Development of Detection Methods for Main Effect DIF and Interactive DIF in Cognitive Diagnostic Assessment: A Recursive Partitioning Perspective

Authors: Liu Kai, Guo Zhichen, Wang Qin, Wang Daxun, Cai Yan, Dongbo Tu, Wang Daxun, Cai Yan, Dongbo Tu

Date: 2025-12-09T00:00:00+00:00

Abstract

In cognitive diagnostic assessment, Differential Item Functioning (DIF) detection constitutes an important technical approach for evaluating test fairness and measurement validity. However, existing cognitive diagnostic DIF detection methods are limited to main-effect DIF detection from a single covariate perspective and lack effective detection means for interaction DIF arising from the interaction of multiple covariates. To address this limitation, this study draws upon the core idea of recursive partitioning techniques and proposes a novel method (denoted as ISRPM) capable of simultaneously detecting both main-effect DIF and interaction DIF in cognitive diagnostic assessment. Simulation study results demonstrate that ISRPM not only yields overall performance comparable to traditional methods in main-effect DIF detection, but more importantly, exhibits superior performance in interaction DIF detection compared to traditional approaches. Empirical studies further substantiate the method' s usability, revealing that ISRPM demonstrates high consistency with traditional DIF detection methods in detection results while showing potential advantages in identifying interaction DIF. Overall, the proposed ISRPM is expected to further enhance the accuracy of cognitive diagnostic DIF detection and facilitate the promotion and application of cognitive diagnostic assessment in psychological and educational measurement practice.

Full Text

Development of Main Effect DIF and Interactive DIF Detection Methods in Cognitive Diagnosis Assessments: A Recursive Partitioning Perspective

LIU Kai[1],[2], GUO Zhichen[1], WANG Qin[1], WANG Daxun[1], CAI Yan[1], TU Dongbo[1]

[1] School of Psychology, Jiangxi Normal University, Nanchang 330022, China
[2] College of Psychology, Liaoning Normal University, Dalian 116029, China

Abstract

With the growing recognition of the advantages of cognitive diagnosis (CD) in psychological and educational measurement, applying the CD framework to test development has become an important research direction in the field. In the development of cognitive diagnostic assessments, detecting differential item functioning (DIF) remains a crucial quality control procedure to ensure test fairness and validity. However, existing CD-based DIF detection methods typically focus on a single covariate at a time. While these approaches are effective for identifying main effect DIF induced by a single covariate, they are limited in detecting interactive DIF caused by the interaction among multiple covariates. Such limitations may compromise the fairness and interpretability of assessment outcomes. To address this issue, the present study integrates CD modeling with recursive partitioning techniques by proposing a novel DIF detection method, namely the Item-based Sequential Recursive Partitioning Method (ISRPM). Building on the core principles of recursive partitioning, the ISRPM allows the simultaneous consideration of multiple covariates within a single DIF detection procedure and facilitates the identification of both main effect DIF and interactive DIF in cognitive diagnostic assessments.

To evaluate the performance of the proposed method, a series of Monte Carlo simulation studies were conducted focusing on two key objectives: (1) examining how factors such as sample size per group, DIF magnitude, DIF type, item quality, correlations among attributes, and the influence of demographic covariates on attribute mastery distribution affect the performance of ISRPM; and (2) comparing ISRPM with several existing DIF detection methods across varied experimental conditions. In addition, to illustrate its practical utility, ISRPM was applied to a cognitive diagnostic version of the Schizotypal Personality Questionnaire (DC-SPQ) and compared with five established DIF detection methods.

The results showed that (1) sample size, DIF magnitude, and item quality substantially influenced the performance of all methods; and (2) when items exhibited interactive DIF, ISRPM achieved higher detection accuracy than the Wald, LR, FS-Wald, FS-LR, and Mantel–Haenszel (MH) approaches. When only the

main effect DIF was present, the overall performance of ISRPM was comparable to that of the existing methods.

These findings suggest that ISRPM provides a flexible and effective framework for identifying both main effect DIF and interactive DIF in cognitive diagnostic assessments, thereby contributing to methodological advancements in fairness evaluation and the broader application of CD-based measurement in psychological and educational measurement.

Keywords: cognitive diagnosis assessments, differential item functioning, main effect DIF, interactive DIF, recursive partitioning

1. Introduction

In today' s rapidly evolving information society, the rise of cognitive diagnosis (CD) has brought significant and substantive transformation to the field of psychological and educational measurement [**?**, **?**, **?**]. Unlike traditional test scoring systems that focus solely on overall ability levels, cognitive diagnosis emphasizes revealing individuals' internal psychological processing mechanisms and cognitive structures, thereby better achieving the core goal of promoting individual development through assessment [**?**]. In psychological assessment, cognitive diagnosis can be used not only to evaluate individuals' cognitive functional status but also to accurately identify symptom characteristics, providing important data support for clinicians to implement precise treatment and early intervention [**?**, **?**, **?**]. In educational measurement, while traditional academic achievement tests typically evaluate student ability based on total scores or proficiency levels, cognitive diagnosis focuses more on the learning process itself. By precisely locating individuals' strengths and weaknesses across different cognitive components, it provides effective references for educators to develop targeted instructional strategies and knowledge remediation plans [**?**]. Overall, against the backdrop of rapid information technology development, cognitive diagnosis has not only broadened the research perspective of psychological and educational measurement but also provided effective technical support for implementing personalized instruction and precise psychological treatment.

In recent years, cognitive diagnosis has become a research frontier in psychological and educational measurement both domestically and internationally, and has been widely applied to psychological and educational test development [**?**, **?**, **?**, **?**], owing to its unique advantage in providing fine-grained diagnostic information. During the development of cognitive diagnostic assessments, test developers are particularly concerned about whether measurement results produce systematic bias against specific groups, thereby placing those groups at an undue advantage or disadvantage. This issue essentially involves evaluating test fairness. Within the psychometric framework, concepts closely related to test fairness primarily include measurement invariance (MI; [**?**]) and differential item functioning (DIF; [**?**]). Measurement invariance refers to the property that

a test maintains consistent measurement characteristics across different examinee groups (e.g., gender and cultural background). When systematic differences exist in measurement characteristics across groups, this indicates measurement noninvariance (MN). When such noninvariance manifests at the item level, it means the item exhibits DIF. Within the cognitive diagnosis framework, DIF is typically defined as: under the condition of identical attribute mastery patterns, examinees from different groups exhibit systematic differences in the probability of correctly answering the same item [**?**, **?**]. Previous research has shown that the presence of DIF not only undermines the measurement fairness of cognitive diagnostic tests but may also reduce measurement validity [**?**]. Furthermore, DIF can lead to biased estimation of item parameters, thereby causing misclassification of examinees' attribute mastery patterns and ultimately resulting in potentially misleading assessment outcomes [**?**]. Therefore, conducting DIF analysis during the development and validation stages of cognitive diagnostic tests has become a widely recognized critical step among psychometric researchers [**?**, **?**, **?**]. This process constitutes an important component of test quality control and a necessary condition for ensuring test fairness and measurement validity.

Currently, researchers have proposed various DIF detection methods applicable to cognitive diagnostic assessments, which can be broadly categorized into parametric and nonparametric approaches. Nonparametric methods have the advantages of low sample size requirements, straightforward and intuitive operation, and ease of understanding; however, their detection accuracy is generally lower compared to parametric methods. Representative methods of this type include the Mantel-Haenszel procedure and SIBTEST developed by [**?**]. Parametric methods require estimation of specific cognitive diagnosis model parameters. Although relatively complex to implement and computationally costly, they demonstrate superior accuracy in DIF detection results. Typical parametric methods include the Wald test [**?**, **?**, **?**], logistic regression [**?**], and likelihood ratio test [**?**]. Notably, the detection performance of these parametric methods has been fully validated in simulation studies, providing reliable theoretical and technical support for DIF analysis in cognitive diagnostic tests. Given the precision advantages of parametric methods in DIF detection, they have received more attention in recent years compared to nonparametric methods. Therefore, this study focuses on the development of parametric cognitive diagnosis DIF detection methods.

Despite the generally good performance of existing cognitive diagnosis DIF detection methods, they still have certain limitations. Specifically, these methods typically can only independently evaluate whether a single covariate induces DIF, without fully considering that interactions among multiple covariates may also lead to DIF. For example, the interaction between gender and household registration may affect examinees' response patterns on specific items, thereby causing DIF. To clearly distinguish DIF caused by interactions among multiple covariates from DIF caused solely by a single covariate, this study defines the former as "interactive DIF"and the latter as "main effect DIF."Previous research has shown that interactive DIF may be prevalent in psychological and educa-

tional tests, further increasing the complexity of measurement bias sources. For instance, [**?**] found that measurement bias caused by interactions among covariates was evident in psychological tests assessing adolescent delinquent behavior. Similarly, [**?**] discovered in intelligence structure tests that DIF in certain items might stem from the interaction between gender and age. Therefore, we have reason to speculate that interactive DIF is also likely to exist in cognitive diagnostic tests and may adversely affect test fairness and measurement validity.

In recent years, numerous researchers have emphasized the importance and theoretical-practical value of interactive DIF detection in test quality analysis [**?**, **?**, **?**]. First, interactive DIF detection helps more comprehensively reveal the complex sources that cause test functioning differences [**?**], thereby providing test developers with more valuable references for item revision. [**?**] pointed out in their research that individual identity is the result of the intersection of multiple demographic characteristics. This perspective suggests that when exploring measurement bias, in addition to considering the main effects of single demographic variables, potential interactions among variables should also be taken into account. Based on this understanding, compared to main effect DIF detection from a single covariate perspective, conducting interactive DIF detection helps identify more potential sources of measurement bias, thereby providing test developers with more precise and targeted item revision bases. Second, the concealed nature of interactive DIF makes its detection particularly necessary. Compared to main effect DIF, interactive DIF is more difficult for traditional methods to identify. This is primarily because traditional DIF detection methods typically assume independence among different demographic variables, so they can only identify DIF forms caused by a single covariate and have difficulty revealing measurement bias caused by interactions among multiple covariates. During test development, researchers often pre-assume possible sources of measurement bias based on single demographic variables (e.g., gender, ethnicity) while neglecting the impact of variable interactions on test fairness. This neglect may lead to some measurement bias being overlooked, thereby affecting test fairness. In view of this, accurately identifying interactive DIF helps further enhance the fairness of cognitive diagnostic assessment results. Finally, the identification of interactive DIF is also valuable for further improving measurement validity. The presence of interactive DIF not only may affect examinees' performance on specific items but also may hinder accurate estimation of their attribute mastery patterns. If DIF caused by interactions among covariates is not identified, the quality of cognitive diagnostic assessment results will be compromised. In summary, interactive DIF detection is valuable for comprehensively revealing sources of measurement bias, safeguarding test fairness, and improving measurement validity. However, existing cognitive diagnosis DIF detection methods still have obvious deficiencies in identifying interactive DIF, which poses a challenge to the effective assurance of test fairness and measurement validity. Therefore, developing a method that can simultaneously identify both main effect DIF and interactive DIF within the cognitive diagnosis framework not only helps improve the theory and methodology of cognitive diagnosis

DIF detection but also holds important value for promoting the reasonable application of cognitive diagnostic assessments in practice. Based on this research need, this study aims to propose a method that can simultaneously identify main effect DIF and interactive DIF, thereby providing more comprehensive technical support for fairness evaluation in cognitive diagnostic tests.

With the continuous development of modern data processing technology, data mining (DM) techniques have been widely applied in psychological and educational measurement. [**?**] pointed out that when conducting DIF analysis, the process of identifying covariates that cause DIF is highly similar to variable selection in regression modeling. This provides a theoretical basis for introducing variable selection methods into DIF detection and offers new ideas for improving and innovating DIF detection methods. Compared with traditional DIF detection methods, DM techniques have advantages such as high efficiency, strong flexibility, and the ability to handle multiple covariates simultaneously, showing great potential for application in identifying complex DIF forms [**?**]. Based on the application potential of variable selection methods in DIF detection, researchers have begun to combine item response theory (IRT) with variable selection techniques to develop a series of new methods capable of identifying complex DIF forms (e.g., [**?**, **?**, **?**]), providing important methodological references for conducting interactive DIF detection in cognitive diagnostic contexts. Among these, recursive partitioning (RP) is the most representative variable selection technique in this category.

The basic principle of this technique is to recursively divide the feature space covered by predictor variables into several subregions and fit a relatively simple model within each region [**?**]. Through continuous data partitioning and modeling, RP methods can intuitively reveal the relationship between covariates' main effects and their interactions and test item parameters, thereby providing effective technical support for exploring complex sources of measurement bias. According to [**?**], [**?**], and [**?**], the advantages of RP technology in DIF analysis are specifically reflected in three aspects: (1) It breaks through the limitation of traditional methods that require artificially dividing focal and reference groups before analysis, and can automatically identify optimal covariate grouping criteria in a data-driven manner, thereby reducing the risk of potential DIF being missed due to improper group settings. (2) Traditional methods typically have difficulty examining interactions among covariates, thus having significant limitations in interactive DIF detection; in contrast, RP technology can not only reveal complex interactions among multiple covariates but also further evaluate their impact on measurement bias, thereby helping to improve the detection accuracy of interactive DIF. (3) RP technology can flexibly handle various types of covariates, including continuous, multicategorical, ordinal, and dichotomous variables, thereby broadening the applicability of DIF detection methods. In summary, RP technology, with its flexible, intuitive, and robust characteristics, provides an effective framework for DIF detection and demonstrates unique advantages in handling complex situations involving interactions among covariates.

To date, the development of DIF detection methods based on RP technology has mainly been carried out within the item response theory (IRT) framework. Such methods can be roughly divided into two categories: global-level RP methods and item-level RP methods. Global-level methods judge whether the main effects of single covariates and interactions among multiple covariates cause measurement bias by testing parameter instability across the space covered by covariates. However, such methods can only identify covariates that cause DIF but cannot further determine which specific items exhibit DIF. In existing research, Rasch Trees [**?**] and polytomous Rasch Trees [**?**] have become typical representatives of this category. Compared with global-level methods, item-level RP methods can not only identify covariates that induce DIF but also locate specific DIF items, thus demonstrating higher flexibility and practicality in DIF detection. Typical representatives of this category include item-focused trees (IFT) based on the Rasch model [**?**] and item-focused trees based on the partial credit model (PCM-IFT; [**?**]). Currently, item-level RP methods have received more attention due to their dual advantages in identifying covariates and locating DIF items. This is specifically reflected in: such methods can simultaneously handle multiple covariates in a single analysis [**?**] and further explore at the item level whether interactions among these covariates induce differential functioning.

Although existing research has confirmed that RP technology has great research potential and application value in main effect DIF and interactive DIF detection, current research work has only verified its effectiveness in DIF detection within the IRT framework. More importantly, there are currently no published studies that systematically explore interactive DIF detection in cognitive diagnostic tests. Notably, IRT and cognitive diagnostic theory have obvious differences in model assumptions, measurement objectives, and data analysis methods, which poses new challenges for directly applying RP technology to cognitive diagnostic contexts. Therefore, how to extend RP technology to main effect DIF and interactive DIF detection in cognitive diagnostic assessments and verify whether it can maintain existing detection accuracy and applicability in cognitive diagnostic assessments remains an important issue to be explored. Based on this, this study develops and validates a new DIF detection method applicable to cognitive diagnostic tests by drawing on the core ideas of RP technology. This method aims to provide effective technical support for identifying main effect DIF and interactive DIF in cognitive diagnostic assessments, thereby further improving the test fairness evaluation system of cognitive diagnosis and ultimately promoting the in-depth application of cognitive diagnosis technology in psychological and educational measurement.

2. Development of the Item-Based Sequential Recursive Partitioning Method (ISRPM)

The main purpose of this paper is to develop a novel method for detecting main effect DIF and interactive DIF within the cognitive diagnosis framework, namely the Item-based Sequential Recursive Partitioning Method (ISRPM). This method combines recursive partitioning technology with cognitive diagnosis models, treating DIF test statistics constructed based on model parameter estimation as criteria for covariate splitting. By comparing the splitting effects of different covariate splitting schemes to identify the optimal splitting method, it generates a recursive partitioning tree for each item that can reflect its DIF manifestation pattern. Specifically, ISRPM compares the statistics corresponding to potential splitting schemes of candidate covariates at each level of the tree, selects the covariate and splitting point that maximize between-group differences in item parameters, and recursively partitions the examinee sample and expands the tree structure level by level, ultimately revealing which covariates and in what manner may induce DIF. The following first briefly introduces the cognitive diagnosis model adopted in this study and its DIF definition, then elaborates on the operational steps and theoretical foundation of ISRPM.

2.1 The Generalized DINA (G-DINA) Model Within the cognitive diagnosis framework, cognitive diagnosis models (CDMs) are a class of psychometric models that fully integrate cognitive variables. As a core technical component of cognitive diagnostic assessment, the quality of CDMs directly determines the validity of cognitive diagnostic results [**?**]. Currently, researchers have developed various CDMs with good diagnostic performance that can be applied to different test situations and theoretical assumptions. This study adopts the generalized deterministic input, noisy "and" gate (G-DINA) model. This model, proposed by [**?**], is a generalized psychometric model that extends the deterministic input, noisy "and" gate (DINA) model.

In the G-DINA model, after examinee i completes item j, they are classified into $2^{K_j^*}$ categories, where K_j^* is the number of attributes measured by item j. To facilitate the introduction of its mathematical expression, we can assume that the first K_j^* attributes are those required to correctly answer item j, and let α_v represent the v-th examinee attribute mastery pattern, where $v = 1, \ldots, 2^{K_j^*}$. Then, within the G-DINA model framework, the conditional probability of examinee i correctly answering item j can be expressed as:

$$P_{ij} = P(X_{ij} = 1 \mid \alpha_i) = f(\delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ik} + \sum_{k=1}^{K_j^*-1} \sum_{k'=k+1}^{K_j^*} \delta_{jkk'}\alpha_{ik}\alpha_{ik'} + \cdots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik})$$

where $f(\cdot)$ is a link function. Depending on the link function adopted, the G-DINA model has different representations. Three commonly used link functions

are identity, log, and logit link functions; δ_{j0} is the intercept term for item j, representing the probability of correctly answering item j when the examinee has not mastered all measured attributes (this value is generally nonnegative); δ_{jk} is the main effect of attribute k on item j, generally nonnegative, where a larger value indicates greater contribution of mastering this attribute to correctly answering item j; α_{ik} represents examinee i' s mastery status of attribute k in the v-th attribute mastery pattern (1 if mastered, 0 otherwise); $\delta_{jkk'}$ is the interaction effect between attributes k and k' on item j; and $\delta_{j12...K_j^*}$ is the interaction effect among all attributes measured by item j. It is important to emphasize that to maintain consistency with previous similar studies (e.g., [**?**, **?**, **?**]), this study adopts the identity link function.

2.2 DIF Definition Under the G-DINA Model Unlike DIF detection within the IRT framework, DIF in CDMs needs to be redefined. This is because CDMs provide examinees' mastery status on discrete attributes rather than locating examinees on a continuous latent trait continuum. According to [**?**], DIF within the G-DINA model framework can be expressed as:

$$\Delta_{jv} = P_{jF}(\alpha_v) - P_{jR}(\alpha_v)$$

where $P_{jF}(\alpha_v)$ and $P_{jR}(\alpha_v)$ represent the probabilities of correctly answering item j for examinees with attribute mastery pattern α_v in the focal group (F group) and reference group (R group), respectively; $\Delta_{jv} = 0$ indicates that item j does not exhibit DIF, otherwise, DIF exists.

2.3 Development of ISRPM: Basic Principles and Detection Steps This section details the operational procedure and key technical aspects of ISRPM. ISRPM fully integrates recursive partitioning technology with the G-DINA model, achieving identification of main effect DIF and interactive DIF through recursive partitioning of covariates.

The basic idea is as follows: For each item, given a set of candidate covariates, ISRPM first identifies all possible splitting points for each covariate and partitions the examinee response data into several subsamples (e.g., response datasets for different genders) based on these splitting points. Subsequently, the G-DINA model is fitted and model parameters are estimated separately within each subsample. Next, the obtained parameters are substituted into a prespecified splitting criterion (i.e., DIF test statistic) for calculation, and the splitting scheme (i.e., covariate and its splitting point) corresponding to the maximum statistic value is identified as the first optimal splitting variable and its splitting point, based on which the initial partition is executed. After completing the first-round partition, ISRPM repeats the above process in the newly generated subsamples, sequentially determining the optimal splitting scheme at each level and executing data partitioning until preset variable search termination rules are met. Ultimately, ISRPM generates a recursive partitioning tree diagram for

each item identified as exhibiting DIF, visually displaying on which covariates and in what manner the item shows DIF. The following uses a single item j as an example to 详细介绍 ISRPM' s main operational steps and key technical details.

Step 1: Identify Covariates of Interest and Their Potential Splitting Points

Before conducting cognitive diagnosis DIF detection, it is first necessary to specify all covariates of interest and their potential splitting points. The goal of ISRPM is to build a recursive partitioning tree for each item exhibiting differential functioning, where the root node includes all covariates of interest and their variable levels. These variable levels together constitute the complete feature space covered by the covariates, while child nodes correspond to subsets of this feature space. From the perspective of examinee grouping logic, this structure of gradually subdividing the feature space is consistent with the idea of grouping examinees based on covariate levels in traditional DIF detection. In traditional DIF detection, researchers partition examinees into focal and reference groups based on each covariate of interest and its levels. Typically, numbers 1 and 2 represent the focal and reference groups, respectively, corresponding to different feature levels of the same demographic characteristic. Let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ denote m covariates related to examinees, with each covariate containing two levels. For a dichotomous covariate x_m, two corresponding subsets A_1 and A_2 can be defined based on its two levels, expressed as $A_1 \cup A_2 = A$, where A represents the entire feature space jointly covered by all covariates, corresponding to the focal and reference groups defined by covariate x_m. Since this variable contains only two levels, it has one and only one potential splitting point.

Step 2: Search for the First Optimal Splitting Covariate Based on Selected Covariate Splitting Criteria

After identifying covariates of interest and their potential splitting points, ISRPM needs to estimate parameters for the grouped datasets corresponding to each potential splitting scheme. To ensure reliable estimation results, this study adopts the widely validated G-DINA model to complete this estimation process. After obtaining parameter estimates, the key task is to determine appropriate covariate splitting criteria to search for the first optimal splitting covariate and thereby judge whether the current item exhibits DIF on that variable. According to [**?**], commonly used splitting criteria in recursive partitioning frameworks mainly include two categories: (1) splits based on impurity measures, such as Gini Index or Shannon Entropy; and (2) splits based on test statistics, such as the log-likelihood test statistic. The former achieves partitioning by measuring the homogeneity of samples within nodes, while the latter conducts statistical tests on parameter differences to determine whether further node splitting is needed based on between-group differences. In the context of parametric cognitive diagnosis DIF detection, splitting based on test statistics better aligns with the research objectives; therefore, this study adopts this type of method as the covariate splitting criterion for ISRPM.

When searching for the first optimal splitting covariate, ISRPM generates candidate child nodes for all potential splitting points of each candidate covariate x_m. These nodes constitute the left and right potential child nodes of the first level of the recursive partitioning tree. Based on all feature levels of covariate x_m, the datasets for the focal and reference groups defined by its potential splitting points can be obtained, and the G-DINA model is fitted on each group to estimate the corresponding item parameters according to Equation (4):

$$\hat{\delta}_{j,\text{node}} = \sum_{g \in \{l,r\}} I(\mathbf{x}_i \in A_g) \cdot \hat{\delta}_{j,g}$$

where node represents the G-DINA model item parameter values for item j under different child nodes in the recursive partitioning tree; $I(\cdot)$ is an indicator function that equals 1 if condition d is satisfied, otherwise 0; subscripts l and r are abbreviations for the English words "left"and "right,"respectively; A_l and A_r represent the focal and reference groups defined by the covariate; $\hat{\delta}_{j0,l}$ represents the intercept parameter corresponding to the left child node established based on covariate x_m (i.e., A_l), which can also be regarded as the intercept parameter for the focal group defined by x_m on the current item, with the rest interpreted analogously.

After obtaining the grouped item parameter estimates for the current item under each candidate covariate x_m, it is necessary to further calculate the DIF test statistic corresponding to each potential splitting point for each covariate. Generally, any test statistic that has been validated as effective in cognitive diagnosis DIF detection research can be adopted. Previous studies have found that, within the cognitive diagnosis framework, Wald statistic-based detection methods demonstrate superior performance in controlling Type I error rates and statistical power compared to MH, SIBTEST, and likelihood ratio test methods [**?**]. Therefore, this study selects the Wald statistic as the covariate splitting criterion. The Wald statistic follows a chi-square distribution, and its mathematical expression is:

$$W = (\hat{\delta}_{j,R} - \hat{\delta}_{j,F})'(\hat{\Sigma}_{j,R} + \hat{\Sigma}_{j,F})^{-1}(\hat{\delta}_{j,R} - \hat{\delta}_{j,F})$$

where $\hat{\delta}_{j,R}$ and $\hat{\delta}_{j,F}$ represent the item parameter estimate vectors for the reference group (R) and focal group (F) under the G-DINA model framework for item j, respectively; $\hat{\Sigma}_{j,R}$ and $\hat{\Sigma}_{j,F}$ are the sampling variance-covariance matrices corresponding to the item parameter estimates for the reference and focal groups on item j. Notably, before using the Wald statistic for DIF detection, item purification is typically implemented to place parameters from different groups on the same scale, thereby ensuring comparability of item parameters across groups [**?**]. However, this study does not conduct item purification, primarily for three reasons: First, from the perspective of measurement scale consistency, unlike

IRT models based on continuous latent traits, model parameters in the cognitive diagnosis framework are naturally on the same scale, thus having lower dependence on item purification. Specifically, the G-DINA model adopted in this study measures examinees' mastery of discrete attributes rather than continuous ability levels, so no additional parameter scale adjustment is needed [**?**]. Second, from the perspective of purification procedure applicability, although item purification helps improve DIF detection accuracy, it faces multiple challenges in practice: (1) The purification process involves multiple rounds of parameter estimation and DIF statistic calculation and requires dynamic adjustment of the anchor set, leading to increased computational complexity and cumbersome procedures [**?**]; (2) Item purification cannot completely ensure that the anchor set contains no DIF items [**?**]; (3) When tests contain multiple DIF items, the purification process may be disturbed by "masking effects" and "swamping effects," thereby weakening the accuracy of anchor set construction [**?**, **?**]. Finally, from the perspective of internationally published literature, the vast majority of studies do not adopt item purification procedures when using Wald tests for cognitive diagnosis DIF detection (e.g., [**?**, **?**, **?**, **?**, **?**]). To maintain consistency with previous research, this study also does not conduct item purification.

After calculating the Wald statistics for all candidate covariates and their potential splitting points, ISRPM ranks the covariates in descending order based on the statistical values and treats the covariate corresponding to the maximum value as the first optimal splitting covariate. For convenience in subsequent steps, this step assumes that the first optimal splitting covariate is the first covariate among all covariates, denoted as x_1. Based on the optimal splitting point identified from covariate x_1, the examinee sample can be partitioned into two subsets: A_l and A_r, which correspond to the left and right child nodes of the first level of the recursive partitioning tree. It should be noted that once the first optimal splitting covariate is determined, the DIF detection process for item j on this covariate is considered complete. Therefore, in subsequent covariate search and partitioning processes, variable x_1 will be removed from the candidate covariate set to avoid redundant splitting or variable redundancy.

Step 3: Recursively Search for Optimal Splitting Covariates and Their Optimal Splitting Points Among Remaining Covariates

After determining the first optimal splitting covariate, ISRPM does not terminate DIF detection for item j but continues to perform recursive partitioning on the response data based on the two generated child nodes (i.e., A_l and A_r) to gradually expand the hierarchical structure of the recursive partitioning tree. At this stage, ISRPM continuously searches for optimal splitting covariates and their optimal splitting points for subsequent levels among the remaining covariates. Specifically, this recursive search process includes the following components:

First, based on the optimal splitting point of the optimal splitting covariate obtained in the previous step, ISRPM partitions the examinee sample into two parts corresponding to the two feature levels defined by this splitting point.

Subsequently, these two levels are fully crossed with the potential splitting points (or levels) of the remaining covariates to generate new feature level combinations. For example, when covariate x_1 has been identified as the first optimal splitting covariate, its two levels will be crossed with the two feature levels of a remaining covariate (e.g., x_2) to form four new level combinations.

Second, ISRPM redefines the focal and reference groups based on these newly generated feature combinations. For example, on the basis of the right child node A_r established in Step 2, after crossing it with the two feature levels of covariate x_2, two new feature combinations are obtained: $A_r \cap A_{2,1}$ and $A_r \cap A_{2,2}$. The former can be regarded as the new focal group, while the latter serves as the new reference group, both of which are used for subsequent model fitting and DIF test statistic calculation.

Next, ISRPM selects the covariate and its corresponding splitting point with the largest Wald statistic as the optimal splitting scheme for the current level by comparing the Wald statistics corresponding to each candidate splitting scheme. For example, if the calculation results show that the Wald statistic is largest for the crossing combinations between the right child node established in Step 2 (i.e., A_r) and all levels of covariate x_2, ISRPM will further generate two new child nodes under this node, denoted as $A_{r,l}$ and $A_{r,r}$. As the partitioning tree structure further expands, ISRPM updates Equation (4) based on the new splitting level to obtain Equation (6) below:

$$\hat{\delta}_{j,\text{new node}} = \sum_{g \in \{l,r\}} I(\mathbf{x}_i \in A_{\text{parent},g} \cap A_{m,g}) \cdot \hat{\delta}_{j,g}$$

where $A_{\text{new node}}$ represents the newly established child node; $\hat{\delta}_{j0,\text{new node}}$ represents the intercept parameter corresponding to the newly established child node, with the rest interpreted analogously.

Step 4: Repeat Steps 2 to 3 Until Splitting Termination Rules Are Met

During the process of repeatedly searching for optimal splitting covariates and their splitting points through Steps 2 and 3, ISRPM performs multiple rounds of partitioning on the response data. As the splitting level deepens, the sample size allocated to each subsequent child node gradually decreases. To ensure that each child node has a sufficient sample size, thereby guaranteeing the accuracy of parameter estimation and the validity of DIF detection results as much as possible, this study sets the following two types of splitting termination rules for ISRPM: (1) No more covariates are available in the covariate search space for recursive partitioning; (2) The sample size of any child node falls below a preset minimum threshold (set at 100 in this study). When either of the above conditions is met, ISRPM immediately terminates the covariate search and recursive partitioning process for the current item.

Step 5: Terminate DIF Detection and Output Recursive Partitioning Tree Diagram Displaying the Current Item' s DIF Pattern

When the search for optimal splitting covariates and splitting points meets the termination conditions, ISRPM immediately stops DIF detection for the current item and decides whether to output the recursive partitioning tree diagram based on the final detection results. Specifically, for a single item, there are only two possible DIF detection outcomes: (1) If no substantively meaningful covariate splitting occurs during the variable search process (i.e., the p-value corresponding to the Wald statistic is higher than the preset significance level), the item is determined to have no DIF, and no recursive partitioning tree needs to be generated; (2) If at least one covariate splitting is substantively meaningful (i.e., the p-value corresponding to the Wald statistic does not exceed the preset significance level), the item is determined to exhibit DIF, and ISRPM outputs the corresponding recursive partitioning tree structure diagram, which visually displays on which covariates and in what manner the item shows DIF.

[Figure 1: see original paper] shows schematic diagrams of recursive partitioning trees for different items under conditions with two dichotomous covariates. To facilitate understanding of the information in the figure, several points need to be clarified: (a) Each child node displays the estimated G-DINA model item parameters for different groups; (b) Arrow types reflect whether the Wald statistic reaches the preset significance level, where solid arrows indicate significance and dashed arrows indicate nonsignificance; (c) Arrow types are also used to identify whether the current item exhibits DIF on the corresponding covariate, with solid arrows indicating DIF presence and dashed arrows indicating DIF absence. Overall, the four items in Figure 1 present three different DIF patterns. Item 1 shows main effect DIF only on variable x_1, specifically manifested as solid arrows appearing only on all child nodes at the first level of the recursive partitioning tree. Item 2 exhibits a special DIF pattern: First, if the difference between $A_{l,1}$ and $A_{l,2}$ is roughly equivalent to the difference between $A_{r,1}$ and $A_{r,2}$, it indicates that the item shows only main effect DIF on the two covariates without interactive DIF; however, when these two differences are significantly unequal, it indicates that Item 2 exhibits both main effect DIF and interactive DIF on the two covariates simultaneously. Item 3 exhibits only interactive DIF, mainly reflected as: dashed arrows appear only on the first-level child nodes of the recursive partitioning tree structure, while solid arrows appear on all second-level child nodes, indicating that DIF can only be identified after crossing all levels of the two covariates. In Item 4, solid arrows appear on all first-level child nodes and the left half of the second-level child nodes of the recursive partitioning tree, while the right half of the second level shows dashed arrows, indicating that the item exhibits both main effect DIF and interactive DIF simultaneously.

2.4 Type I Error Control When conducting DIF detection based on multiple covariates, ISRPM inevitably involves multiple testing problems. In statistical literature, an important concept is the familywise error rate (FWER).

According to [**?**], FWER refers to the probability of at least one erroneous rejection when simultaneously testing a set of null hypotheses. Within the methodological framework of ISRPM, this means: for the same item, when hypothesis testing is conducted simultaneously on multiple covariates, as long as one erroneous rejection occurs, the item can be considered to have been incorrectly determined as exhibiting DIF, that is, a Type I error has occurred. To effectively control the FWER of ISRPM during the DIF detection process at a given global significance level, this study follows the recommendation of [**?**] by adopting a more stringent significance level for each individual test, that is, adjusting the local significance level through the overall significance level to maintain overall Type I error control.

Specifically, this study employs the Bonferroni adjustment method to achieve Type I error control. This method obtains the local significance level corresponding to each covariate by dividing the overall significance level α (set at 0.05 in this study) by the number of hypothesis tests performed simultaneously. According to recommendations from [**?**], [**?**], and [**?**], for a single item, when simultaneously detecting potential DIF across multiple covariates, the overall significance level should be proportionally adjusted according to the number of covariates to counteract the increased false positive probability caused by multiple testing, thereby ensuring that the probability of incorrectly classifying a DIF-free item as DIF does not exceed α. Based on the Bonferroni adjustment method, the local significance level for ISRPM can be calculated according to the following formula:

$$\alpha_{\text{local}} = \frac{\alpha}{m}$$

where α refers to the overall significance level, set at $\alpha = 0.05$ in this study, and m refers to the number of covariates on which DIF detection is performed.

---

3. Simulation Study

3.1 Experimental Design The main purpose of this simulation study is to evaluate the performance of the proposed ISRPM method on commonly used DIF detection performance metrics under various experimental conditions and to compare it with several widely used cognitive diagnosis DIF detection methods internationally. Referencing the research design of [**?**], this simulation study sets the following eight manipulated variables: (1) sample size per group (three levels: 500, 1000, and 2000); (2) DIF magnitude (two levels: 0.05 and 0.1, representing small and large DIF, respectively); (3) DIF pattern (three levels: main effect DIF only, interactive DIF only, and both main effect DIF and interactive DIF); (4) item quality (two levels: high quality and medium quality, where parameters $P(\alpha_v = \mathbf{0})$ and $1 - P(\alpha_v = \mathbf{1})$ for high-quality items are randomly drawn from a uniform distribution $U(0.05, 0.15)$, while corresponding

parameters for medium-quality items are drawn from $U(0.15, 0.25)$); (5) DIF detection method (six levels: ISRPM, Wald, LR, Wald-FS, LR-FS, and MH); (6) attribute mastery pattern distribution (two levels: uniform distribution and multivariate normal distribution); (7) correlation among attributes (two levels: 0 and 0.3, representing zero correlation and moderate correlation, respectively); and (8) whether demographic covariates influence the attribute mastery pattern distribution (two levels: influence and no influence). It should be noted that when attribute mastery patterns follow a uniform distribution, it is not possible to directly manipulate the correlation among attributes or the influence of demographic covariates on the attribute mastery pattern distribution. Therefore, the manipulation of these two factors is only implemented under the condition that attribute mastery patterns follow a multivariate normal distribution. Specifically, when demographic covariates have no influence on the attribute mastery pattern distribution of the focal and reference groups, both groups' attribute mastery patterns are drawn from a multivariate normal distribution $MVN(\mathbf{0},\ )$; under the influence condition, the focal group' s attribute mastery patterns are drawn from $MVN(\mathbf{0},\ )$, while the reference group' s attribute mastery patterns are drawn from $MVN(\mathbf{0.5},\ )$.

After fully crossing the above eight manipulated factors, this study initially obtained $3 \times 2 \times 3 \times 2 \times 6 \times 2 \times 2 \times 2 = 1728$ experimental conditions. Since the condition of uniform attribute mastery pattern distribution could not be effectively combined with two of the manipulated factors (namely, correlation among attributes and whether demographic covariates influence attribute mastery pattern distribution), 648 invalid combinations related to these factors were eliminated, leaving 1080 valid simulation conditions. During the experiment, each condition was simulated 100 times. All simulation experiments were implemented in the R environment [**?**]. Additionally, this study set the following fixed conditions: (1) number of cognitive attributes, fixed at 5 following [**?**]; (2) test length, fixed at 30 items, which is common in cognitive diagnosis DIF detection research [**?**, **?**, **?**, **?**]; (3) number of attributes measured by a single item, limited to a maximum of 3 attributes per item following [**?**, **?**]; (4) DIF item proportion, fixed at 20% of total test length following [**?**], meaning 6 DIF items out of 30; (5) test Q-matrix, using the widely adopted 30-item Q-matrix from previous research [**?**, **?**, **?**], which includes 5 cognitive attributes and employs a balanced design where each attribute is measured by an equal number of items, and the numbers of items measuring 1, 2, and 3 attributes are also equal; (6) DIF item positions, following the practice of [**?**], items 6, 9, 12, 14, 24, and 25 were designated as DIF items across all experimental conditions; (7) number of covariates of interest, set at 2 following [**?**] and [**?**], denoted as x_1 and x_2, with each covariate containing two levels.

3.2 Simulation of Response Data with Different DIF Patterns and DIF Detection This study conducted simulation of response data with different DIF patterns and DIF detection in the R environment [**?**]. All code is available through the online OSF platform (link:

https://osf.io/7ykqj/?view_{only}=e873ab9c2d2c408385a90e945952f296). The specific operational steps are as follows: (1) Use the `simGDINA` function in the GDINA package [**?**] to generate examinee parameters and item parameters for the reference group, and simulate the reference group' s response data based on these. (2) Manipulate DIF and generate response data for the focal group. First, based on the correct response probabilities of the reference group examinees on each item $P_{jR}(\alpha_v)$, calculate the correct response probabilities for the focal group examinees on the same items $P_{jF}(\alpha_v)$ according to different DIF magnitudes. In this study, the calculation methods for $P_{jF}(\alpha_v)$ differ across different DIF patterns, as detailed in Table 1 . Then, use the obtained $P_{jF}(\alpha_v)$ to simulate the focal group examinees' response data. (3) After generating response data for both groups, apply the six DIF detection methods considered in this study (ISRPM, Wald, LR, Wald-FS, LR-FS, and MH) to conduct DIF detection and output corresponding results. It should be noted that none of the detection methods in this study implemented item purification procedures.

Table 1 DIF Simulation Methods Under Different DIF Patterns

DIF Pattern	Calculation Method for $P_{jF}(\alpha_v)$
Main effect DIF only	$P_{jF}(\alpha_v) = P_{jR}(\alpha_v) + z$
Interactive DIF only	$P_{jF}(\alpha_v) = P_{jR}(\alpha_v) + z \cdot I(x_1 \neq x_2)$
Both main effect and interactive DIF	$P_{jF}(\alpha_v) = P_{jR}(\alpha_v) + z + z \cdot I(x_1 \neq x_2)$

Note: z represents DIF magnitude, taking values of 0.05 and 0.1 in this study, representing small and large DIF, respectively.

3.3 Evaluation Metrics Previous cognitive diagnosis DIF studies have commonly used true positive rate (TPR) and false positive rate (FPR) to evaluate DIF detection method performance. TPR is equivalent to statistical power, i.e., the proportion of DIF items correctly identified; FPR is equivalent to Type I error rate, i.e., the proportion of DIF-free items incorrectly identified as DIF. However, calculating these two metrics only at the item level may not fully reflect the DIF detection performance of the proposed method or allow for reasonable comparison with traditional methods. This is because previous studies typically only considered DIF caused by a single covariate, where TPR and FPR could only evaluate detection performance on a single covariate. In contrast, this study further considers situations where multiple covariates simultaneously cause DIF. In this context, evaluating detection method performance should not only focus on whether DIF items can be correctly identified but also examine whether the method can further accurately locate the specific covariates causing

DIF after identifying DIF items. Therefore, in addition to calculating TPR and FPR at the item level, this study also calculates these metrics at the combined item-covariate level to more comprehensively evaluate the overall performance of DIF detection methods.

Let $\mathbf{DIF}_j = (DIF_{j1}, DIF_{j2}, \ldots, DIF_{jm})$ represent the DIF detection result vector for item j on m covariates, where $DIF_{jm} = 1$ indicates that item j has DIF on covariate m, while $DIF_{jm} = 0$ indicates no DIF. If any element in vector $\mathbf{DIF}_j$ is 1, then item j is a DIF item; if all elements in the vector are 0, i.e., $\mathbf{DIF}_j = \mathbf{0}$, then item j is a DIF-free item. Based on this, the calculation methods for each evaluation metric are as follows:

(1) Item-level TPR, denoted as TPR_I:

$$\mathrm{TPR}_I = \frac{\sum_{j=1}^{J} I(\mathbf{DIF}_j^{\text{detected}} \neq \mathbf{0} \mid \mathbf{DIF}_j^{\text{true}} \neq \mathbf{0})}{\sum_{j=1}^{J} I(\mathbf{DIF}_j^{\text{true}} \neq \mathbf{0})}$$

(2) Item-level FPR, denoted as FPR_I:

$$\mathrm{FPR}_I = \frac{\sum_{j=1}^{J} I(\mathbf{DIF}_j^{\text{detected}} \neq \mathbf{0} \mid \mathbf{DIF}_j^{\text{true}} = \mathbf{0})}{\sum_{j=1}^{J} I(\mathbf{DIF}_j^{\text{true}} = \mathbf{0})}$$

(3) Combined item-covariate level TPR, denoted as TPR_{IC}:

$$\mathrm{TPR}_{IC} = \frac{\sum_{j=1}^{J}\sum_{m=1}^{M} I(DIF_{jm}^{\text{detected}} = 1 \mid DIF_{jm}^{\text{true}} = 1)}{\sum_{j=1}^{J}\sum_{m=1}^{M} I(DIF_{jm}^{\text{true}} = 1)}$$

(4) Combined item-covariate level FPR, denoted as FPR_{IC}:

$$\mathrm{FPR}_{IC} = \frac{\sum_{j=1}^{J}\sum_{m=1}^{M} I(DIF_{jm}^{\text{detected}} = 1 \mid DIF_{jm}^{\text{true}} = 0)}{\sum_{j=1}^{J}\sum_{m=1}^{M} I(DIF_{jm}^{\text{true}} = 0)}$$

In the above formulas, $I(d)$ represents the indicator function, which equals 1 if condition d is satisfied, otherwise 0.

3.4 Experimental Results Due to space limitations, the main text only reports the overall performance of each cognitive diagnosis DIF detection method in terms of statistical power and Type I error rate under different experimental conditions when attribute mastery patterns follow a uniform distribution. The relevant results are shown in Figure 2 [Figure 2: see original paper] and Figure 3 [Figure 3: see original paper]; the complete numerical results are listed in Appendix Tables 1 and 2 . Furthermore,

when attribute mastery pattern distribution follows a multivariate normal distribution, the statistical power and Type I error rate results of each detection method under different experimental conditions are uniformly presented in Appendix Tables S1 through S8 in the online appendix (https://osf.io/7ykqj/?view_{only}=e873ab9c2d2c408385a90e945952f296). Overall, ISRPM demonstrates acceptable detection performance under most experimental conditions and outperforms the existing Wald, LR, FS-Wald, FS-LR, and MH methods. The following sections further analyze the effects of different manipulated variables on statistical power and Type I error rate.

3.4.1 Statistical Power Figure 2 shows the statistical power results of each cognitive diagnosis DIF detection method at the item level and the combined item-covariate level under different experimental conditions. However, when an item' s DIF is jointly caused by main effects and/or interactions of multiple covariates, relying solely on the item-level TPR_I metric may not fully reflect a method' s detection capability. Specifically, the TPR_I metric only measures whether a detection method can correctly identify DIF items without further considering which covariates cause these DIF effects. In other words, even if a method only identifies partial sources of DIF, TPR_I will still consider it a correct identification, potentially overestimating the method' s detection performance to some extent. In contrast, the TPR_{IC} metric not only examines whether DIF items are correctly identified but also further evaluates whether the detection method can accurately locate the covariates that induce DIF. Therefore, TPR_{IC} provides a more detailed evaluation perspective and better reflects detection method performance in multi-covariate DIF contexts.

The results in Figure 2 show that when interactive DIF exists in the test, ISRPM' s statistical power outperforms the existing Wald, LR, FS-Wald, FS-LR, and MH methods under most conditions. As sample size per group increases, the statistical power of all methods improves significantly, indicating that sample size is an important factor affecting DIF detection performance. When sample size per group is 2000, ISRPM' s overall performance in statistical power is superior to the other five traditional methods, especially under conditions where only interactive DIF exists. However, under the condition of 500 examinees per group, ISRPM' s statistical power is somewhat limited. Specifically, under this sample size condition, when the test contains only main effect DIF or both main effect DIF and interactive DIF, although all methods have relatively low statistical power, ISRPM' s overall performance is inferior to the other five methods. This result may be related to the small sample size allocated to child nodes during ISRPM' s recursive partitioning process, which reduces parameter estimation precision and consequently affects DIF detection accuracy. For example, with 500 examinees per group and two dichotomous covariates, ISRPM allocates 500 examinees to each child node in the first-level partition; if proceeding to the second-level partition, the sample size allocated to the four child nodes decreases to 250. In this situation, insufficient sample size may increase parameter estimation error and affect detection result accuracy. Notably, under

the same sample size condition, ISRPM' s detection results are relatively better when DIF is caused only by interactions among covariates. This may be related to the method' s ability to consider interactions among multiple covariates during analysis, thereby demonstrating higher sensitivity in identifying interactive DIF.

As DIF magnitude increases, the statistical power of all methods improves significantly. For example, as shown in Appendix Table 1, when the test contains only main effect DIF and sample size per group is 1000, ISRPM' s TPR_I values range between 0.53 and 0.59 under small DIF conditions, while this indicator rises to 0.98 under large DIF conditions. Similarly, under small DIF conditions, the TPR_I ranges for Wald, LR, FS-Wald, FS-LR, and MH methods are 0.62–0.69, 0.67–0.76, 0.64–0.73, 0.67–0.75, and 0.48–0.49, respectively, while under large DIF conditions, the ranges are 0.99–0.99, 1–1, 0.99–0.99, 1–1, and 0.96–0.96, respectively. Regarding DIF patterns, ISRPM demonstrates relatively higher statistical power in detecting interactive DIF. For example, Appendix Table 1 shows that when the test contains only interactive DIF and sample size per group is 100, ISRPM' s TPR_I range is 0.22–0.81, while the TPR_I ranges for Wald, LR, FS-Wald, FS-LR, and MH methods are 0.05–0.08, 0.08–0.10, 0.06–0.08, 0.08–0.11, and 0.07–0.09, respectively. In situations with only main effect DIF, ISRPM' s performance is similar to other methods under larger sample size conditions. However, when sample size per group decreases to 500, ISRPM' s overall performance in statistical power is inferior to the other five traditional methods. Overall, under conditions with sample sizes of 1000 or more per group, ISRPM' s overall performance in detecting main effect DIF is comparable to traditional methods while demonstrating higher statistical power in situations involving interactive DIF. Additionally, as item quality improves, the statistical power of all methods shows an upward trend. Finally, results from Tables S1, S3, S5, and S7 in the online appendix show that neither the correlation among attributes nor whether demographic covariates influence attribute mastery pattern distribution has a significant impact on ISRPM' s statistical power.

3.4.2 Type I Error Rate [**?**] pointed out that under a nominal significance level of α and n replications in simulation studies, due to sampling error, the observed Type I error rate may not be completely consistent with the nominal level, but there is a 95% probability that it falls within the interval $[\alpha - 1.96\sqrt{\alpha(1-\alpha)/n}, \alpha + 1.96\sqrt{\alpha(1-\alpha)/n}]$. Corresponding to this study $(\alpha = 0.05, n = 100)$, this interval is $[0.007, 0.093]$. If a detection method' s observed Type I error rate falls within this range, its control effect can be considered reasonable. As shown in Figure 3, ISRPM' s control of Type I error rate is relatively stable under most experimental conditions, with overall results falling within the reasonable range. Further comparison reveals that under most experimental conditions, ISRPM' s Type I error rate is lower than the other five traditional methods, indicating that in contexts involving multiple covariates for DIF detection, the new method' s control of Type I error is generally superior to traditional methods. For example, Appendix Table 2 shows that when

the test contains only main effect DIF, ISRPM' s FPR_I values range from 0.01 to 0.07, indicating that ISRPM' s Type I error control effect meets expectations; while the FPR_I ranges for Wald, LR, FS-Wald, FS-LR, and MH methods are 0.04–0.11, 0.11–0.17, 0.06–0.10, 0.11–0.14, and 0.09–0.30, respectively, with some conditions exceeding the upper limit of the reasonable interval (i.e., 0.093). Overall, ISRPM demonstrates high stability in Type I error rate control, with different DIF patterns and DIF magnitudes showing no significant effect on Type I error rates across detection methods. Additionally, the Type I error rates of all methods under high-quality item conditions are generally lower than under medium-quality item conditions, which may be related to higher parameter estimation precision under high-quality items. Finally, results presented in Tables S2, S4, S6, and S8 in the online appendix indicate that neither the correlation among attributes nor whether demographic covariates influence attribute mastery pattern distribution has a significant effect on the Type I error rates of any method.

[Figure 2: see original paper] Statistical power results of cognitive diagnosis DIF detection methods under different experimental conditions
[Figure 3: see original paper] Type I error rate results of cognitive diagnosis DIF detection methods under different experimental conditions

---

4. Empirical Study

4.1 Data Description The empirical analysis in this study uses the Diagnostic Classification Version of the Schizotypal Personality Questionnaire (DC-SPQ) developed by [**?**] as an example to demonstrate the feasibility of the proposed ISRPM in real test data and to compare its results with five existing cognitive diagnosis DIF detection methods internationally. The DC-SPQ contains 74 items, each using dichotomous scoring. This study adopts the test Q-matrix for DC-SPQ provided by [**?**] (see Table A1 in the original manuscript), which includes 9 attributes. The test data provided by [**?**] came from 980 university students from seven universities in three Chinese cities, with an average age of 20.5 years ($SD = 1.79$). For DIF detection, this study selected commonly used gender and household registration as covariates. Regarding gender distribution, female examinees accounted for 62.3% ($N = 611$); regarding household registration distribution, 43.1% ($N = 422$) of examinees were from urban areas.

4.2 DIF Detection Results [**?**] used the Wald test for DIF analysis of the DC-SPQ and set the overall significance level at 0.01. To more comprehensively compare the detection performance of different methods, this study used the proposed ISRPM and five existing cognitive diagnosis DIF detection methods (Wald, MH, LR, FS-Wald, and FS-LR) to analyze the DC-SPQ under both significance levels of 0.05 and 0.01. It should be emphasized that when implementing ISRPM, the search termination rules for optimal splitting covariates and splitting points were consistent with those in the simulation study.

Table 4 presents the DIF detection results for the DC-SPQ by different methods under the two overall significance levels (0.05 and 0.01). Overall, ISRPM shows high consistency with the other five traditional methods in detection results. For example, under the overall significance level of 0.05, items 24, 25, 26, 27, and 40 were all identified by all methods as potentially exhibiting main effect DIF caused by gender. However, the detection results for item 35 showed some differences across methods: ISRPM identified it as potentially exhibiting interactive DIF caused by the interaction between gender and household registration; LR and MH methods identified it as potentially showing only main effect DIF caused by gender; while Wald, FS-Wald, and FS-LR methods did not detect any DIF pattern for this item. Figure 4 [Figure 4: see original paper] shows the recursive partitioning tree structure output by ISRPM for item 35. The figure reveals that both gender-specific paths in the first level are dashed lines, indicating that this item may not exhibit main effect DIF on gender; in the second-level partition, the left branch remains a dashed arrow while the right branch is a solid arrow, indicating that females from different household registration areas show significant differences in response patterns on this item, while the between-group differences on the household registration variable are not significant for males. This result suggests that item 35 may exhibit interactive DIF caused by the interaction between gender and household registration. Although post-hoc interpretation of DIF detection results still involves some uncertainty, the empirical analysis results indicate that ISRPM demonstrates potential advantages in identifying interactive DIF. It should be noted that interpretation of DIF detection results should not rely solely on statistical tests. A more reasonable strategy is to combine statistical analysis with expert judgment: first use statistical methods to screen for items that may exhibit DIF, then have content domain experts conduct rigorous review of these items' content and wording. Through this combination of quantitative and qualitative analysis, test developers can more comprehensively evaluate the substantive significance of detection results, thereby providing more evidence-based references for test revision and quality control.

Table 2 DIF Detection Results for DC-SPQ Items by Different Methods at Overall Significance Level 0.05 (0.01)

Item	ISRPM	Wald	LR	MH	FS-Wald	FS-LR	First-Level Variable
24	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	Gender
25	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	Gender
26	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	Gender
27	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	Gender

Item	ISRPM	Wald	LR	MH	FS-Wald	FS-LR	First-Level Variable
35	√ (×)	× (×)	√ (×)	√ (×)	× (×)	× (×)	Gender
40	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	√ (√)	Gender

Note: This table only presents results for items where at least one method detected DIF. "First-Level Variable" refers to the first optimal splitting variable automatically identified by ISRPM for the current item, i.e., the first-level splitting variable in the recursive partitioning tree structure; different items may show different first-level optimal splitting variables depending on each variable's splitting effect. "√" indicates that the current method flagged the item as exhibiting DIF, while "×" indicates it flagged the item as DIF-free. Results outside parentheses are under overall significance level 0.05; results inside parentheses are under overall significance level 0.01.

[Figure 4: see original paper] Recursive partitioning tree for item 35 based on ISRPM output

5. Discussion and Conclusions

5.1 Discussion As emphasized in the *Standards for Educational and Psychological Testing* [**?**], to ensure the fairness and accuracy of test results, conducting differential item functioning (DIF) analysis has become an important component in evaluating test quality and validity during psychological and educational test development. Therefore, DIF detection remains an indispensable quality assessment step in cognitive diagnostic test development and validation. However, existing cognitive diagnosis DIF detection methods typically can only conduct detection for a single covariate in a single analysis, thus being limited to identifying main effect DIF. Although these methods are reliable in identifying main effect DIF, they have difficulty effectively detecting interactive DIF caused by interactions among multiple covariates, which directly affects the measurement fairness and validity of cognitive diagnostic assessments. To address this issue, this study combines cognitive diagnosis models with recursive partitioning techniques from data mining and proposes a new cognitive diagnosis DIF detection method, namely the Item-based Sequential Recursive Partitioning Method (ISRPM). This method aims to provide more effective methodological support for conducting main effect DIF and interactive DIF detection in cognitive diagnostic tests. Overall, the main contribution of this study is that the proposed ISRPM enables simultaneous processing of multiple covariates during cognitive diagnosis DIF detection and can effectively detect both main effect DIF and interactive DIF, thereby promising to further improve the detection accuracy of cognitive diagnosis DIF detection methods.

To evaluate the detection performance of ISRPM, this study conducted a series of Monte Carlo simulation experiments, focusing on two important aspects: (1) the effects of factors such as sample size per group, DIF magnitude, DIF pattern, item quality, correlation among attributes, and whether demographic covariates influence attribute mastery pattern distribution on the new method' s detection performance; and (2) comprehensive performance comparison results between the new method and five existing cognitive diagnosis DIF detection methods internationally (Wald, LR, FS-Wald, FS-LR, and MH). Additionally, to demonstrate and verify the feasibility of the new method in real testing environments, this study conducted an empirical analysis using the cognitive diagnostic version of the Schizotypal Personality Questionnaire as an example and compared its results with those of five existing DIF detection methods. The research results show that: (1) Sample size per group, DIF magnitude, and item quality are all important factors affecting ISRPM' s DIF detection performance; (2) When items exhibit interactive DIF, the proposed ISRPM demonstrates superior performance in statistical power and Type I error rate compared to traditional Wald, LR, FS-Wald, FS-LR, and MH methods; (3) In situations with only main effect DIF, when sample size reaches 1000 or more per group, ISRPM's detection performance is generally comparable to the other five DIF detection methods, while when sample size decreases to 500 per group, ISRPM' s performance is somewhat limited; (4) Neither the correlation among attributes nor the influence of demographic covariates on attribute mastery pattern distribution has a significant effect on ISRPM' s detection performance.

Through simulation and empirical studies, several advantages of ISRPM can be summarized: (1) It promises to further improve the DIF detection accuracy of cognitive diagnostic tests. Simulation results show that compared with existing detection methods, ISRPM generally has better performance in identifying DIF items and the covariates causing DIF. (2) It can simultaneously identify both main effect DIF and interactive DIF. This characteristic of the new method that can handle interactive DIF simultaneously promises to help researchers more comprehensively analyze the complex demographic sources that lead to measurement bias in tests. (3) It promises to provide more detailed DIF diagnostic information for test developers. Compared with traditional methods, ISRPM can not only accurately identify items exhibiting DIF but also reveal the effects of different covariates' main effects and their interactions on DIF, thereby providing more targeted references for item revision.

Based on the comprehensive results of simulation and empirical studies, this paper offers the following recommendations for potential users of ISRPM: (1) Regarding the choice of cognitive diagnosis model. This study verifies the applicability of ISRPM under saturated cognitive diagnosis models (i.e., the G-DINA model) through simulation and empirical analysis. It should be noted that this method can also be extended to several simplified forms of the G-DINA model, such as DINA, DINO, and ACDM, so researchers can flexibly choose according to research objectives and data characteristics. (2) Regarding the number and type of covariates. Simulation results show that ISRPM can support simulta-

neous DIF detection for two or more covariates. When research involves multiple covariates, ISRPM demonstrates good applicability. However, it should be noted that as the number of covariates increases, to maintain the stability and detection accuracy of the new method, users should increase sample size as much as possible. Therefore, in scenarios involving multiple covariates, users need to balance the relationship between sample size and the number of covariates. To further verify the new method' s performance under conditions with multicategorical and continuous covariates, this paper presents two supplementary experiments in the online appendix (see Supplementary Experiments 1 and 2 in the online appendix). The results show that ISRPM can still maintain acceptable detection performance under conditions with more than two covariates and demonstrates acceptable detection effects for different types of covariates. (3) Regarding sample size. Simulation results indicate that sample size is an important factor affecting ISRPM' s detection performance. When sample size per group reaches 1000 or more, ISRPM demonstrates relatively ideal detection accuracy under most conditions; when sample size decreases to 500 or less, its statistical power shows a downward trend, and overall performance is inferior to traditional methods. Therefore, in practical applications, this paper recommends ensuring sample size is no less than 1000 per group to obtain relatively robust detection results. (4) Regarding the choice of DIF detection method. When selecting appropriate DIF detection methods in cognitive diagnostic tests, key factors such as research objectives, sample size, and the number of covariates of interest should be comprehensively considered. Although ISRPM shows better performance in identifying interactive DIF, its performance is not superior to traditional methods in all DIF patterns. For research focusing only on main effect DIF with small sample sizes per group (e.g., less than 500), traditional methods (such as the Wald test) are more recommended; in situations with larger sample sizes per group ($ $1000) or when research simultaneously focuses on both main effect and interactive DIF, ISRPM is more recommended. It should be reminded that when sample size is small, the precision of DIF detection results may be limited regardless of the method used, so conclusions should not be based solely on statistical test results. (5) Regarding the handling of DIF items. When conducting DIF detection in cognitive diagnostic tests, researchers should not rely solely on statistical test results. A more robust approach is to combine statistical analysis with expert evaluation: first use ISRPM or other methods to screen for items that may exhibit DIF, then invite domain experts to review the content and wording of these items. Through this combination of quantitative analysis and qualitative judgment, it is possible to more accurately determine whether the detected DIF has practical significance, thereby providing more evidence-based decision support for test revision and quality control.

5.2 Distinctions and Connections Between Cognitive Diagnosis DIF Detection Methods and DIF Detection Methods Under CTT and IRT Frameworks In the field of test fairness research, DIF detection methods

mainly rely on three measurement frameworks: Classical Test Theory (CTT), Item Response Theory (IRT), and Cognitive Diagnosis (CD). DIF detection methods under these three frameworks share certain commonalities in core objectives, detection logic, and basic principles, but show obvious differences in theoretical foundations, equating 处理方法, and ability matching criteria.

First, CTT, IRT, and CD frameworks maintain consistency in the core objective of DIF detection: all aim to identify items that may cause measurement bias to safeguard test result fairness. Second, regardless of the measurement framework adopted, DIF detection revolves around a common logic: under the condition of controlling ability levels, examine whether there are systematic differences in response performance on the same item across different examinee groups. Finally, these methods all follow the "conditional matching" principle, i.e., comparison is made under the condition of controlling examinee ability levels rather than directly judging DIF based on raw scores. Specifically, CTT methods use observed total scores for matching, IRT methods use latent trait scores for matching, while CD methods use examinees' attribute mastery patterns for matching.

Despite maintaining consistency in core objectives, basic logic, and basic principles of DIF detection, CTT, IRT, and CD still show obvious differences in theoretical foundations, equating 处理方法, and ability matching criteria. First, from the perspective of theoretical foundations, CTT methods are based on classical measurement theory, IRT methods rely on item response theory, while CD methods are built upon cognitive diagnostic theory. Second, from the perspective of equating 处理方法, DIF detection under CTT and IRT frameworks usually requires equating processing to place ability scores from different groups on the same scale for fair comparison. In contrast, cognitive diagnosis models (CDMs) used in CD methods naturally possess parameter equivalence characteristics because CDMs mainly measure examinees' mastery of discrete cognitive attributes rather than locating their abilities on a continuous latent trait scale. Therefore, DIF detection under the CD framework usually does not require additional equating processing. Finally, regarding ability matching criteria, CTT methods match through observed total scores, IRT methods match through latent trait scores, while CD methods match based on examinees' mastery patterns of cognitive attributes.

5.3 Research Limitations and Future Directions Although this study has made substantial progress in main effect DIF and interactive DIF detection in cognitive diagnostic assessments, there are still some obvious limitations and areas for improvement, mainly including the following aspects:

(1) Similar to existing Wald, LR, FS-Wald, FS-LR, and MH methods, the ISRPM proposed in this study still has room for further improvement in DIF detection performance under small sample conditions. This means that the performance of ISRPM under small sample conditions needs further exploration and optimization.

(2) This study mainly focused on the application of ISRPM in dichotomously scored cognitive diagnosis models (i.e., the G-DINA model), but the performance of recursive partitioning technology in polytomously scored cognitive diagnosis models and multiple-strategy cognitive diagnosis models remains an unknown area. Future research could further attempt to extend ISRPM to DIF detection under polytomous cognitive diagnosis models and multiple-strategy models.

(3) Consistent with existing IRT framework-based DIF detection methods using recursive partitioning technology [**?**, **?**], ISRPM currently only supports binary splitting. This may limit its ability to identify complex patterns where DIF is exhibited across multiple intervals within a single covariate. Therefore, future research could attempt to further extend ISRPM to DIF detection methods capable of performing multivariate splitting to enhance its applicability in more complex testing situations.

(4) Although this study systematically evaluated ISRPM' s performance through simulation experiments, it is still necessary to recognize that there are certain differences between simulation studies and empirical studies. Simulation study conditions are carefully designed and allow for strict variable control, enabling examination of detection methods' rationality and validity under different test conditions. However, empirical data structures are usually more complex and may be affected by measurement error, heterogeneity of examinee characteristics, and potential confounding factors. Therefore, simulation study results should be regarded as preliminary validation of the new method' s DIF detection performance rather than direct evidence of its application effects in actual tests. Future research should further validate the applicability and robustness of ISRPM based on more types and larger sample sizes of empirical data.

5.4 Conclusions The main conclusions of this study are as follows:

(1) This study proposes a new method for simultaneously detecting main effect DIF and interactive DIF within the cognitive diagnosis framework, namely the Item-based Sequential Recursive Partitioning Method (ISRPM). This method achieves full integration of recursive partitioning technology and cognitive diagnosis models, enabling simultaneous processing of multiple covariates in a single detection and identifying both items exhibiting DIF and the covariates causing DIF.

(2) Overall, ISRPM demonstrates acceptable DIF detection performance. In detecting main effect DIF, its overall performance is largely consistent with existing Wald, LR, FS-Wald, FS-LR, and MH methods; while in detecting interactive DIF, ISRPM shows relatively superior identification capability.

(3) Comprehensive simulation results show that when sample size per group is large, DIF effects are large, and item quality is high, ISRPM can maintain

relatively ideal detection performance.

References

[?] American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: AERA Publications.

[?] Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. Hoboken: Wiley.

[?] Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507–526.

[?] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289–300.

[?] Belzak, W. C. (2023). The multidimensionality of measurement bias in high-stakes testing: Using machine learning to evaluate complex sources of differential item functioning. *Educational Measurement: Issues and Practice, 42*(1), 24–33.

[?] Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological methods, 25*(6), 1-15.

[?] Bollmann, S., Berger, M., & Tutz, G. (2018). Item-focused trees for the detection of differential item functioning in partial credit models. *Educational and Psychological Measurement, 78*(5), 781–804.

[?] Collins, P. H. (1990). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. UnwinHyman.

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

[?] de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development, 51*(4), 281–296.

[?] DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31a review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics, 26*, 979–1030.

[?] Finch, W. H., Hernández Finch, M. E., & French, B. F. (2015). Recursive partitioning to identify potential causes of differential item functioning in cross-national data. *International Journal of Testing, 16*(1), 21–53.

[?] Holland, P. W, & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

[?] Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651–674.

[?] Hou, L. (2013). *Differential item functioning assessment in cognitive diagnostic modeling: Applying the Wald test to investigate DIF in the generalized DINA model framework* (Unpublished doctoral dissertation), University of Delaware.

[?] Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98–125.

[?] Komboz, B., Strobl, C., & Zeileis, A. (2016). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement, 78*(1), 128–166.

[?] Leighton, J. P., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and Applications.* Cambridge, UK: Cambridge University Press.

[?] Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning* (Unpublished doctoral dissertation). University of Georgia.

[?] Li, L., Zhou, X., Huang, J., Tu, D., Gao, X., Yang, Z., & Li, M. (2020). Assessing kindergarteners' mathematics problem solving: The development of a cognitive diagnostic test. *Studies in Educational Evaluation, 66*, 100888.

[?] Li, X., & Wang, W.-C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*(1), 28–54.

[?] Liu, Y., Xin, T., Li, L., Tian, W., & Liu, X. (2016). An improved method for differential item functioning detection in cognitive diagnosis models: An application of Wald statistic based on observed information matrix. *Acta Psychologica Sinica, 48*(05), 588–598.

[?] Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1–26.

[?] Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement, 45*(1), 37–53.

[?] Magis, D., Béland, S., Tuerlinckx, F., & Boeck, P. d. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862.

[?] Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*(5), 1016–1031.

[?] Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

[?] Nichols, P.D., Chipman, S.F., & Brennan, R.L. (1995). *Cognitively diagnostic assessment* (1st ed.). Routledge.

[?] Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement, 44*(4), 267–281.

[?] R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[?] Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

[?] Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289–316.

[?] Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323–348.

[?] Sun, X., Liu, Y., Wang, S., Xin, T., Song, N., & Zhou, M. (2022). Using information matrix-based method to detect differential item functioning with multiple groups in cognitive diagnostic test. *Journal of Psychological Science, 45*(03), 710–717.

[?] Tan, Z., de La Torre, J., Ma, W., Huh, D., Larimer, M. E., & Mun, E.-Y. (2023). A tutorial on cognitive diagnosis modeling for characterizing mental health symptom profiles using existing item responses. *Prevention Science: The Official Journal of the Society for Prevention Research, 24*(3), 480–492.

[?] Tay, L., Huang, Q., & Vermunt, J. K. (2015). Item response theory with covariates (IRT-C): Assessing item recovery and differential item functioning for the three-parameter logistic model. *Educational and Psychological Measurement, 76*(1), 22–42.

[?] Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

[?] Tu, D., Cai, Y., Gao, X., & Wang, D. (2019). *Advanced cognitive diagnosis*. Beijing: Beijing Normal University Publishing Group.

[?] Tutz, G., & Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika, 81*(3), 727–750.

[?] Wang, D., Gao, X., Cai, Y., & Tu, D. (2019). Development of a new instrument for depression with cognitive diagnosis models. *Frontiers in Psychology, 10*, 1306.

[?] Wang, X. (2019). Development and verification of cognitive diagnostic test for cross-grade pupils' mathematics learning ability, *Chinese Exam, 8*, 71–78.

[?] Wang, Z., Guo, L., & Bian, Y. (2014). Comparison of DIF detecting methods in cognitive diagnostic test. *Acta Psychologica Sinica, 46*(12), 1923–1932.

[?] Xi, C., Cai, Y., Peng, S., Lian, J., & Tu, D. (2020). A diagnostic classification version of schizotypal personality questionnaire using diagnostic classification models. *International journal of methods in psychiatric research, 29*(1), e1807.

[?] Yuan, K. H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test. *Psychometrika, 86*(2), 345–377.

[?] Zhang, J. (2006). *Detecting differential item functioning using the Mantel-Haenszel procedure*. Unpublished manuscript.

Appendix Table 1 Statistical Power of Different Cognitive Diagnosis DIF Detection Methods Under Uniform Attribute Mastery Pattern Distribution

Sample Size	DIF Magnitude	DIF Pattern	Method	TPR_I	TPR_{IC}
500	0.05	Main only	ISRPM	0.53-0.59	0.48-0.55
500	0.05	Main only	Wald	0.62-0.69	0.58-0.65
...	...	...	...	...	...

Appendix Table 2 Type I Error Rates of Different Cognitive Diagnosis DIF Detection Methods Under Uniform Attribute Mastery Pattern Distribution

Sample Size	DIF Magnitude	DIF Pattern	Method	FPR_I	FPR_{IC}
500	0.05	Main only	ISRPM	0.01-0.07	0.02-0.08
500	0.05	Main only	Wald	0.04-0.11	0.05-0.12
...	...	...	...	...	...

Note: Complete tables are available in the online appendix.

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