# A Comparative Study of Item Selection Strategies for CD-CAT Based on the Nominal Response Model

**Authors:** Zhang Jie, Luo Zhaosheng, Yu Xiaofeng, Qin Chunying, Luo Zhaosheng, Yu Xiaofeng

**Date:** 2022-05-12T16:44:20+00:00

## Abstract

Currently, most research on CD-CAT has been conducted based on 0-1 scoring data, whereas in actual educational and psychological testing, there also exists a large amount of nominal response data. This study develops a CD-CAT suitable for nominal response data (hereinafter referred to as NCD-CAT) based on the Nominal Response Cognitive Diagnosis Model (NR-cRUM), and introduces seven item selection methods from 0-1 scoring CD-CAT into NCD-CAT. The effects of different item selection methods on examinee classification accuracy and test efficiency are compared under various conditions. Results indicate that PWKL-family new methods such as NR_{PWCDI} and NR_{MPWKL}, as well as the NR_{SHE}/MI method, are well-suited for NCD-CAT and outperform the baseline method NR_{PWKL} under most conditions. This study expands the item selection methods for nominal polytomous scoring CD-CAT.

## Full Text

## A Comparative Study of Item Selection Methods in CD-CAT Based on Nominal Response Models

**Zhang Jie**[1], **Luo Zhaosheng**[1], **Yu Xiaofeng**[1], **Qin Chunying**[2] [1]School of Psychology, Jiangxi Normal University, Nanchang, 330022 [2]School of Mathematics and Information Science, Nanchang Normal University, Nanchang, 330032

### Abstract

Currently, most research on CD-CAT has been conducted based on 0-1 scoring data. However, a substantial amount of nominal response data exists in practical

educational and psychological testing. This study develops a CD-CAT suitable for nominal response data (hereinafter referred to as NCD-CAT) based on the nominal response cognitive diagnosis model (NR-cRUM) and introduces seven item selection methods from 0-1 scoring CD-CAT into NCD-CAT. The effects of different selection methods on classification accuracy and test efficiency under various conditions are compared. Results indicate that PWKL-family methods such as NR_{PWCDI} and NR_{MPWKL}, along with NR_{SHE}/MI methods, are well-suited for NCD-CAT and outperform the baseline method NR_{PWKL} under most conditions. This research expands the item selection methods for nominal polytomous CD-CAT.

**Keywords:** CD-CAT; nominal responses; NR-cRUM; item selection methods; multiple-choice items

## 1. Introduction

Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT; Cheng, 2009) integrates the advantages of Cognitive Diagnostic Assessment (CDA; Leighton & Gierl, 2007; von Davier & Lee, 2019) and Computerized Adaptive Testing (CAT) (Luo et al., 2015; Yu et al., 2019). Depending on the scoring method, CD-CAT can be divided into dichotomous and polytomous CD-CAT. At present, most CD-CAT research has been based on dichotomous scoring data. Nevertheless, numerous polytomous items exist in practical educational and psychological tests (Liu et al., 2015; Wang et al., 2016). Based on whether ordering or grading exists among response categories, polytomous items can be further classified as nominal polytomous or ordinal polytomous. Nominal polytomous scoring data, commonly found in multiple-choice items (MCI) and personality or attitude scales where different tendencies are reflected without definitive correct answers, refers to data where multiple response categories are relatively independent without ordering or grading among them. Nominal polytomous data can be considered the most general and lowest measurement-level data type, with both ordinal polytomous and dichotomous data viewed as special cases of nominal polytomous data (Mellenbergh, 1995).

To analyze and extract information from nominal response data, researchers have developed corresponding nominal response models. Within the IRT framework, Bock (1972) developed the Nominal Response Model (NRM) and applied it to analyze nominal-scored multiple-choice items. Results demonstrated that NRM could utilize information from distractors, with ability estimation precision significantly higher than that of conventional dichotomous IRT models. For medium- and low-ability examinees, NRM' s ability estimation precision could achieve the same level as a dichotomous IRT model with twice the test length. Wang et al. (2017) applied NRM to construct a CAT that allows answer changes, using NRM to incorporate both initial responses and subsequent modifications into interim and final ability estimates, thereby providing more item selection information and more accurate ability estimates. This answer-changeable CAT scheme precisely leverages the characteristic that response categories in nominal

response models are nominally scored. Furthermore, Wang et al. (2019) theoretically discussed the feasibility of this NRM-based answer-changeable CAT scheme.

Within the cognitive diagnosis framework, Templin et al. (2008) and de la Torre (2009) developed nominal response cognitive diagnosis models. Based on conventional dichotomous cognitive diagnosis models, examinees can only be classified into mastery and non-mastery groups. In contrast, nominal response cognitive diagnosis models can divide examinees into more categories, thereby improving classification accuracy. Templin et al. (2008) extended the log-linear cognitive diagnosis model (LCDM) to a nominal polytomous version, developing the NR-DM (Nominal Response Diagnostic Model) and its reduced model NR-cRUM (Nominal Response Compensatory Reparameterized Unified Model; detailed introduction of this model appears in Appendix A). Compared with the dichotomous cRUM, using the nominal polytomous NR-cRUM yields higher classification accuracy for examinees. De la Torre (2009) extended the DINA model to a nominal polytomous version, with the extended MC-DINA model demonstrating significantly higher classification accuracy than the DINA model because MC-DINA utilizes diagnostic information from distractors. Guo and Zhou (2021) proposed a class of nonparametric methods for diagnosing distractor information.

After reviewing relevant domestic and international literature, we found that only Yigit et al. (2019) developed CD-CAT based on nominal response models. In that study, the authors evaluated only one item selection method' s efficiency (by comparing it with dichotomous CD-CAT) using the MC-DINA model. MC-DINA is a non-compensatory model applicable only to non-compensatory situations among attributes, and due to limited parameter quantities, the model cannot directly explain the relationship between each attribute and each response category, thereby limiting the generality of model parameter interpretation (Luo et al., 2020). Since the full NR-DM model has numerous parameters requiring large sample sizes for estimation, its reduced form NR-cRUM is more practical (Li, 2014; Templin et al., 2008). Therefore, this study adopts the NR-cRUM model to develop polytomous CD-CAT, extends seven commonly used item selection methods from conventional CD-CAT to be applicable in NR-cRUM-based diagnostic testing, and compares them across different conditions.

## 2. Methods

### 2.1 Initial Item Selection

In the initial stage of CD-CAT, researchers have proposed random selection methods or randomly assigning an attribute mastery pattern (KS) to examinees and then selecting corresponding items (Gao et al., 2017; Yu et al., 2019). The PWCDI and PWACDI selection methods proposed by Zheng and Chang (2016) can completely eliminate the problem of unstable KS estimation in the initial stage, naturally selecting items that meet the requirements of the "T-matrix

method" (Tu et al., 2013) for initial item selection. For fair comparison among selection methods, this study adopts the approach of randomly selecting one item and using this same item as the initial item for all selection methods.

### 2.2 Item Selection Methods for NCD-CAT

In dichotomous CD-CAT, researchers have proposed various item selection methods (Guo et al., 2016; Li et al., 2021; Luo et al., 2015; Cheng, 2009; Guo & Zheng, 2019; Kaplan et al., 2015; Wang, 2013; Yu et al., 2019; Zheng & Chang, 2016). Existing polytomous CD-CAT item selection methods (Gao et al., 2020) have primarily been extended from dichotomous CD-CAT. Following this approach, this study extends several effective selection methods from conventional CD-CAT to NCD-CAT.

Assume there are K attributes dividing examinees into C = 2^K latent classes, and item j has b_j + 1 options (1 correct option and b_j distractors). The NCD-CAT item selection methods involved in this study are introduced below.

**2.2.1 KL Information Matrix and Its Variants** Before introducing the selection methods, we briefly describe the KL information matrix. The KL information matrix, also known as the D-matrix (Henson & Douglas, 2005), is a 2^K × 2^K matrix (where K is the number of attributes) whose elements are the expected KL distances between two KSs given response patterns. The calculation formula is as follows:

When X is dichotomously scored, the D_{juv} calculation formula is:

$$D_{juv} = E_{\alpha u}\left[\log\left(\frac{P_{\alpha u}(X_j)}{P_{\alpha v}(X_j)}\right)\right] = \sum P_{\alpha u}(x_j)\log\left[\frac{P_{\alpha u}(x_j)}{P_{\alpha v}(x_j)}\right]$$

In NCD-CAT, X is nominally scored. Accordingly, when calculating the D-matrix (denoted as NR_D), the expectation should be taken according to the probability of each response category:

$$NR\_D_{juv} = \sum P_{\alpha u}(x_j)\log\left[\frac{P_{\alpha u}(x_j)}{P_{\alpha v}(x_j)}\right]$$

The NR_D_{juv} matrix contains information about the ability of each response category to discriminate between different KSs, thus incorporating more information than the traditional dichotomous D-matrix. The PWKL-family selection methods in this study (NR_{PWKL}, NR_{PWCDI}, NR_{PWACDI}, NR_{MPWKL}) are all based on the NR_D matrix combined with the ideas of corresponding dichotomous CD-CAT selection methods. Similar situations will not be reiterated below.

The item-level D-matrix can represent the item' s information content. Researchers have performed different forms of weighted summation on the D-matrix to obtain CDI and ACDI indices (Henson & Douglas, 2005; Henson

et al., 2008). CDI is the weighted average of all elements in the D-matrix using the Hamming distance between two KSs, while ACDI is the average of elements in the D-matrix with Hamming distance equal to 1. The calculation formulas are as follows:

$$CDI_j = \frac{\sum h(\alpha_u, \alpha_v)^{-1} \cdot D_{juv}}{\sum h(\alpha_u, \alpha_v)^{-1}}$$

$$ACDI_j = \frac{\sum D_{juv}}{\text{all relevant cells}}$$

Where "all relevant cells" refers to cells in the D-matrix where the Hamming distance between two KSs is 1.

**2.2.2 NR_{PWKL}** The PWKL index for NCD-CAT (hereafter NR_{PWKL}) is the nominal polytomous extension of the PWKL (Posterior-Weighted KL) selection method (Cheng, 2009). The PWKL selection method weights each element of the D-matrix by the posterior probability of each KS and sums them to obtain the PWKL index. Similarly, the NR_{PWKL} selection method performs the same weighted summation on each element of the NR_D matrix. The calculation formula is as follows:

$$NR\_PWKLj(\hat{\alpha}_i) = \sum \left\{ \left[ \sum \log \left( \frac{P(Y_{ij} = y|\hat{\alpha}_i)}{P(Y_{ij} = y|\alpha_c)} \right) P(Y_{ij} = y|\hat{\alpha}_i) \right] \pi(\alpha_c|\mathbf{y}_t) \right\}$$

Where $\pi(\alpha\_c|y\_t)$ is the posterior probability that the examinee' s attribute mastery pattern is $\alpha\_c$ given the observed response vector y_t.

**2.2.3 NR_{PWCDI} and NR_{PWACDI}** Zheng and Chang (2016) followed the idea of PWKL, incorporating posterior probabilities of KSs into both rows and columns of the D-matrix to construct the PWD matrix (Posterior-Weighted D matrix), then applied different weighting schemes based on CDI and ACDI indices to obtain PWCDI and PWACDI indices. Similarly, this study incorporates posterior probabilities into both rows and columns of the NR_D matrix to obtain the NR_{PWD} matrix (Formula 7), then applies two different weighting schemes to weight and average the NR_{PWD} and NR_D matrices to obtain NR_{PWCDI} and NR_{PWACDI} indices (Formulas 8 and 9).

$$NR\_PWD_{jic} = \pi(\alpha_i) \times \pi(\alpha_c) \times \left[ \sum \log \left( \frac{P(Y_{ij} = y|\hat{\alpha}_i)}{P(Y_{ij} = y|\alpha_c)} \right) P(Y_{ij} = y|\hat{\alpha}_i) \right]$$

$$NR\_PWCDIj = \frac{\sum h(\alpha_i, \alpha_c)^{-1} \cdot NR\_PWD_{jic}}{\sum h(\alpha_i, \alpha_c)^{-1}}$$

$$NR\_PWACDIj = \frac{\sum NR\_D_{juv}}{\text{all relevant cells}}$$

Similarly, "all relevant cells" refers to cells in the NR_D matrix where the Hamming distance between two KSs is 1.

**2.2.4 NR_{MPWKL}** Studies by Kaplan et al. (2015) and Zheng and Chang (2016) noted that in KL-family methods, weighting only by the posterior probability of the current KS estimate is not conducive to selecting the most appropriate items because the current KS estimate is often inaccurate when the test is short. In addition to the aforementioned PWCDI-family methods, Kaplan et al.' s (2015) MPWKL method effectively addresses this issue. The MPWKL method performs a second weighted summation of the PWKL index according to the posterior probabilities of KSs. This study applies a second weighted summation of the NR_{PWKL} index according to KS posterior probabilities to obtain the NR_{MPWKL} index. The calculation formula is as follows:

$$NR\_MPWKLj(\hat{\alpha}_i) = \sum \left\{ \sum \left[ \sum \log \left( \frac{P(Y_{ij} = y|\hat{\alpha}_i)}{P(Y_{ij} = y|\alpha_c)} \right) P(Y_{ij} = y|\hat{\alpha}_i) \right] \pi(\alpha_c|\mathbf{y}_t)\pi(\alpha_d|\mathbf{y}_t) \right\}$$

As with dichotomous scoring, the above four PWKL-family methods select items with the maximum corresponding index value.

**2.2.5 NR_{SHE}** Applying Shannon Entropy (SHE) to item selection involves choosing items with the minimum expected Shannon entropy from the item bank to minimize uncertainty in the estimated KS posterior probability distribution. Unlike SHE, when calculating the expected Shannon entropy NR_{SHE}, the expectation should be taken according to the probability of each response category (b_j + 1 categories) rather than considering only two response categories (0 and 1). The calculation formula is as follows:

$$NR\_SHE = \sum \left[ \sum \left( \pi(\alpha_c|\mathbf{y}_t, Y_{t+1} = y) \log \pi(\alpha_c|\mathbf{y}_t, Y_{t+1} = y) \right) Pr(Y_{t+1} = y|\mathbf{y}_t) \right]$$

Where Pr(Y_{t+1}=y|y_t) is the conditional probability of obtaining score y on item t+1 given the observed response vector for the first t items. Its calculation formula is:

$$\sum P(Y_{t+1} = y|\alpha_c)\pi(\alpha_c|\mathbf{y}_t)$$

Where $\pi(\alpha\_c|y\_t)$ is the posterior probability that the examinee' s attribute mastery pattern is $\alpha$_c after completing t items.

**2.2.6 NR_{MI}** The Mutual Information (MI) selection method, proposed by Wang (2013), is more suitable for short tests. Similarly, when calculating NR_{MI}, the expectation of MI should be taken according to the probabilities of b_j + 1 response categories. The calculation formula is as follows:

$$NR\_MI = \sum \left[ \sum \left( \pi(\alpha_c|\mathbf{y}_t, Y_{t+1} = y) \log \frac{\pi(\alpha_c|\mathbf{y}_t, Y_{t+1} = y)}{\pi(\alpha_c|\mathbf{y}_t)} \right) \sum P(Y_{t+1} = y|\alpha_c)\pi(\alpha_c|\mathbf{y}_t) \right]$$

This method selects items with the maximum NR_{MI} index from the item bank.

**2.2.7 NR_{GDI}** Kaplan et al. (2015) applied the GDI (G-DINA discrimination index) to item selection in CD-CAT, where P_j is the average score of the other options (b_j options) excluding the first option.

$$NR\_GDI = \sum \pi(\alpha_c|\mathbf{y}_t) \left[\sum P(Y_{ij} = y|\alpha_c) - \bar{P}_j\right]$$

$$\bar{P}_j = \sum \pi(\alpha_c|\mathbf{y}_t) \sum P(Y_{ij} = y|\alpha_c)$$

Like the PWKL-family indices, a larger GDI value indicates stronger ability to discriminate between different KSs. Therefore, NR_{GDI}···

### 2.3 Exposure Control

Zheng and Wang (2017) developed the SDBS and DBS methods based on binary search algorithms from computer science. Compared with existing item exposure control methods (RP, RT, SHTVOR, etc.), these new methods can more efficiently handle the trade-off between test accuracy and exposure control. This study focuses on the performance of each selection method in terms of classification accuracy and test efficiency, thus exposure control issues are not considered.

### 2.4 Parameter Estimation Methods

Three parameter estimation methods exist in CD-CAT: MLE, MAP, and EAP (Huebner & Wang, 2011). EAP calculates the expected posterior probability of examinees' attribute mastery patterns by multiplying all possible KSs with their corresponding posterior probabilities and taking the expectation given the response vector y_t (Formula 16), followed by dichotomous conversion (Tu et al., 2017). This study adopts the EAP method.

$$\widehat{p}_{ik} = \sum \pi(\alpha_c|\mathbf{y}_t)\alpha_{ck}$$

### 2.5 Termination Strategies

Current CD-CAT termination strategies mainly include fixed-length and variable-length approaches. Guo and Zheng (2019) noted that Tatsuoka' s rule and the dual-criterion rule in variable-length termination strategies exhibit instability when the number of attributes varies, and proposed new methods for variable-length termination strategies from an information theory perspective. The focus of this study is on comparing item selection methods; therefore, only fixed-length termination and variable-length termination with maximum posterior probability reaching a fixed precision are considered (Yu et al., 2019).

## 3. Simulation Studies

To investigate and compare the performance of different item selection methods in NCD-CAT, this study conducted two experiments.

Experiment 1 employed a 2 × 4 × 7 three-factor completely randomized design. The independent variables were item quality, test length, and item selection method. Item quality had two levels (high and low), test length had four levels (5, 10, 15, 20 items), and selection method had seven levels (NR_{PWKL}, NR_{PWCDI}, NR_{PWACDI}, NR_{MPWKL}, NR_{SHE}, NR_{MI}, NR_{GDI}).

The detailed experimental procedure is described below.

### 3.1.1 Data Simulation

For examinees, this study assumed 5 independent attributes creating 32 attribute mastery patterns uniformly distributed in the population, with 3,200 examinees simulated. For the item bank, this study adopted a simulation method similar to Ma and de la Torre (2016) to generate high- and low-quality item banks (600 items each). The Q-matrix used was from de la Torre (2009) (see Appendix B). For high- (or low-) quality items, the probability of selecting the correct answer for examinees who have mastered all attributes required by the item was 0.8 (or 0.6). The probability of selecting incorrect answers was simulated using a uniform distribution.

For response simulation, a random number uniformly distributed between (0, 1) was first generated, then compared against cumulative response probability intervals for options A, B, C, D to determine which option was selected as the answer. For example, if an examinee' s probabilities of selecting each option were 0.1, 0.3, 0.5, and 0.1 respectively, the cumulative response probability distribution would be 0.1, 0.4, 0.9, 1. If the random number was 0.63, the simulation would indicate the examinee selected the third option.

### 3.1.2 Evaluation Metrics

Evaluation metrics for fixed-length NCD-CAT include Pattern Match Ratio (PMR), $^2$, and Test Overlap Rate (TOR). Detailed descriptions of each metric appear in Appendix C.

Experiment 1 results are presented in Table 1, Appendix D: Table D-1, Figure D-1, and Figure D-2.

**Table 1** Pattern Match Ratios for Seven Item Selection Methods and Differences Compared to NR_{PWKL}

[Table content showing PMR values and differences across conditions]

**Overall Trends (Table 1):**

(1) As item quality increased from low to high, the PMR of all selection methods improved noticeably, with the improvement being particularly pronounced for short tests (5 and 10 items). For example, when test length was 5 items, the PMR of all methods nearly doubled. (2) As test length increased, the PMR of all methods improved, with the degree of

improvement varying by test length and item quality. Specifically, when test length increased from 5 to 10 items, regardless of item quality, PMR improved by over 20%. When test length increased from 10 to 15 or 20 items, PMR for low-quality item NCD-CAT still improved by nearly or over 10%. Comparatively, improving item quality enhances NCD-CAT classification accuracy more than increasing test length.

**Comparison of Each Method with Baseline NR_{PWKL} (Table 1):**

(1) The PMR of NR_{PWKL}-family methods (NR_{PWCDI}, NR_{PWACDI}, NR_{MPWKL}) and NR_{SHE}/MI methods were higher than or equal to that of NR_{PWKL} under all conditions, especially for short tests with high-quality items. (2) As test length increased, the PMR advantage of other methods over NR_{PWKL} continuously decreased. This trend is consistent with the changes in PWCDI, PWACDI, and MPWKL methods reported in Zheng and Chang (2016).

**Comparison of NR_{PWCDI} and NR_{PWACDI} with NR_{MPWKL} and NR_{SHE}/MI (Table 1):**

(1) NR_{PWCDI} and NR_{PWACDI} methods performed very similarly to NR_{MPWKL}, consistent with Zheng and Chang (2016) because these methods are all PWKL-based improvements. (2) Shannon entropy methods (SHE/MI) showed PMR higher than or equal to NR_{PWKL}-family methods. Wang (2013) noted that the MI selection method performs well in short tests.

The NR_{GDI} method performed worst among the seven methods under low-quality item conditions, while slightly outperforming the baseline NR_{PWKL} under high-quality item conditions.

**

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