Cognitive Diagnostic Assessment Based on Signal Detection Theory: Development and Application

Authors: Guo Lei, Qin Haijiang, Guo Lei

Date: 2023-11-13T00:00:00+00:00

Abstract

Responding to multiple-choice items can be conceptualized as a process of extracting signals from noise. This study proposes a cognitive diagnostic model based on signal detection theory (SDT-CDM). The novel model offers several advantages: (1) It eliminates the need for attribute-level coding of options. (2) It can obtain item discrimination and difficulty parameters unavailable from traditional diagnostic models. (3) It can directly express differences in plausibility among options, enabling a more nuanced and comprehensive characterization of item performance. Results from two simulation studies indicate that: (1) The EM algorithm can effectively and conveniently implement the parameter estimation process for the new model. (2) SDT-CDM exhibits good performance, demonstrating high classification accuracy and parameter estimation precision, while additionally providing option-level estimation information for item quality diagnosis and revision. (3) Factors such as the number of attributes, item quality, and sample size influence SDT-CDM's performance. (4) Compared to the nominal response diagnostic model (NRDM), SDT-CDM achieves higher classification accuracy for examinees across all experimental conditions. Empirical research demonstrates that SDT-CDM yields better model-data fit than NRDM, with higher classification accuracy and consistency, particularly exhibiting strong stability when the number of attribute assessments is limited. Furthermore, the difficulty and discrimination parameters show higher correlations with IRT model estimates, suggesting that the model merits broader adoption.

Full Text

Cognitive Diagnostic Assessment Based on Signal Detection Theory: Modeling and Application

Guo Lei[1],[2]; **QIN Haijiang**[1],[3]
([1] Faculty of Psychology, Southwest University, Chongqing, China)
([2] Southwest University Branch, Collaborative Innovation Center of Assessment

toward Basic Education Quality, Chongqing, China)
([3] Guiyang No.37 Middle School, Guiyang, China)

Abstract

Responding to multiple-choice items can be viewed as a process of extracting signals from noise. This study proposes a cognitive diagnostic model based on signal detection theory (SDT-CDM). The new model offers three key advantages: (1) it eliminates the need for attribute-level coding of options; (2) it provides item discrimination and difficulty parameters that traditional diagnostic models cannot offer; and (3) it directly expresses the relative plausibility differences among options, enabling more nuanced and comprehensive characterization of item performance. Two simulation studies demonstrate that: (1) the EM algorithm can effectively estimate the new model's parameters in a convenient and efficient manner; (2) SDT-CDM exhibits good performance with high classification accuracy and parameter estimation precision, while additionally providing option-level information for item quality diagnosis and revision; (3) factors such as the number of attributes, item quality, and sample size affect SDT-CDM's performance; and (4) compared to the nominal response diagnostic model (NRDM), SDT-CDM achieves higher classification accuracy across all experimental conditions. Empirical research shows that SDT-CDM demonstrates better model-data fit than NRDM, with higher classification accuracy and consistency—particularly showing strong stability when attributes are assessed infrequently. Additionally, its difficulty and discrimination parameters correlate more highly with IRT model estimates, making it worthy of broader application.

Keywords: signal detection theory, cognitive diagnosis, multiple-choice items, EM algorithm

1 Introduction

Since Kelly (1916) first introduced the multiple-choice (MC) test format, it has remained one of the most popular item types due to its objectivity, efficiency, and convenience, and continues to be widely used in standardized assessments such as TIMSS, PISA, NAEP, and TOEFL. MC items offer numerous advantages: they are free from subjective scoring errors, enhance test reliability, facilitate rapid scoring, and enable content balancing (Guo & Zhou, 2021). Traditionally, MC response data are treated as dichotomously scored (correct/incorrect), which results in the loss of information embedded in distractors. To fully exploit the diagnostic information in distractors and improve the precision of individual knowledge state classification, researchers have proposed various methods, including the multiple-choice DINA model (MC-DINA; de la Torre, 2009) and its extended structured version (Ozaki, 2015), the Scaling Individuals and Classifying Misconceptions Model (SICM; Bradshaw & Templin, 2014), the Generalized Diagnostic Classification Models for Multiple Choice Option-Based Scoring

(GDCM-MC; DiBello et al., 2015), and nonparametric cognitive diagnostic methods at the option level (Guo et al., 2021; Wang et al., 2023). These approaches aim to classify examinees within a knowledge state space to identify their mastery of academic knowledge or cognitive attributes—a methodology known as cognitive diagnostic assessment. However, these MC processing methods share a common prerequisite: they require coding of distractors to represent latent categories distinct from the correct option. Although early research required distractor coding to be a subset of the correct option's coding with hierarchical relationships among distractors (Guo et al., 2013), recent work has relaxed this constraint, allowing distractor coding to be non-nested within the correct option's coding (Wang et al., 2023), thereby advancing the field.

In essence, MC testing can also be conceptualized as a signal detection task in which examinees must select the signal (correct answer) from a series of noises (all options). During the response process, examinees exist in one of two possible states: they either "know" the answer or they "do not know." From the perspective of signal detection theory (SDT), the response process comprises two stages: (1) a perception stage where examinees develop varying degrees of plausibility judgments for each option after comprehending the item, represented by plausibility parameters that follow certain distributions; and (2) a decision stage where examinees select the option they perceive as most plausible.

Based on this framework, DeCarlo (2021) integrated SDT with item response theory (IRT) for MC item analysis, enabling the estimation of relative plausibility parameters for each option as well as item discrimination and difficulty parameters. Research indicates that SDT-derived difficulty parameters are largely consistent with those from two- and three-parameter IRT models, though discrimination parameters correlate highly only with the two-parameter model (correlation with the three-parameter model is as low as 0.04). Moreover, SDT provides richer information, such as examinees' plausibility tendencies toward each option—particularly distractors—and the perceived plausibility differences among options. Consequently, SDT offers more nuanced item analysis, revealing overall item characteristics at the option level. Its value lies in: (1) enabling targeted option-level revisions to increase item difficulty when items are too easy, based on estimated option plausibility parameters; and (2) diagnosing problematic items. When examinees who "know" the answer still show greater tendency to select distractors than the correct option, this signals item quality issues—capabilities beyond the reach of two- and three-parameter models. Additionally, SDT provides a more parsimonious and interpretable analysis of MC items than the nominal response model (NRM; Bock, 1972). While NRM can also analyze option-level data, it introduces multiple discrimination parameters that complicate parameter estimation and result interpretation. Incorporating guessing behavior into NRM requires additional parameters, increasing model complexity and estimation difficulty (Thissen & Steinberg, 1997). In contrast, SDT can characterize guessing behavior without extra parameters, offering greater simplicity. Empirical research by DeCarlo (2021) further demonstrates that SDT

achieves better model-data fit than NRM.

Although Templin et al. (2008) extended NRM to the nominal response diagnostic model (NRDM) for cognitive diagnostic analysis, and Ma and de la Torre (2016) proposed the sequential G-DINA framework that includes NRDM for handling ordered and nominal data, these models retain NRM's limitations, such as excessive item parameters—requiring estimation of intercepts, main effects, and interaction terms for each option. Therefore, applying SDT to analyze option-level diagnostic data and exploring its practical value is significant. SDT offers several advantages for cognitive diagnostic assessment: (1) it eliminates the need for option-level coding, saving substantial resources; (2) it provides more parsimonious model expression with better interpretability than NRDM while delivering option-level analysis; (3) its simpler structure may improve model-data fit; and (4) it provides difficulty and discrimination parameters unavailable in traditional diagnostic models. In summary, an SDT-based approach to MC cognitive diagnostic assessment offers numerous benefits. This paper explores the methodology and techniques of SDT-based MC cognitive diagnostic assessment, develops the SDT-CDM, derives its parameter estimation methods, and evaluates the new model's performance and validity through simulation and empirical studies. The paper is structured as follows: first, we introduce the logical background of SDT models; second, we elaborate on the development of the SDT cognitive diagnostic model (SDT-CDM) and its parameter estimation; third, we examine SDT-CDM's performance through simulation and empirical studies; and finally, we discuss the results and future directions.

2 Introduction to the SDT Model

When responding to MC items, examinees first develop perceptions of each option and then convert these perceptions into plausibility tendencies regarding which option is correct. To model this process, we can assume that examinees' plausibility tendencies for each option follow a probability distribution, as illustrated in Figure 1 [Figure 1: see original paper].

Figure 1 presents the SDT response process. Consider a four-option MC item with options A, B, C, and D, where B is the correct answer. On one hand, if an examinee does not know the answer, they respond based on their perceived plausibility of each option, with more plausible options having distributions further to the right (solid lines in Figure 1). Following this perception stage is the decision process: examinees select the option they perceive as most plausible. In this example, the perceived plausibility ordering is C > B > D > A, making option C the most likely choice as it is positioned furthest right. To enable parameter estimation, one option's plausibility parameter must be fixed as a reference group, typically the last option (D) is fixed at 0. Thus, relative differences in option plausibility are represented by distances from "0," with parameters for options A, B, and C labeled as b_1, b_2, and b_3, respectively—analogous to threshold parameters in polytomous models. On the other hand, if an examinee "knows" the answer, their perceived plausibility for the correct

option B becomes strongest, shifting B's distribution to the rightmost position (dashed line in Figure 1). Since B is now furthest right, the examinee will select it.

As shown in Figure 1, the distance between distributions B and B (denoted d) represents the difference in selecting the correct option between "knowing" and "not knowing" states—this is the item discrimination parameter d, analogous to the discrimination parameter a in IRT. Clearly, larger d values indicate higher item discrimination. A negative d suggests item problems: examinees who "do not know" are more likely to answer correctly than those who "know," warranting item revision or removal. Additionally, DeCarlo (2021) defined two easiness parameters based on "not knowing" and "knowing" states: e_{DK} (easiness don't-know) and e_K (easiness know). Both represent the difference between the perceived plausibility of the correct option and the highest remaining plausibility. Specifically, in Figure 1: (1) for "not knowing" states, plausibility values for options A, B, C, and D are b_1, b_2, b_3, and b_4, respectively, with easiness parameter $e_{DK} = b_2 - b_3$; (2) for "knowing" states, plausibility values are b_1, $b_2 + d$, b_3, and b_4, with easiness parameter $e_K = b_2 + d - b_3 = e_{DK} + d$.

In SDT, discrimination d measures item quality, with $d = e_K - e_{DK}$. For e_{DK}, negative values indicate that examinees who "do not know" have lower probability of selecting the correct option and are more likely to choose distractors—consistent with test logic. Positive e_{DK} values violate test logic. Conversely, for e_K, positive values indicate that examinees who "know" have higher probability of selecting the correct option—also consistent with test logic, while negative values indicate problems. Therefore, when e_{DK} is positive and large, or e_K is negative and small, item quality is compromised and revision or removal should be considered.

This understanding of parameters d, e_{DK}, and e_K highlights SDT's advantages in evaluating item quality and guiding revisions. In practice, MC item quality is not guaranteed; even large-scale assessments exhibit high guessing probabilities. In DeCarlo's (2021) analysis of 32 SAT items, 17 had positive e_{DK} values and 2 had negative e_K values. SDT enables efficient screening of problematic items and guides targeted revisions, offering valuable capabilities beyond traditional models. Moreover, SDT provides a more parsimonious and interpretable analysis of MC items than the nominal response model (NRM; Bock, 1972). While NRM can also analyze option-level data, it introduces multiple discrimination parameters that complicate parameter estimation and result interpretation. Incorporating guessing behavior into NRM requires additional parameters, increasing model complexity and estimation difficulty (Thissen & Steinberg, 1997). In contrast, SDT can characterize guessing behavior without extra parameters, offering greater simplicity. Empirical research by DeCarlo (2021) further demonstrates that SDT achieves better model-data fit than NRM.

Based on this theoretical foundation, the SDT model is essentially a mixture model, as shown in equation (1) (detailed derivation in DeCarlo, 2021):

$$P_{jm}(b_{jm}, d_j, X_{jm}, \lambda_i) = \lambda_i \frac{e^{b_{jm}+d_j X_{jm}}}{\sum_{h=1}^{M} e^{b_{jh}+d_j X_{jh}}} + (1-\lambda_i)\frac{e^{b_{jm}}}{\sum_{h=1}^{M} e^{b_{jh}}}$$

where P_{jm} represents the probability of examinee i selecting option m ($m = 1, \ldots, M$) on item j, M is the total number of options, λ_i is a mixing parameter representing the probability that examinee i knows the answer (ranging 0-1), b_{jm}/b_{jh} are plausibility parameters for option m/h on item j, d_j is item j's discrimination, and X_{jm}/X_{jh} are indicator functions (1 if the option is correct, 0 otherwise). The first term represents the probability of selecting option m when knowing the answer, while the second term represents the probability when not knowing.

3 Construction and Parameter Estimation of SDT-CDM

To apply SDT to cognitive diagnostic assessment and develop the SDT-CDM, three requirements must be met: (1) the model must represent examinees' knowledge states for diagnostic classification; (2) interactions between knowledge states and item q-vectors must be reflected, with different knowledge states differentially affecting the probability of knowing the answer to enable identification; and (3) the model must be identifiable with parameters estimable via common algorithms such as EM or MCMC. Based on these requirements, we propose SDT-CDM, as shown in equation (2):

$$P_{jm}(b_{jm}, d_j, X_{jm}, \alpha_l) = \lambda_{lj} \frac{e^{b_{jm}+d_j X_{jm}}}{\sum_{h=1}^{M} e^{b_{jh}+d_j X_{jh}}} + (1-\lambda_{lj})\frac{e^{b_{jm}}}{\sum_{h=1}^{M} e^{b_{jh}}}$$

where α_l represents knowledge state l ($l = 1, 2, \ldots, 2^K$), K is the number of attributes, and λ_{lj} represents the probability that examinees with knowledge state α_l know the answer to item j. Unlike equation (1), SDT-CDM's advantage lies in its ability to characterize interactions between examinees and different q-vector item types, while relaxing the strong assumption of traditional SDT models that only reflect overall examinee ability (λ_i) rather than examinee-item interactions, making the model more flexible. Other symbols remain the same as in equation (1). The mixing parameter λ_{lj} is calculated as:

$$\lambda_{lj} = \frac{\sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k=1}^{K_j^*-1}\sum_{k'=k+1}^{K_j^*} \delta_{jk'k}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk}}{\sum_{k=1}^{K_j^*} \delta_{jk}q_{jk} + \sum_{k=1}^{K_j^*-1}\sum_{k'=k+1}^{K_j^*} \delta_{jk'k}q_{jk}q_{jk'} + \cdots + \delta_{j12\ldots K_j^*}\prod_{k=1}^{K_j^*} q_{jk}}$$

Although the numerator structure resembles the G-DINA model (de la Torre, 2011), the parameter meanings are fundamentally different. First, λ_{lj} calculation lacks an intercept term, meaning $\lambda_{0j} = 0$ when examinees have not

mastered any attributes required by the item, consistent with SDT's conceptualization. While δ_{jk} can be viewed as the main effect of attribute k for item j, its meaning is the contribution to the probability of "knowing" the item when the attribute is mastered—not the contribution to the probability of correct response, which is the essential difference between SDT-CDM and G-DINA. $\delta_{jk'k}$ represents second-order interactions, and $\delta_{j12\ldots K_j^*}$ represents the highest-order interaction, with similar interpretations as main effects. K_j^* is the number of attributes measured by item j. The denominator represents the sum of effects for all attributes measured by item j, while the numerator represents the sum of effects for attributes the examinee has mastered. Therefore, the more attributes mastered, the higher the probability of "knowing" the item, with different knowledge states yielding different probabilities of knowing the same item—another advantage of SDT-CDM over traditional SDT models.

Clearly, when examinees have not mastered any attributes required by item j, $\lambda_{0j} = 0$, and when all attributes are mastered, $\lambda_{1j} = 1$. Thus, $\lambda_{lj} = \frac{\text{sum of effects for mastered attributes}}{\text{sum of effects for all required attributes}}$, and equation (3) can be rewritten as:

$$\lambda_{lj} = \frac{\sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k=1}^{K_j^*-1}\sum_{k'=k+1}^{K_j^*} \delta_{jk'k}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk}}{\sum_{k=1}^{K_j^*} \delta_{jk} + \sum_{k=1}^{K_j^*-1}\sum_{k'=k+1}^{K_j^*} \delta_{jk'k} + \cdots + \delta_{j12\ldots K_j^*}}$$

SDT-CDM parameters can be estimated using the MMLE/EM algorithm. Detailed algorithm derivation and standard error calculations are provided in the online appendix.

4 Simulation Study 1

4.1 Research Purpose

Monte Carlo simulation was employed to examine SDT-CDM's classification accuracy and parameter estimation precision across different experimental conditions.

4.2 Experimental Design

This study employed a five-factor fully crossed design with the following independent variables: number of attributes ($K = 3, 5$), test length ($J = 20, 40$), item quality (high vs. low), sample size ($N = 1000, 2000$), and attribute distribution (higher-order vs. multivariate normal). All conditions were replicated 200 times to reduce random error.

4.2.1 Item Simulation Q-matrices were generated by randomly sampling q-vectors from all possible vectors while ensuring two identity matrices were included, enabling identifiability of knowledge states (Xu, 2017; Fang et al.,

2019) and random Q-matrix simulation. Since no prior research provided reference ranges for SDT-CDM parameters, we followed previous cognitive diagnostic studies (Guo et al., 2016). High-quality item parameters were generated with e_{DK} drawn from $U[-2.5,-1]$ and e_K from $U[2.5,3.5]$. When $e_{DK}=-2.5$ and $e_K=3.5$, $(1-P_1)$ and $P_0 \cong 0.05$; when $e_{DK}=-1$ and $e_K=2.5$, $(1-P_1)$ and $P_0 \cong 0.15$, equivalent to $(1-P_1)$ and P_0 drawn from $U[0.05,0.15]$ in traditional CDMs. Low-quality item parameters used e_{DK} from $U[-1,-0.5]$ and e_K from $U[1.8,2.5]$. When $e_{DK}=-1$ and $e_K=2.5$, $(1-P_1)$ and $P_0 \cong 0.15$; when $e_{DK}=-0.5$ and $e_K=1.8$, $(1-P_1)$ and $P_0 \cong 0.25$, equivalent to $(1-P_1)$ and P_0 drawn from $U[0.15,0.25]$. To maximize randomness and generalizability, option plausibility parameters b_{jm} and attribute effects δ were not strictly constrained. Since b_{jm} only affects option selection probabilities through relative magnitude (as shown in Figure 1), they were drawn from standard normal distributions to create random plausibility relationships. Attribute effects δ were constrained only to satisfy the assumption that more mastered attributes yield higher probabilities of "knowing" the item. Additionally, all MC items were fixed at four options, consistent with most real-world MC items.

4.2.2 Examinee Simulation Examinee knowledge states were generated using higher-order and multivariate normal distributions. The higher-order distribution followed Ma et al. (2016):

$$P_{ik}=\frac{\exp[1.7\times(\theta_i-\delta_k)]}{1+\exp[1.7\times(\theta_i-\delta_k)]}$$

where θ_i is examinee ability drawn from a standard normal distribution, and δ_k is the difficulty of mastering attribute k, equally spaced from -1 to 1 across attributes (e.g., for three attributes: $\delta_1=-1$, $\delta_2=0$, $\delta_3=1$).

The multivariate normal distribution followed Chiu (2013), defining a K-dimensional vector $\theta_i=(\theta_{i1},\theta_{i2},\ldots,\theta_{iK})$ representing continuous ability values for each attribute, drawn from $MVN(\mathbf{0},\ )$ with covariance matrix off-diagonal elements set to 0.5 to describe attribute correlations. True knowledge states were generated as:

$$\alpha_{ik}=\begin{cases}1, & \text{if } \theta_{ik}>\Phi^{-1}\left(\frac{1+K}{2K}\right)\\ 0, & \text{otherwise}\end{cases}$$

4.3 Evaluation Metrics

Parameter estimation precision was evaluated using average bias and root mean squared error (RMSE), calculated as:

$$\text{Bias}=\frac{\sum_{r=1}^{R}(\omega-\widehat{\omega}_r)}{R}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{r=1}^{R}(\omega - \widehat{\omega}_r)^2}{R}}$$

where ω represents the true parameter value, $\widehat{\omega}_r$ the estimated value in replication r, and R the total number of replications. Bias approaching zero indicates minimal estimation bias, while smaller RMSE indicates better estimation accuracy.

Classification accuracy of attribute mastery was evaluated using average attribute correct classification rate (AACCR) and pattern correct classification rate (PCCR):

$$\text{AACCR} = \frac{\sum_{k=1}^{K}\text{ACCR}_k}{K}$$

$$\text{PCCR} = \frac{\sum_{i=1}^{N} p_{ir}}{N}$$

where $\text{ACCR}_k = \frac{\sum_{i=1}^{N}\alpha_{mikr}}{N}$ and $\alpha_{mikr} = 1$ indicates correct classification of examinee i's attribute k in replication r, while $p_{ir} = 1$ indicates correct classification of examinee i's knowledge state in replication r.

4.4 Results

Figures 2 [Figure 2: see original paper] and 3 [Figure 3: see original paper] present overall parameter estimation bias and RMSE for SDT-CDM across experimental conditions. For brevity, results are shown as means across multiple parameters (plausibility, main effects, interaction effects). Overall, parameter estimation precision was high: plausibility parameters showed bias ranging from -0.003 to 0.007 (mean = 0.002) and RMSE from 0.119 to 0.261 (mean = 0.173); discrimination parameters showed bias from -0.054 to -0.001 (mean = -0.022) and RMSE from 0.145 to 0.385 (mean = 0.253); easiness parameter e_K showed bias from -0.014 to 0.075 (mean = 0.027) and RMSE from 0.181 to 0.334 (mean = 0.260).

Different factors affected parameter estimation precision differently. First, higher-order attribute distributions yielded slightly better precision than multivariate normal distributions: mean bias (RMSE) for parameters b, d, e_{DK}, e_K, δ-M, and δ-I were 0.002 (0.160), -0.022 (0.234), 0.046 (0.245), 0.025 (0.248), 0.001 (0.078), and -0.001 (0.154) under higher-order distributions, compared to 0.002 (0.187), -0.022 (0.271), 0.051 (0.267), 0.029 (0.271), 0.008 (0.126), and -0.009 (0.236) under multivariate normal distributions. Second, more attributes slightly decreased precision: when K increased from 3 to 5, mean bias changed from 0.009 to 0.010, but mean RMSE increased from 0.189

to 0.224 (18.5% increase). However, test length had minimal impact: when J increased from 20 to 40, mean bias changed from 0.008 to 0.010 and mean RMSE from 0.203 to 0.210. Third, item quality had a substantial effect: when quality decreased from high to low, mean bias increased from 0.000 to 0.019 and mean RMSE from 0.192 to 0.221 (15.1% increase). Finally, sample size had the greatest impact: when N decreased from 2000 to 1000, mean bias increased from 0.007 to 0.010 and mean RMSE from 0.179 to 0.234 (30.7% increase).

Figure 4 [Figure 4: see original paper] presents SDT-CDM's AACCR and PCCR results. Overall, the new model accurately classified examinees, with classification precision affected by experimental factors. Among the five factors, item quality had the largest impact: under low quality, AACCR ranged from 0.902 to 0.988 (mean = 0.951) and PCCR from 0.609 to 0.964 (mean = 0.816); under high quality, AACCR ranged from 0.973 to 1.000 (mean = 0.990) and PCCR from 0.876 to 0.999 (mean = 0.957), representing a 17.4% improvement in PCCR. The number of attributes was the second most influential factor: when $K = 3$, mean AACCR and PCCR were 0.983 and 0.951, respectively; when $K = 5$, AACCR decreased by 2.5% while PCCR decreased by 15.7%. Test length was the third factor: with $J = 20$, mean AACCR and PCCR were 0.958 and 0.841; with $J = 40$, they improved to 0.984 and 0.932 (2.7% and 10.8% increases, respectively). Attribute distribution and sample size had minimal effects: mean AACCR and PCCR were 0.969 and 0.882 under higher-order distributions vs. 0.972 and 0.891 under multivariate normal distributions; with $N = 1000$, mean AACCR and PCCR were 0.970 and 0.883 vs. 0.972 and 0.890 with $N = 2000$.

5 Simulation Study 2

5.1 Research Purpose

Monte Carlo simulation was used to compare classification accuracy between SDT-CDM and NRDM across experimental conditions. The NRDM is expressed as:

$$P_{jm}(Y_j = m_j|\alpha_l) = \frac{\exp[\gamma_{0,j,m_j} + \gamma_{j,m_j}^T \mathbf{h}(\alpha_l, \mathbf{q}_j)]}{\sum_{m_j \in M_j} \exp[\gamma_{0,j,m_j} + \gamma_{j,m_j}^T \mathbf{h}(\alpha_l, \mathbf{q}_j)]}$$

where $\gamma_{j,m_j}^T \mathbf{h}(\alpha_l, \mathbf{q}_j) = \sum_{k=1}^{K_j^*} \gamma_{1,j,k,m_j}(\alpha_{lk} q_{jk}) + \sum_{k=1}^{K_j^*-1} \sum_{c=k+1}^{K_j^*} \gamma_{2,j,k,c,m_j}(\alpha_{lk}\alpha_{lc} q_{jk} q_{jc}) + \cdots$. Here γ_{0,j,m_j} are intercepts, γ_{1,j,k,m_j} are main effects, γ_{2,j,k,c,m_j} are second-order interactions, etc. See Templin et al. (2008) for NRDM details. Comparing equations (2) and (12) reveals that the models' item parameters have different meanings and are not interchangeable: SDT-CDM uses option plausibility parameters b_{jm}, while NRDM uses main effects and interactions. Additionally, parameter ranges differ—SDT-CDM parameters can exceed 1, while NRDM

parameters are bounded 0-1—making direct precision comparisons unfair. Therefore, we focused on classification accuracy differences.

5.2 Experimental Design

Independent variables matched Study 1. Both SDT-CDM and NRDM served as true models to generate data, with both models then fit to each dataset. NRDM item quality parameters were set as: high-quality items had $1-P(1)$ and $P(0)$ drawn from $U[0.05, 0.15]$; low-quality items had $1-P(1)$ and $P(0)$ drawn from $U[0.15, 0.25]$, where $P(1)$ and $P(0)$ are correct response probabilities for fully mastered and non-mastered knowledge states, respectively. Other settings matched Study 1.

5.3 Results

Online Appendix Figures A1 and A2 present PCCR and AACCR results when each model served as the true model. Regardless of the true model, SDT-CDM outperformed NRDM. When SDT-CDM was the true model, attribute distribution minimally affected both models, while sample size moderately affected only NRDM (PCCR increased 7.6% with larger samples). Increasing attributes from 3 to 5 decreased mean PCCR by 12.9% for SDT-CDM and 10.3% for NRDM. Decreasing item quality reduced mean PCCR by 14.3% for SDT-CDM and 29.4% for NRDM. Notably, test length affected the models differently: increasing items improved SDT-CDM's mean PCCR by 9.2% but decreased NRDM's by 18.2%. A likely explanation is that more items substantially increase NRDM's parameter count (see equation (12)), requiring larger samples for accurate estimation. When sample size is insufficient, parameter estimation suffers, reducing classification accuracy. This is supported by Templin et al. (2008), where even a reduced compensatory NRDM required 5,000 examinees for adequate precision. In contrast, SDT-CDM shows the expected pattern: more items yield higher classification accuracy, demonstrating its suitability for diagnostic testing with nominal data. This also explains why NRDM performed poorly even when it was the true model. When NRDM was the true model, factor effects were similar but differences between models were smaller: decreasing item quality reduced mean PCCR by 6.2% for SDT-CDM vs. 14.8% for NRDM, indicating SDT-CDM's greater stability.

Online Appendix Table A1 further shows factor effects on model differences. Regardless of the true model, test length had the greatest impact: with $J = 20$, performance was similar, but with $J = 40$, SDT-CDM's mean PCCR was 42.29% and 21.04% higher than NRDM's under the two true-model conditions, respectively, suggesting NRDM is unsuitable for longer tests without large samples, while SDT-CDM can handle them with moderate samples. Item quality was the second most important factor: under low quality, SDT-CDM's mean PCCR was 36.06% and 16.52% higher than NRDM's. Sample size was third: with small samples, SDT-CDM's mean PCCR was 24.72% and 14.93% higher, indicating better small-sample performance. Other factors showed varying degrees

of influence.

Overall, SDT-CDM outperformed NRDM across all conditions, demonstrating broader applicability and greater stability.

6 Empirical Study

Empirical data came from TIMSS 2011 mathematics assessment data previously used by Ma and de la Torre (2020), containing 23 mathematics items. We analyzed 14 MC items with responses from 748 U.S. examinees. Missing values were replaced with random incorrect responses. The Q-matrix contained six attributes: A1) whole numbers; A2) fractions, decimals, and ratios; A3) expressions, equations, and functions; A4) lines, angles, and shapes; A5) location and movement; and A6) data organization, representation, and interpretation (Table 1). Diagnostic reliability and validity were evaluated using attribute-level and pattern-level classification consistency and accuracy indices proposed by Wang et al. (2015), where higher values indicate better reliability and validity. NRDM was included for comparison.

Table 2 presents relative model-data fit indices: -2 log likelihood, AIC, and BIC, where smaller values indicate better fit. Results show SDT-CDM outperformed NRDM on all three indices (bolded results), with 71 freely estimated parameters compared to NRDM's 87, making it more parsimonious.

Online Appendix Tables A2 and A3 present parameter estimates. Table A2 shows all 14 items had positive discrimination d, indicating proper differentiation between "knowing" and "not knowing" states. Theoretically, larger d indicates better item quality, but DeCarlo (2021) found that excessively large d values (above 6) could inflate standard errors. In this study, only item 7 had $d > 6$ (6.244) with a standard error of 4.044—much smaller than 8, indicating stable estimation.

Theoretically, high-quality items should have negative e_{DK} values (smaller is better), meaning the most plausible option among b_1 to b_4 should be a distractor, not the correct answer; otherwise, "not knowing" examinees would have high guessing probability. Nine items had negative e_{DK}, indicating low guessing probability. However, five items had positive e_{DK}, suggesting the correct option appeared more plausible to "not knowing" examinees. For example, item 11 had $e_{DK} = b_{\text{correct}} - b_{\text{largest distractor}} = 2.227 - 1.391 = 0.836$, meaning "not knowing" examinees had a 50.9% probability of selecting the correct answer ($\frac{e^{2.227}}{e^{1.369}+e^{1.391}+e^{2.227}+e^{0}} = 0.509$), indicating a highly guessable item. This aligns closely with NRDM results (Table 5), which showed $P(0) = 0.498$ for non-masters, confirming that SDT-CDM's e_{DK} accurately identifies excessive guessing.

Similarly, high-quality items should have positive e_K values (larger is better), meaning the correct option should be most plausible after accounting for discrimination. For item 11, $e_K = (b_{\text{correct}} + d) - b_{\text{largest distractor}} = e_{DK} + d =$

$0.836 + 3.454 = 4.290$, indicating masters had a 98.7% probability of selecting correctly ($\frac{e^{2.227+4.290}}{e^{1.369}+e^{1.391}+e^{2.227+4.290}+e^{0}} = 0.987$). However, item 12 had the smallest e_K (0.41), giving masters only a 35.8% correct response probability, indicating high slipping probability and need for revision.

These analyses demonstrate SDT-CDM's utility for guiding item and option revision. Among the 14 items, five showed excessive guessing (positive e_{DK}), suggesting distractors need more 吸引力. While no items showed logical anomalies, items 12, 6, and 13 had $e_K < 1$ and low discrimination ($d = 0.971, 0.884, 1.123$), indicating opportunities to increase the plausibility difference between "knowing" and "not knowing" states.

Online Appendix Table A4 shows SDT-CDM's attribute main effects and interaction estimates. For item 1, $\delta_1 = 0.999$ indicates that mastering only the first required attribute yields a 99.9% probability of "knowing" the item, while mastering only the second yields 66.5%. Mastering both increases this probability by 0.1% and 33.5% relative to the single-attribute cases.

Table 3 presents classification accuracy and consistency indices. In classification accuracy, SDT-CDM outperformed NRDM on all attributes except A1, with a 39.13% improvement in pattern accuracy and 23.77% improvement for A6. In classification consistency, SDT-CDM again outperformed NRDM except on A1, with a 28.63% improvement for A6. Since A6 was assessed only once (see Q-matrix in Table 2), NRDM was more affected, while SDT-CDM maintained high accuracy and consistency even with limited assessments, demonstrating robustness.

SDT-CDM provides difficulty and discrimination parameters unavailable in traditional CDMs. Correlations with two- (2PL) and three-parameter (3PL) IRT models were: $r(-e_{DK}, \beta_{2PL}) = 0.63^{*}$, $r(-e_{DK}, \beta_{3PL}) = 0.71^{**}$, $r(-e_K, \beta_{2PL}) = 0.89^{***}$, and $r(-e_K, \beta_{3PL}) = 0.79^{***}$. Following Cohen's (1988) criteria ($r \geq 0.5$ indicates large effect) and Zhang and Xu's (2015) guidelines ($0.6-0.8$ = strong, > 0.8 = very strong), all correlations were significant and substantial, confirming SDT-CDM's ability to characterize item difficulty like IRT models. Since NRDM cannot provide difficulty parameters, we compared discrimination parameters: $r(d, a_{2PL}) = 0.66^{**}$, $r(d, a_{3PL}) = 0.79^{***}$, $r(\text{GDI}, a_{2PL}) = 0.20^{ns}$, $r(\text{GDI}, a_{3PL}) = 0.15^{ns}$. SDT-CDM's discrimination d correlated strongly and significantly with IRT models, while NRDM's generalized discrimination index (GDI) showed low, non-significant correlations.

SDT-CDM identified knowledge states for 748 examinees from 64 possible states. Figure 5 [Figure 5: see original paper] shows the top 10 most frequent knowledge states, accounting for 79.3% of examinees. Correlations between attribute mastery and total scores were 0.87 for SDT-CDM and 0.76 for NRDM (both $p < .001$), further supporting the new model's superiority.

7 Discussion and Outlook

The MC response process can be viewed as signal detection, where examinees develop plausibility perceptions for each option and always select the most plausible. This study integrated SDT into CDMs, yielding several key findings. First, SDT-CDM eliminates the need for option-level coding, instead assigning plausibility parameters to each option to characterize option differences. These parameters can be combined to compute difficulty and discrimination parameters unavailable in traditional diagnostic models, enabling item quality diagnosis and revision. Simulation and empirical studies confirm these advantages and demonstrate successful model development. Second, two comprehensive simulation studies examining five factors found: (1) item quality and sample size substantially affected parameter estimation precision, while attribute distribution, number of attributes, and test length had smaller effects; (2) item quality, number of attributes, and test length substantially affected classification accuracy, while attribute distribution and sample size had minimal effects; (3) model comparisons showed SDT-CDM outperformed NRDM regardless of the true model, attributable to NRDM's large sample requirements, thereby proving SDT-CDM's practical applicability and robustness. Third, TIMSS 2011 empirical analysis showed SDT-CDM achieved better model fit, higher classification accuracy and consistency (especially with infrequently assessed attributes), stronger correlations with IRT difficulty and discrimination parameters, and higher correlations between attribute mastery and total scores. Furthermore, the two easiness parameters (e_{DK} and e_K) and discrimination parameter d provide targeted indicators for item quality diagnosis and revision—capabilities unavailable in NRDM. Several issues warrant further discussion.

7.1 Utilization of Distractor Information

Most current MC cognitive diagnostic research codes distractors (de la Torre, 2009; DiBello et al., 2015; Guo et al., 2021; Ozaki, 2015; Wang et al., 2023) to maximize diagnostic information from q-vectors. However, this approach increases item development difficulty, and when distractor q-vectors are similar or some options are uncodable, the additional diagnostic information becomes limited. Although SDT-CDM does not require distractor coding, it transforms traditional dichotomous scoring into nominal data processing and provides option-level parameters (b_{jm}), constituting option-level information processing. Simulation and empirical studies show SDT-CDM achieves better diagnostic classification accuracy, consistency, and model fit than NRDM. This study represents an initial introduction of SDT to CDA; future work should explore extensions that incorporate distractor information. One potential approach is refining the mixing parameter λ_{lj} to the option level to characterize interactions between knowledge states and option q-vectors, comprehensively reflecting the probability of "knowing" each option.

7.2 EM Algorithm Improvements and Standard Error Computation

This study derived the EM algorithm for SDT-CDM, but various EM variants exist (Chalmers, 2012), including standard EM with fixed quadrature, Monte Carlo EM, stochastic EM, MH-RM algorithm, and minimum chi-square EM (Zhu et al., 2006). Most have been applied in IRT research and are available in the mirt package. However, CDM applications have relied primarily on MMLE/EM since de la Torre (2009). While MMLE/EM is simple and efficient, exploring new algorithms with higher precision, faster convergence, or other advantages is worthwhile. Future research should consider introducing mature IRT algorithms to SDT-CDM.

Additionally, CDM standard errors are typically computed via the inverse information matrix, but multiple information matrix types exist (Liu, 2022), including empirical cross-product (XPD), observed (Obs), and sandwich-type (Sw) matrices. This study used XPD, but future research should explore how different information matrices affect SDT-CDM standard error estimation.

7.3 Integration with Process Data

Advances in computer technology have facilitated recording examinee response process data, leading researchers to explore how such information can improve diagnostic precision and reveal response styles or strategies. Examples include integrating response times (Zheng et al., 2023), eye-tracking data (Zhan, 2022), and action sequences (Zhan & Qiao, 2022) into CDMs, demonstrating feasibility and effectiveness for multimodal data analysis. While mining process data is accepted, consensus on optimal analysis methods is lacking (He et al., 2021), and various models exist for different data types—e.g., Poisson, negative binomial, zero-inflated, and hurdle models for count data; latent space models (Chen et al., 2022), RNN-based autoencoders (Tang et al., 2021), and multidimensional scaling (Tang et al., 2020) for action sequences. Different feature extraction methods affect diagnostic classification. Future research should examine the effectiveness of various process data models and feature extraction methods when combined with SDT-CDM.

7.4 Integration with Longitudinal Diagnosis

Longitudinal cognitive diagnostic research is a recent hotspot, enabling characterization of learning trajectories and targeted remedial instruction. Current longitudinal CDMs include latent transition analysis-based models (Wang et al., 2018; Zhang & Chang, 2020) and higher-order latent structure models (Lee, 2017; Zhan et al., 2019). Future work should integrate SDT into longitudinal CDMs to track knowledge states and monitor item quality changes over time.

This study has limitations. First, SDT-CDM was only compared to NRDM, limiting deeper exploration due to the scarcity of CDMs that handle option-level data without requiring option coding. Second, while XPD is an analytic information matrix, it can encounter non-positive definite matrices or negative

variance-covariance diagonal elements, preventing standard error computation. Liu (2022) proposed the "parallel bootstrap method" as a better alternative that avoids these issues through Monte Carlo-like computation, but its effectiveness for SDT-CDM was not explored. Third, while MMLE/EM is efficient, it may converge to local optima; Zeng et al. (2023) proposed the Tensor-EM algorithm to address this limitation, offering a promising approach for complex models.

7.5 Conclusions

This study proposed the SDT-CDM. Based on simulation and empirical results, we conclude:

1. SDT-CDM parameters can be estimated via EM algorithm. Beyond providing difficulty and discrimination parameters unavailable in traditional CDMs, it estimates option-level plausibility parameters for item revision.
2. Simulation results show good parameter estimation precision, with classification accuracy affected by experimental factors. The importance ranking for classification precision is: item quality, number of attributes, and test length, while attribute distribution and sample size have smaller effects.
3. Empirical results show SDT-CDM achieves better model-data fit, higher pattern/attribute classification accuracy and consistency (especially with infrequently assessed attributes), and higher correlations between attribute mastery and total scores than NRDM, without requiring distractor coding. The two easiness parameters (e_{DK} and e_K) and discrimination parameter d enable targeted item quality diagnosis and revision.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: a psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, 79(3), 403–425.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37(8), 598–618.

Chen, Y., Zhang, J., Yang, Y., & Lee, Y-S. (2022). Latent space model for process data. *Journal of Educational Measurement*, 59(4), 517–535.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Erlbaum.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement*, 39(1), 62–79.

DeCarlo, L. T. (2021). A signal detection model for multiple-choice exams. *Applied Psychological Measurement*, 45(6), 423–440.

de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179–199.

Fang, G., Liu, J., & Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika*, 84, 19–40.

Guo, L., Yuan, C. Y., & Bian, Y. F. (2013). Discussing the development tendency of cognitive diagnosis from the perspective of new models. *Advances in Psychological Science*, 21(12), 2256–2264. [郭磊, 苑春永, 边玉芳. (2013). 从新模型视角探讨认知诊断的发展趋势. 心理科学进展, 21(12), 2256–2264.]

Guo, L., Zheng C., Bian Y., Song N., & Xia L. (2016). New item selection methods in cognitive diagnostic computerized adaptive testing: combining item discrimination indices. *Acta Psychologica Sinica*, 48(7), 903–914. [郭磊, 郑蝉金, 边玉芳, 宋乃庆, 夏凌翔. (2016). 认知诊断计算机化自适应测验中新的选题策略: 结合项目区分度指标. 心理学报, 48(7), 903–914.]

Guo, L., & Zhou, W. J. (2021). Nonparametric methods for cognitive diagnosis to multiple-choice test items. *Acta Psychologica Sinica*, 53(9), 1032–1043. [郭磊, 周文杰. (2021). 基于选项层面的认知诊断非参数方法. 心理学报, 53(9), 1032–1043.]

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, 166: 104170.

Kelly, F. J. (1916). The kansas silent reading tests. *Journal of Educational Psychology*, 7, 63–80.

Lee, S. Y. (2017). *Growth curve cognitive diagnosis models for longitudinal assessment*. Unpublished doctoral thesis, University of California, Berkeley.

Liu, Y. (2022). Standard errors and confidence intervals for cognitive diagnostic models: parallel bootstrap methods. *Acta Psychologica Sinica*, 54(6), 703–724. [刘彦楼. (2022). 认知诊断模型的标准误与置信区间估计: 并行自助法. 心理学报, 54(6), 703–724.]

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, 69(3), 253–275.

Ma, W., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and*

Statistical Psychology, 73(1), 142–163.

Ozaki, K. (2015). DINA models for multiple-choice items with few parameters: considering incorrect answers. *Applied Psychological Measurement*, 39(6), 431–447.

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, 85, 378–397.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoder. *British Journal of Mathematical and Statistical Psychology*, 74(1), 1–33.

Templin, J., Henson, R., Rupp, A., Jang, E., & Ahmed, M. (2008). Cognitive diagnosis models for nominal response data. Annual Meeting of the National Council on Measurement in Education, New Brunswick, New Jersey.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–65). Springer.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, 52, 457–476.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics*, 43, 57–87.

Wang, Y., Chiu, C.-Y., & Kohn, H. F. (2023). Nonparametric classification method for multiple-choice items in cognitive diagnosis. *Journal of Educational and Behavioral Statistics*, 48(2), 189–219.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, 45, 675–707.

Xu, X., Chang, H., & Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Quebec, Canada.

Zeng, Z., Gu, Y. & Xu, G. (2023). A tensor-EM method for large-scale latent class analysis with binary responses. *Psychometrika*, 88, 580–612.

Zheng, T. P., Zhou, W. J., & Guo, L. (2023). Cognitive diagnosis modelling based on response times. *Journal of Psychological Science*, 46(2), 478–490. [郑天鹏, 周文杰, 郭磊. (2023). 基于题目作答时间信息的认知诊断模型. 心理科学, 46(2), 478–490.]

Zhan, P. D. (2022). Joint-cross-loading multimodal cognitive diagnostic modeling incorporating visual fixation counts. *Acta Psychologica Sinica*, 54(11),

1416–1432. [詹沛达. (2022). 引入眼动注视点的联合-交叉负载多模态认知诊断建模. 心理学报, 54(11), 1416–1432.]

Zhan, P. D., Jiao, H., Liao D. D., & Li, F. M. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics*, 44(3), 251–281.

Zhan, P. D., & Qiao, X. (2022). Diagnostic classification analysis of problem-solving competence using process data: an item expansion method. *Psychometrika*, 87(4), 1529–1547.

Zhang, H. C., & Xu, J. P. (2015). *Modern psychology and educational statistics* (4th ed.). Beijing: Beijing Normal University Press. [张厚粲, 徐建平. (2015). 现代心理与教育统计学 (第 4 版). 北京: 北京师范大学出版社.]

Zhang, S. S., & Chang, H. H. (2020). A multilevel logistic hidden markov model for learning under cognitive diagnosis. *Behavior Research Methods*, 52, 408–421.

Zhu W., Ding S., Chen X. (2006). Minimum chi-square/EM estimation under IRT. *Acta Psychologica Sinica*, 38(3), 453–460. [朱玮, 丁树良, 陈小攀. (2006). IRT 中最小化 2/EM 参数估计方法. 心理学报, 38(3), 453–460.]

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