 1, 0}	7.65 (0.24)	2.92 (0.07)	0.47 (0.07)	0.26 (0.22)	12.78 (0.29)	3.96 (0.13)	1.76 (0.07)	0.3 (0.07)	-0.21 (0.12)	7 (0.2)
{0, 1, 1}	8.29 (0.44)	2.82 (0.06)	0.49 (0.07)	0.28 (0.2)	13.27 (0.45)	4.62 (0.22)	1.68 (0.07)	0.28 (0.06)	-0.2 (0.09)	7.56 (0.28)
{1, 0, 0}	7.32 (0.26)	2.84 (0.06)	0.56 (0.08)	0.52 (0.26)	12.33 (0.34)	3.9 (0.13)	1.77 (0.07)	0.41 (0.07)	-0.08 (0.12)	7.06 (0.18)
{1, 0, 1}	8.8 (0.45)	2.9 (0.07)	0.43 (0.06)	0.16 (0.16)	13.9 (0.49)	4.73 (0.26)	1.72 (0.06)	0.23 (0.06)	-0.25 (0.08)	7.7 (0.3)
{1, 1, 0}	8.72 (0.35)	2.89 (0.06)	0.42 (0.06)	0.13 (0.16)	13.8 (0.38)	4.75 (0.29)	1.69 (0.07)	0.24 (0.06)	-0.22 (0.09)	7.66 (0.32)
{1, 1, 1}	9.54 (0.16)	3.45 (0.07)	0.23 (0.05)	-0.17 (0.1)	15.47 (0.25)	4.31 (0.09)	1.97 (0.08)	0.05 (0.04)	-0.46 (0.06)	7.56 (0.2)
High										
{0, 0, 0}	8.66 (0.18)	4.09 (0.06)	0.38 (0.05)	0 (0.13)	15.76 (0.3)	4.09 (0.12)	2.64 (0.04)	0.39 (0.05)	-0.25 (0.11)	8.8 (0.11)
{0, 0, 1}	7.22 (0.24)	3.33 (0.05)	0.51 (0.06)	0.23 (0.2)	13.15 (0.31)	3.61 (0.14)	2.31 (0.04)	0.65 (0.06)	0.17 (0.18)	7.84 (0.18)
{0, 1, 0}	7.29 (0.24)	3.34 (0.06)	0.52 (0.06)	0.27 (0.2)	13.23 (0.31)	3.59 (0.15)	2.31 (0.04)	0.65 (0.06)	0.19 (0.18)	7.86 (0.18)
{0, 1, 1}	8.56 (0.41)	3.21 (0.05)	0.52 (0.07)	0.28 (0.25)	14.26 (0.47)	4.35 (0.27)	2.22 (0.04)	0.65 (0.06)	0.2 (0.17)	8.55 (0.29)
{1, 0, 0}	6.98 (0.23)	3.22 (0.06)	0.54 (0.08)	0.33 (0.26)	12.69 (0.28)	3.51 (0.17)	2.27 (0.04)	0.74 (0.07)	0.4 (0.22)	7.73 (0.21)
{1, 0, 1}	8.91 (0.44)	3.33 (0.05)	0.49 (0.06)	0.18 (0.18)	14.84 (0.47)	4.39 (0.27)	2.29 (0.04)	0.61 (0.06)	0.13 (0.16)	8.62 (0.28)
{1, 1, 0}	8.85 (0.45)	3.33 (0.07)	0.49 (0.06)	0.19 (0.18)	14.76 (0.52)	4.4 (0.27)	2.29 (0.04)	0.61 (0.07)	0.11 (0.16)	8.66 (0.28)
{1, 1, 1}	8.68 (0.18)	4.1 (0.07)	0.38 (0.05)	-0.02 (0.12)	15.8 (0.3)	4.18 (0.12)	2.69 (0.04)	0.37 (0.05)	-0.29 (0.1)	8.87 (0.11)

Note. The values in the parentheses represent the standard deviations across replications. RCI = response conformity index.

Simulation Study 2: The Type I Errors and Power of I_z and RCI

A second simulation study was conducted to examine the Type I errors of I_z and RCI as well as their detection power in detecting five types of misfitting response patterns for each of the 12 tests used in the above simulation study. The first type of misfit was associated with high-ability students' creative responding behavior. As discussed by Meijer (1994), creative responding behavior causes a high-ability student to produce incorrect responses to easy items. In this study, easy items are operationalized as those that measure only one attribute. Creative item responses were simulated by generating item responses from high-ability students who had mastered all the attributes measured by the test and inputting incorrect responses to items that solely measured one attribute. The second type of misfit dealt with students having low test-taking motivation that led to random guessing responses, which has been proposed as being possibly associated with person misfit (e.g., Petridou & Williams, 2007). Students' true attribute patterns were randomly generated by assuming that each student had a probability of .80 mastering each attribute. Their item responses were randomly generated with a probability of .25 for a correct response. The probability of .25 here corresponds to the probability of producing a correct response by blindly guessing to a multiple-choice item with four options. The third type of misfit examines the "sleeping" behavior, defined by Wright and Stone (1979) as failing to answer correctly the first several questions in a test due to problems of getting started. This type of misfit was generated by setting responses to the first 25% of items to zero. The fourth type of misfit was related to fatigue. As discussed by Petridou and Williams (2007), external factors such as fatigue may also influence student responses thereby resulting in person misfit. These responses were generated by setting the last 25% of item responses to zero. It should be noted that the discriminating power of test items was randomly generated according to the distribution outlined in the previous section, and therefore item characteristics were expected to distribute evenly among the first, middle, and last parts of the test. The last type of misfit was related to the misspecification of the Q matrix. This type of misfit is unique to cognitive diagnostic tests in which the Q matrix is used to specify the attributes measured by each item. According to Kunina-Habenicht, Rupp, and Wilhelm (2012), invalid Q-matrix specifications could lead to notably decreased measurement accuracy of diagnostic tests. In addition, as reported by Cui and Roberts (2013), the misspecification of the Q matrix could also result in person misfit, which threatens the validity of score inferences. To generate this type of misfit, we assumed that 25% of Q-matrix entries were misspecified for some students, thereby creating misfitting responses.

For each type of misfit, 1,000 misfitting response patterns were randomly generated, and values I_z and RCI were calculated for each pattern using the estimated attribute pattern. For both I_z and the RCI, empirical critical values were found based on the empirical distribution of fitting response patterns at significance level of .05 using the simulated data of the previous section. The power rates of I_z and RCI were indicated by the average percentage of misfitting patterns classified as misfit by each statistic across the 1,000 replications of each simulation condition. To examine the Type I error rates, we generated 1,000 fitting responses for each test. Students' true attribute patterns were randomly generated by assuming that each student had a probability of .80 mastering each attribute. The empirical critical values of the RCI and I_z associated with the estimated attribute pattern was used to classify the response pattern. Type I error rates of each statistic were reflected by the percentage of fitting responses that were classified as misfit by each statistic across 100 replications.

For the purpose of comparison, Liu et al.'s (2009) marginal likelihood ratio test was also included in the simulation. To use the test for detecting misfits, hypothesis about the type of misfitting behavior, spuriously high or spuriously low, needs to be specified. Given that four

out of five types of misfitting behavior simulated in this study (creative responding, random responding due to low motivation, sleeping, and fatigue) tended to result in students failing to correctly answer as many items as they should, spuriously low responding was assumed. The MATLAB codes developed by Liu et al. were requested and implemented in the current study.

Table 3 gives the Type I error rates and the detection power of the three person fit statistics under different simulation conditions. In terms of Type I errors, results suggested the Type I error rates of both the RCI and the I_z were mostly reasonable, around the nominal level, .05. One exception was the Type I error of the RCI was .01 for the condition of 40 items, mixed discriminating power, and six attributes. However, it appeared that Liu et al.'s likelihood ratio test was associated with inflated Type I error rates, with values greater than .05 under most of simulation conditions (values ranging from .05 to .24). This finding was not consistent with the results of Liu et al.'s (2009) study where the Type I error rates were found to be reasonable, typically around the nominal level of .05. This could be possibly related to the current use of C-RUM model in the simulation where the DINA (deterministic input noisy and) model (de la Torre & Douglas, 2004) was initially used by Liu et al. as the basis for data generation.

For the power study, results suggested that test length and item discrimination tended to show positive impact on the power of all three statistics, while the number of attributes was found to negatively affect the detection rates. This finding is expected because misfitting patterns tend to become more detectable with more reliable tests as a result of having more and higher discriminating items but measuring fewer attributes. This result is in agreement with that from Cui and Leighton (2009). Of the three person fit statistics, not surprisingly, Liu et al.'s marginal likelihood ratio test was associated with the highest detection power under most of the simulation conditions given that the test, according to the Type I error results, was found to be overly liberal. It should be noted that when the Type I errors of the likelihood ratio test were close to .05 under the simulation condition of low discriminating items and six attributes, its detection rates tended to be higher than RCI and I_z . More detailed results of the marginal likelihood test will not be discussed further. The detection rates of the RCI ranged from .58 to 1.00 for the creative responding behavior, from .08 to .41 for random guessing, from .24 to .94 for the "sleeping" behavior, from .03 to .59 for fatigue, and from .03 to .51 for misfitting due to the misspecification of the Q matrices. The rates for I_z were between .27 and 1.00 for creative responding, between .04 and .38 for random guessing, between .23 and .93 for the "sleeping" behavior, between .12 and .70 for fatigue, and between .04 and .28 for Q misspecifications. Results suggested that RCI and I_z were most effective in detecting creative behavior. The comparison of RCI and I_z revealed that the RCI tend to outperform I_z in detecting creative responding, guessing, sleeping, and Q misspecification, while I_z outperformed the RCI in detecting misfitting responses due to fatigue.

Summary and Conclusion

This article aims to examine ways of detecting person misfit for CDMs. We first investigated whether the well-known I_z statistic, developed under IRT, could be extended for use in the context of CDMs. The authors conducted a simulation study to investigate the distributional properties of the I_z statistic. Results show that for tests with high discriminating items, the distributions of I_z are very close to standard normal distributions with the means close to zero and the standard deviations close to one. One might find surprising that the distributions of I_z are very close to standard normal even when estimated ability parameters are used in the calculation of I_z as this finding is inconsistent with the previous results (e.g., van Krimpen-Stoop & Meijer, 1999).

One plausible explanation is that the negative impact of the bias of the point estimates of student ability parameters on the distributions of I_z tends to be reduced when the coarser classification results produced by CDMs (i.e., the mastery or non-mastery of a set of attributes) are used in calculating I_z . When the items are mixed or low discriminating, the means and standard deviations deviate from 0 and 1, especially when the test is short with 20 items. This is somewhat expected as the item and student parameter estimates likely become less accurate with poorer items. And as a result, the empirical distributions of I_z will deviate from the theoretical standard normal distribution when estimated item and student parameters are used in calculating I_z .

The authors also introduced a new person fit statistic, called the response conformity index (RCI), to specifically examine the consistency of observed response patterns relative to the expectations based on the Q matrix. Simulated data suggest the distributions of the RCI be fairly normal, with values of skewness and kurtosis close to zero. Also, the distributions of the RCI tend to be more comparable across different attribute patterns for the same test. However, the means and standard deviations of the distribution of the RCI depend greatly on the number of items, with greater values as the number of items increases. To evaluate and compare the performance of the RCI and I_z , we conducted a second simulation study. Liu et al.'s (2009) marginal likelihood ratio test was also examined in the simulation for the purpose of comparison. Although our simulation is based on the C-RUM model, it is important to note that the two statistics can be applied to many other CDMs in which the item response function is explicitly defined. According to the simulation results, higher detection rates of the three statistics are found with tests with more and higher discriminating items, which is consistent with expectations. Simulation results indicate that the marginal likelihood ratio test appears to be liberal with inflated Type I errors under most of the simulation conditions considered in the current study. As expected, its detection rates are mostly higher than the RCI and I_z . In comparison, the Type I error rates of RCI and I_z are mostly reasonable, around the nominal level of .05. Between RCI and I_z , the RCI tends to outperform I_z in detecting creative responding, guessing, sleeping, and Q misspecifications. For fatigue, the performance of the I_z is superior. It should be noted that the two statistics do not compete but rather compensate each other as the two statistics examine the misfit of student responses from different perspectives. The RCI measures directly the degree to which the observed item responses deviate from the ideal responses based on estimated student attribute patterns. If large discrepancies exist between student observed and ideal responses, the responses are more likely to be detected by the RCI as misfitting. For I_z , the focus is on the comparison between the student observed responses and the model-calculated probabilities of the observed responses. When the mismatch between a student's observed responses and the ideal responses is not obvious but large discrepancies are found between the observed response and the calculated probability, the use of I_z is more advantageous.

In the simulation, the authors examined the effect of test length, number of attributes, and item discriminating power on the Type I errors and power of the RCI, I_z , and Liu et al.'s likelihood ratio test in detecting different types of misfit. One line of the future research is to expand our simulation by including other factors such as different CDMs and sample size to examine their impact on the Type I error rates and detection power of different person fit statistics for cognitive diagnostic assessment. In addition, it is worth investigating why the marginal likelihood ratio test was found to be liberal within the current simulation conditions. The distributional properties of the likelihood ratio test should be further examined to see whether the asymptotic chi-square distribution holds under different settings.

Appendix

Q Matrices Used in the Simulation.

3 attributes 20 items	6 attributes 20 items	3 attributes 40 items	6 attributes 40 items
100	100000	100	100000
010	010000	010	010000
001	001000	001	001000
100	000100	100	000100
010	000010	010	000010
001	000001	001	000001
100	010011	100	101110
010	000011	010	010000
001	111100	001	011010
110	001011	110	000010
101	010110	101	111100
011	111110	011	100101
110	111110	110	100001
101	101010	101	011101
110	001101	110	100101
101	001010	101	011011
011	100111	011	100101
111	011000	111	000010
111	101001	111	100100
111	010100	111	001110
		100	100110
		010	111011
		001	000011
		100	001001
		010	001011
		001	010110
		100	001010
		010	000001
		001	000101
		110	101101
		101	000111
		011	000010
		110	010111
		101	000110
		110	101010
		101	100011
		011	010001
		111	100001
		111	111001
		111	100010

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this article was supported by a grant from the Social Sciences and

Humanities Research Council of Canada. Grantees undertaking such projects are encouraged to express freely their professional judgment. This article, therefore, does not necessarily represent the positions or the policies of the Canadian government, and no official endorsement should be inferred.

References

- Choi, H.-J., Templin, J. L., Cohen, A. S., & Atwood, C. H. (2010). *The impact of model misspecification on estimation accuracy in diagnostic classification models*. Paper presented at the meeting of the National Council on Measurement in Education (NCME), Denver, CO.
- Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, 49, 19-38.
- Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, 46, 429-449.
- Cui, Y., & Roberts, M. R. (2013). Validating student score inferences with person-fit statistic and verbal reports: A person-fit study for cognitive diagnostic assessment. *Educational Measurement: Issues and Practice*, 32, 34-42.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement*, 34, 267-285.
- DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 979-1027). Amsterdam, The Netherlands: Elsevier.
- 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, 67-86.
- Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, 10, 318-341.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- 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, 59-81. doi:10.1111/j.1745-3984.2011.00160.x
- Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, 27, 3-16.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoaka's rule-space approach. *Journal of Educational Measurement*, 41, 205-237.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269-290.
- Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, 33, 579-598.
- Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, 18, 311-314.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107-135.
- Meijer, R. R., & Tendeiro, J. N. (2012). The use of the Iz and Iz* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics*, 37, 758-766.
- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55, 75-106.

- Petridou, A., & Williams, J. (2007). Accounting for aberrant test response patterns using multilevel models. *Journal of Educational Measurement, 44*, 227-247.
- Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55*, 3-38.
- Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*, 219-262.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Seo, D. G., & Weiss, D. J. (2013). Iz person-fit index to identify misfit students with achievement test data. *Educational and Psychological Measurement, 73*, 994-1016.
- Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 51*, 337-350.
- Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum.
- Tatsuoka, K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York, NY: Routledge.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-346.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26*, 164-180.
- von Davier, M., & Molennar, I. (2003). A person-fit index for polytomous rasch models, latent class models and their mixture generalizations. *Psychometrika, 68*, 213-228.
- Wright, B. D., & Stone, M. H. (1979). *Best test design (Rasch measurement)*. Chicago, IL: Mesa Press.