Research on Person Fit in Cognitive Diagnostic Assessment

Authors: Yu Xiaofeng, Tang Qian, Qin Chunying, Li Yujun, YU Xiaofeng

Date: 2022-05-12T16:42:54Z

Abstract

Typically, cognitive diagnosis necessitates diagnostic evaluation of examinees through cognitive diagnosis models. The validity of diagnostic results generated by cognitive diagnosis models depends on the congruence between examinees' response patterns and the selected model. Consequently, when evaluating diagnostic outcomes, subject-fit analysis is required to examine the fit between individual examinees' responses and the model, thereby preventing erroneous or ineffective remedial interventions. This study proposes a novel subject-fit index, R, for cognitive diagnosis assessment based on weighted score residuals. Simulation research indicates that the R index demonstrates satisfactory stability in Type I error rates and high statistical power for detecting four types of aberrant response patterns: random responding, fatigue, sleeping, and creative responding. Furthermore, the R index is applied to empirical fraction subtraction data to demonstrate its practical implementation in operational testing.

Full Text

Research on Person-Fit in Cognitive Diagnostic Assessment

Yu Xiaofeng[1], Tang Qian[1], Qin Chunying[2], Li Yujun[1]
([1]School of Psychology, Jiangxi Normal University, Nanchang, 330022)
([2]School of Mathematics and Information Science, Nanchang Normal University, Nanchang, 330032)

Abstract
Cognitive Diagnostic Assessment (CDA) has been widely used in psychological and educational measurement to analyze whether test-takers have mastered specific knowledge points or skills, thereby providing guidance for further learning and instruction (Leighton & Gierl, 2007; Rupp et al., 2010; Tatsuoka, 1983). The validity of diagnostic results generated by cognitive diagnostic models depends on whether examinees'response patterns fit the selected model. Therefore,

when evaluating diagnostic outcomes, person-fit analysis must be conducted to examine the alignment between individual response patterns and the model, thereby avoiding erroneous or ineffective remedial measures. This study proposes a new person-fit index, denoted as R, based on weighted score residuals. Simulation results demonstrate that the R index maintains stable Type I error rates and exhibits high statistical power for detecting four types of aberrant examinees: random responders, fatigued examinees, sleepers, and creative responders. The R index is further applied to empirical fraction-subtraction data to illustrate its practical utility in real testing situations.

Keywords: cognitive diagnosis, person-fit, DINA model, aberrant response

1. Introduction

In recent years, Cognitive Diagnostic Assessment (CDA) has gained widespread application in psychological and educational measurement, enabling analysis of whether examinees have mastered specific knowledge points or skills to guide further learning and instruction (Leighton & Gierl, 2007; Rupp et al., 2010; Tatsuoka, 1983). Cognitive diagnostic models serve as statistical tools in this process, facilitating inferences about examinees' attribute mastery patterns (von Davier & Lee, 2019). The fit between cognitive diagnostic models and test data directly affects the accuracy of diagnostic results and influences the overall reliability and validity of the assessment. Consequently, evaluating model-data fit is essential in cognitive diagnostic evaluation. Standard 5.19 of the *Standards for Educational and Psychological Testing* (p. 107) explicitly requires fit testing between selected item response models and response data in educational and psychological measurement.

In educational measurement, test scores are used to assess examinees' ability levels. However, due to potential aberrant behaviors, these scores may not accurately reflect true skill or knowledge levels. In psychometrics, methods that quantify discrepancies between observed responses and model-predicted responses are termed person-fit statistics (Meijer & Sijtsma, 2001). Person-fit analysis examines the degree to which individual response patterns align with cognitive diagnostic models. An appropriate cognitive diagnostic model should accurately reflect the psychological processing characteristics of examinees during item responding to enable valid inferences about attribute mastery. When examinee response patterns fit the selected cognitive diagnostic model, this is called person-fit; conversely, when aberrant response patterns fail to fit the model, this is termed person-misfit. Person-misfit can lead to uninterpretable or invalid inferences about attribute mastery patterns for misfitting examinees, potentially resulting in inappropriate remedial measures. Additionally, misfitting data may compromise the overall reliability and validity of the test, making person-fit analysis particularly crucial.

Previous person-fit research has predominantly focused on Item Response Theory (IRT; Baker & Kim, 2004), with relatively limited attention to person-fit

within the cognitive diagnostic framework. Existing studies in this area include: Liu et al. (2009), who proposed a likelihood ratio test statistic for identifying aberrant responders based on marginal and joint likelihood ratio tests, introducing an aberrant response probability variable ρ_i and using indicator variable A to define aberrant response types, though this approach has limitations as aberrant examinees and response types are difficult to define in practice; Cui and Leighton (2009), who developed the Hierarchical Consistency Index (HCI) to measure the match between observed and ideal response patterns under attribute hierarchy models, which is limited when attributes have only partial hierarchical relationships or no hierarchical relationships; Liu et al. (2009) demonstrated that their likelihood ratio statistic effectively detects spuriously high and low scores when using the DINA model; Cui and Li (2015) extended the l_z index to the cognitive diagnostic framework and proposed the Response Conformity Index (RCI); and review studies on person-fit in cognitive diagnostic testing (Chen et al., 2016; Tu et al., 2014).

Given the importance of person-fit research in diagnostic testing, this study aims to develop a person-fit index for cognitive diagnostic assessments and compare its performance with the l_z and RCI indices under various conditions. Detailed introductions to the l_z and RCI indices are provided in Appendix A.

2. Development of the Person-Fit Index R in Cognitive Diagnostic Assessment

Residuals represent a fundamental concept in regression analysis, defined as deviations between observed values and expected (fitted) values in mathematical statistics. The underlying logic of residual application involves identifying anomalies by contrasting ideal versus actual situations. Expected deviations inflate residual statistics, aligning with the conceptual framework of person-fit testing. This study proposes constructing a residual-based person-fit statistic, the R index, for diagnostic testing. We first define standardized residuals below.

2.1 Definition of Standardized Residuals

Numerous studies on Rasch models and other IRT applications have utilized standardized residuals of the form $\frac{x_{ij}-E(X_{ij}|\theta_i)}{\sqrt{Var(X_{ij}|\theta_i)}}$, where $Var(X_{ij}|\theta_i)$ represents the variance of random variable X_{ij} given ability value θ_i (Masters & Wright, 1997). Summing standardized residuals across items for each examinee yields a person-fit evaluation metric. On one hand, standardized residuals can be viewed as weighted residuals where weights are the inverse of conditional standard errors of item responses, approximately following a standard normal distribution. On the other hand, since person-fit focuses on consistency between observed and model-predicted responses, severe inconsistencies result in low probabilities for the observed response pattern. Because this probability appears in the denominator as an inverse weight, it artificially inflates residual values. Based on these

considerations, this study uses the inverse of observed response probabilities as weights for the person-fit statistic.

2.2 Definition of the R Index

The mathematical expression for the R index is:

$$R_i = \sum \log \left[\frac{x_{ij} - E(X_{ij}|\alpha_i)}{P(x_{ij}|\alpha_i)} \right]$$

where x_{ij} denotes the observed score of examinee i on item j, and α_i represents examinee i' s attribute mastery pattern. In practice, true attribute mastery patterns are unobservable, so this study employs estimated attribute mastery patterns. $E(X_{ij}|\alpha_i)$ indicates the expected score of examinee i with attribute mastery pattern α_i on item j. In the DINA model (de la Torre, 2009), each item has two parameters: slipping parameter s and guessing parameter g. If examinee i has mastered all attributes required by item j, then $E(X_{ij}|\alpha_i) = 1 - s_j$; if examinee i has not mastered at least one required attribute, then $E(X_{ij}|\alpha_i) = g_j$. The numerator represents the difference between observed and expected responses.

The denominator $P(x_{ij}|\alpha_i)$ denotes the probability of examinee i with attribute mastery pattern α_i obtaining score x_{ij} on item j. When an examinee with pattern α_i has mastered all required attributes and responds correctly, $P(x_{ij} = 1|\alpha_i) = E(X_{ij}|\alpha_i)$. Smaller $P(x_{ij}|\alpha_i)$ values indicate greater person-misfit, further amplifying residuals between observed and expected responses. R_i represents the sum of R values across all items for examinee i, with larger values indicating poorer fit. For a "well-fitting" examinee, R_i is expected to be relatively small. Importantly, the R index is not dependent on a specific diagnostic model; the DINA model is used as an example due to its simplicity, ease of use, and availability in numerous open-source software packages. For detailed information on the DINA model, see de la Torre (2009), Junker & Sijtsma (2001), and von Davier & Lee (2019).

In the DINA model, each examinee' s completed items can be categorized into four types based on attribute mastery and response accuracy: mastered attributes with correct response (η_{11}) or incorrect response (η_{10}); incomplete mastery with incorrect response (η_{00}) or correct response (η_{01}). Here, η represents the count of corresponding item types, with the first subscript indicating complete attribute mastery and the second indicating correct response (value of 1 indicates mastery or correct response). Thus, Formula 1 can be rewritten as:

$$R_i = \sum \log \left[\log \left[\log \left[\log \left[\right] \right] \right] \right], \quad (2)$$

where J_{11}, J_{10}, J_{00}, and J_{01} correspond to the numbers of items for η_{11}, η_{10}, η_{00}, and η_{01}, respectively. Furthermore, when both s_j and g_j are less than 0.5,

Formula 2 transforms to:

$$R_i = 2\left\{\sum \log[] + \sum \log[] + \sum \log[\log[]]\right\}, \quad (3)$$

For a "well-fitting"examinee, J_{10} and J_{01} should be small, making log [] and log [] negative values and resulting in a smaller R_i. For a "poorly-fitting" examinee, J_{10} and J_{01} are relatively larger, making log [] and log [] positive values and yielding a larger R_i.

3. Study 1: Comparative Analysis of R, l_z, and RCI Indices

To evaluate the performance of the R index in person-fit testing for diagnostic assessments, we conducted a simulation study comparing the R index with the l_z and RCI indices. Cui and Li (2015) demonstrated that the RCI index outperformed Liu et al.' s (2009) likelihood ratio statistic, so the latter was not included as a comparison.

3.1 Research Design

This study examined Type I error rates and statistical power of the R_i, l_z, and RCI indices under varying conditions of test length, item quality, and aberrant examinee types in the DINA model. Item length and quality are critical factors affecting diagnostic measurement accuracy (Cui et al., 2012). Type I error rate (false positive rate) refers to the proportion of normal examinees incorrectly flagged as misfitting, while statistical power represents the proportion of correctly identified aberrant examinees.

Experimental Design: A 2×2×6 fully randomized factorial design was employed with three factors: test length (20 vs. 40 items), item quality (high vs. low discrimination), and aberrant examinee type (creative responding, random responding, fatigue, sleeping, cheating, and random cheating; Cui & Li, 2015; Santos et al., 2020). High-discrimination items had slipping parameters s and guessing parameters g drawn from a uniform distribution $U(0.05, 0.25)$, while low-discrimination items used $U(0.25, 0.40)$. Following Cui and Li (2015), creative responding was defined as high-ability examinees (those who mastered all attributes) incorrectly answering easy items (those measuring only one attribute). The simulation assumed each examinee had an 80% probability of mastering each attribute, with attribute mastery patterns generated randomly, and examinees were set to answer single-attribute items incorrectly. Random responding represented low-motivation examinees guessing randomly, operationalized as a 25% probability of correct response per item (Yu & Cheng, 2019). Sleeping behavior was simulated as incorrect responses on the first 25% of items, while fatigue was incorrect responses on the last 25% of items. Cheating was defined as low-ability examinees (those mastering fewer than 2 attributes with 20% mastery probability per attribute) correctly answering difficult items (those requiring 3+ attributes). Random cheating represented low-ability examinees

correctly answering 10% of difficult items with 90% probability (Santos et al., 2020).

Control variables included: 1,000 examinees, the DINA model as the cognitive diagnosis model, six attributes, and a fixed Q-matrix (see Appendix B for details). Examinee knowledge states and item parameters were generated and estimated using R with the DINA model. The simulation was replicated 30 times, evaluating Type I error rates and statistical power at significance level $\alpha = 0.05$. Type I error was calculated as the proportion of misfit flags among 1,000 normal response patterns generated from the DINA model under each condition. Statistical power was calculated as the proportion of detected aberrant examinees among 1,000 simulated aberrant examinees for each type. Final results were averaged across the 30 replications.

For the l_z and RCI indices, critical values were determined from theoretical distributions at the 0.05 significance level: the 5th percentile for l_z and the 95th percentile for RCI. For the R index, empirical critical values were used: given the Q-matrix and DINA model, 10,000 normal examinee response patterns were simulated assuming uniformly distributed knowledge states. Item parameters were estimated using MMLE/EM (de la Torre, 2009), R_i values were computed for each examinee, sorted ascending, and the 95th percentile was used as the critical value.

3.2 Results

Table 1 presents Type I error rates and statistical power for the three indices across experimental conditions, while Table 2 shows pattern correct classification rates (PCCR) and attribute correct classification rates (ACCR). Type I error results indicate that the R index maintains good control, stable at 0.05, whereas the l_z and RCI indices show slight inflation, with RCI approaching reasonable levels at 40 items. This differs somewhat from Cui et al. (2015), who found normal Type I error rates for l_z and RCI, likely due to our use of the DINA model versus their C-RUM model.

Regarding statistical power, all indices showed improved detection as item discrimination increased, with l_z demonstrating particularly notable gains for fatigue, sleeping, creative responding, and random responding—consistent with Cui and Li (2015). Increasing test length from 20 to 40 items generally improved power, though l_z showed slight decreases for fatigue and sleeping, and R showed slight decreases for random cheating.

Across aberrant types, the R index performed best for random responding and random cheating. The l_z index was superior for fatigue, sleeping, and creative responding, though R approached l_z performance as test length increased, likely due to improved pattern and attribute classification accuracy. For low-discrimination items, R outperformed both l_z and RCI for fatigue and sleeping. For cheating behavior, RCI performed best and most stably, while l_z performed poorly.

Overall, detection rates improved with increased test length and item quality, with creative responding being most easily detected. RCI is best suited for detecting cheating, l_z for fatigue and sleeping, while R shows good power for creative responding, random responding, and cheating, and remains most robust under low item discrimination.

4. Study 2: Application of the R Index to Empirical Data

Educational assessment tools should reflect students' learning status and provide feedback for instructional improvement. Cognitive diagnostic assessment classifies examinees' mastery levels on tested attributes, identifying which attributes require remediation. Person-fit testing ensures the accuracy and validity of these classifications. To further examine the practical feasibility of the R index, this section applies it to person-fit testing and analysis using fraction subtraction data.

4.1 Empirical Data Source

This study uses the widely-cited Tatsuoka fraction subtraction dataset, comprising 536 examinees and 11 items (Henson et al., 2009). The test assesses three attributes: A1) borrowing from whole number, A2) separating whole number from fraction, and A3) finding common denominator. The test Q-matrix is shown in Table 3.

4.2 Methods and Results

Using the fraction subtraction Q-matrix and response data, item parameters and examinee attribute mastery patterns were estimated via the DINA model using the GDINA package in R (item parameters shown in Table 4). Based on these estimates, 10,000 normal examinee response patterns were simulated, and the 95th percentile of R_i values was used as the critical value for flagging aberrant examinees. The R index was then applied to the empirical data, with results compared to those from RCI and l_z indices (see Appendix C for details).

Results identified 23 examinees (4.29%) with aberrant response patterns. Table 5 summarizes selected cases. Examinees 24, 48, and 97 correctly answered items 5, 6, 9, and 10, which measure attribute 1 four times and attribute 2 twice, but not attribute 3. Their estimated attribute mastery pattern was [110], with an ideal response pattern of [10011100111], yet they incorrectly answered items 1, 4, and 11, which test attributes A1 and A2, suggesting possible incomplete mastery of attribute 2.

Examinee 137 showed response pattern [00001011111] with estimated mastery pattern [111]. Theoretically, mastering all attributes should yield perfect performance, but the first four incorrect responses suggest possible "sleeping" behavior.

Examinee 230 had estimated mastery pattern [000] but observed response pattern [01100100110], correctly answering items 2, 3, 6, 9, and 10, suggesting

potential cheating.

5. Discussion and Future Directions

This study proposes the R index for person-fit analysis in cognitive diagnostic assessment and compares it with l_z and RCI indices. Simulation results show the R index maintains reasonable Type I error rates around 0.05, making it suitable for detecting aberrant response patterns. As expected, detection rates improve with increased test length and item discrimination. However, l_z showed slightly inflated Type I error and decreased power for fatigue and sleeping as test length increased, diverging from Cui et al. (2015), possibly due to model differences requiring further investigation.

Second, since the theoretical distribution of the R index remains unknown, this study employed empirical critical values, which may limit practical applicability. Exploring the statistical properties of the R index and deriving its theoretical null or approximate distribution (Andrews, 1993) would facilitate broader application.

Third, the current R index sums across items for each examinee. An alternative formulation summing across examinees for each item could enable item-fit testing (Drasgow et al., 1985), representing a worthwhile extension.

Fourth, item quality substantially impacts person-fit testing, yet this study did not fully incorporate item quality considerations—a limitation for future research. Additionally, Wang et al. (2018) attempted to identify specific types of aberrant behavior, requiring deeper exploration. In empirical studies, using existing datasets precluded further analysis and remediation of flagged examinees. Moreover, person-fit indices alone cannot determine the actual causes of aberrant responses, necessitating auxiliary information such as verbal reports, seating arrangements, and testing time for comprehensive analysis.

Finally, dichotomous models only assess whether knowledge or skills are mastered, not the degree of mastery. Real-world educational and psychological assessments include varied item formats (e.g., constructed-response, essays, Likert scales) that produce polytomous data (Ding et al., 2014; Xia et al., 2018; Wang et al., 2019) or data with multiple attribute levels (Ding et al., 2015; Zhan et al., 2017). Future research should extend person-fit testing to polytomous scoring and multiple-attribute cognitive diagnosis.

6. Conclusion

This study proposes the person-fit index R for cognitive diagnostic frameworks. Through simulation studies comparing Type I error rates and statistical power of RCI, l_z, and R indices, and applying R to empirical data, we find: (1) The R index maintains reasonable Type I error rates, while l_z and RCI show slight inflation; (2) Statistical power improves with increased item discrimination and

test length; (3) RCI is optimal for detecting cheating, l_z for fatigue and sleeping, while R shows strong detection capability for creative responding, random responding, and cheating behaviors.

Appendix A: The l_z and RCI Indices

(1) l_z Index

Cui and Li (2015) adapted the l_z index (Drasgow et al., 1985) for cognitive diagnostic testing. Originally an IRT-based person-fit statistic derived from the likelihood function l_0 (Levine & Rubin, 1979), l_z standardizes l_0:

$$l_{0i} = \ln\left\{\prod P_j(\theta_i)^{X_{ij}}[1-P_j(\theta_i)]^{1-X_{ij}}\right\} \quad \text{(A-1)}$$

where X_{ij} is the dichotomous (0,1) observed response of examinee i to item j, and $P_j(\theta_i)$ is the probability of correct response for examinee i with ability θ_i on item j. Small l_{0i} values indicate low probability of observing response pattern X_i for ability θ_i under the specified IRT model. Standardizing l_{0i} yields:

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

where

$$E(l_0) = \sum\{P_j(\theta)\ln[P_j(\theta)] + [1-P_j(\theta)]\ln[1-P_j(\theta)]\} \quad \text{(A-3)}$$

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

Cui and Li (2015) replaced $P_j(\theta_i)$ with $P_j(\alpha_i)$ from cognitive diagnostic models. Their simulations revealed that l_z based on estimated attribute mastery patterns showed negative skewness, consistent with IRT findings (Molenaar & Hoijtink, 1990; Reise, 1995).

(2) Response Conformity Index (RCI)

Since HCI depends on attribute hierarchies and becomes inapplicable when no such relationships exist, Cui and Li (2015) proposed RCI to assess consistency between Q-matrix-predicted and observed responses:

$$RCI_i = \sum|RCI_{ij}| = \sum\left|\ln\left[\frac{X_{ij}-P_j(\alpha_i)}{I_j(\alpha_i)-P_j(\alpha_i)}\right]\right| \quad \text{(A-5)}$$

where α_i is examinee i' s attribute mastery pattern, $P_j(\alpha_i)$ is the probability of correct response for pattern α_i on item j, and $I_j(\alpha_i)$ is the ideal response (1 if all required attributes are mastered, 0 otherwise). X_{ij} is the observed response (0 or 1).

For each item, RCI_i measures deviation between observed response X_{ij} and ideal response $I_j(\alpha_i)$. When $X_{ij} = I_j(\alpha_i)$, $RCI_i = 0$, indicating excellent fit.

When $X_{ij} \neq I_j(\alpha_i)$, person-fit depends on the magnitude of differences between $X_{ij} - P_j(\alpha_i)$ and $I_j(\alpha_i) - P_j(\alpha_i)$. Large $X_{ij} - P_j(\alpha_i)$ relative to $I_j(\alpha_i) - P_j(\alpha_i)$ suggests aberrant behavior (e.g., cheating, creative responding), yielding large positive RCI values. Conversely, large $I_j(\alpha_i) - P_j(\alpha_i)$ relative to $X_{ij} - P_j(\alpha_i)$ may indicate poor item quality or use of strategies not specified in the Q-matrix, also producing large RCI values.

Appendix B: Q-Matrices Used in Study 1

Table B-1: Q-matrix for simulated data with $K = 6$, $J = 20$
Table B-2: Q-matrix for simulated data with $K = 6$, $J = 40$

Appendix C: Analysis of Fraction Subtraction Data Using RCI and l_z Indices

In addition to the R index, RCI and l_z were applied to the fraction subtraction data. Results flagged 47 and 35 examinees (8.8% and 6.5%) as aberrant, respectively. Notably, while R identified only 23 aberrant cases, 1 of these was not flagged by l_z but all were flagged by RCI, suggesting R is more "conservative" in flagging—a desirable property for high-stakes testing where decisions require careful consideration and multiple analytical methods. Examinee 137, flagged by l_z but not by R, showed pattern [00001011111] with estimated mastery [111], exhibiting "warm-up" or "sleeping" behavior, corroborating l_z' s dependence on item quality for detecting sleeping in short tests.

Comparing RCI and l_z results, 28 examinees were flagged by both indices, representing 80% of RCI-flagged and 60% of l_z-flagged cases. This proportion confirms l_z' s relatively lenient flagging criteria, corresponding to its slightly inflated Type I error rate observed in Study 1.

Note: Figure translations are in progress. See original paper for figures.