A Style-Aware Polytomous Diagnostic Model for Individual Traits

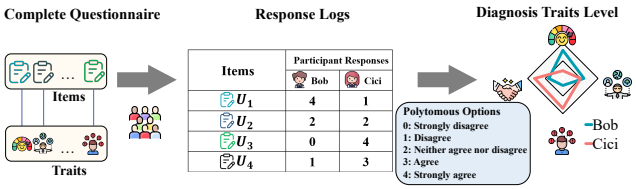
Yixuan Wang[a], Jiale Feng[a], Yue Huang[a], Xuruo Pan[a], Zhongjing Huang[b,*], Zhi Liu[a,c] and Hong Qian[a,d,*]

[a]Shanghai Institute of AI for Education, East China Normal University, Shanghai, China
[b]Faculty of Education, East China Normal University, Shanghai, China
[c]Shanghai Innovation Institute, Shanghai, China
[d]Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, East China Normal University, Shanghai, China

Abstract. Diagnostic models aim to precisely infer individuals' cognitive or non-cognitive competencies from their response logs, such as mathematical or social-emotional skills. While deep learning shows success in cognitive diagnosis, it remains underexplored in the equally important area of non-cognitive trait diagnosis. Accurate non-cognitive trait estimation is critical for individuals' development. Unlike cognitive assessments using right or wrong responses, non-cognitive trait assessments typically use subjective Likert-scale items with ordinal polytomous options to reflect latent trait levels. Furthermore, individual response styles, such as tendencies toward higher or lower options, introduce bias in trait inference, causing estimations that deviate from true trait levels. Thus, maintaining options ordinal semantic structure and mitigating the response style bias in trait estimation are two major challenges for accurate trait diagnosis. To address these issues, this paper proposes a Style-Aware Polytomous Diagnosis (SAPD) model. Specifically, to capture the ordinal semantics of response options, SAPD constructs an Ordinal Option Graph (OOG) that explicitly encodes the ordinal relationship among polytomous options, where higher options reflect higher latent trait levels. To mitigate the bias caused by individual response styles, we first design a Style-Aware Relational Graph (SARG), a heterogeneous graph that integrates multiple interactions among participants, items, options and traits, implicitly embedding response style information within node representations. We then propose a Response Style Corrector (RSC) that explicitly captures individual response tendencies and disentangles response style bias during trait diagnosis, allowing for dynamic and adaptive correction of trait levels. Extensive experiments on five real-world datasets show that SAPD improves accuracy by an average of 4% over competitive methods. Visualizations confirm SAPD effectively disentangles response style effects, leading to more accurate and interpretable trait diagnosis.

1 Introduction

Machine learning-based diagnostic models have rapidly gained significant attention [31, 7, 18, 3], becoming crucial in intelligent education research [25, 4]. By leveraging deep learning's powerful feature extraction capabilities [33], these models efficiently extract informative patterns from sparse data [20], capture complex interactions, and integrate multidimensional information [14]. Consequently, they substantially improve the efficiency, accuracy, granularity, and personalization of diagnostics. In educational and psychological assessment, machine learning methods now significantly surpass traditional diagnostic paradigms, playing an increasingly vital role [28, 16, 17].



(a) An illustrative example of polytomous trait diagnosis.

	Cognitive Diagnosis Model	Polytomous Trait Diagnosis Model
Assessment Scenario	Examination	Questionnaire
Assessment Objective	Knowledge Mastery Level	Latent Traits Level
Assessment Content	Cognitive Ability	Non-cognitive Ability, e.g., Optimism, Cooperation…
Rating Scale	0 or 1 (Dichotomous)	0, 1, 2, … , N (Polytomous Options)

(b) Dichotomous vs. polytomous diagnostic paradigms.

Figure 1. The Comparison of diagnostic models.

Recent advances in deep learning have significantly boosted the development of cognitive diagnosis models (CDMs) [18, 24, 29, 15], most of which focus on binary response data. The cognitive diagnosis infers the learners' proficiency in knowledge concepts from the response logs [29], aiming for fine-grained cognitive measurement. Typical Neural Cognitive Diagnosis Model (NCDM) [31], maps students and exercises to low-dimensional embeddings, enabling neural networks to capture complex interactions effectively. Inspired by NCDM, numerous deep learning CDMs have incorporated Graph Neural Networks (GNNs) to model deeper relational structures between students, exercises, and knowledge concepts. However, these models primarily target binary cognitive data.

In contrast, the diagnosis of individual non-cognitive traits, such as personality traits and social-emotional skills [9], remains equally important but relatively underexplored in deep learning research and faces unique challenges. Unlike cognitive tasks, non-cognitive trait assessments typically use subjective self-report questionnaires with

* Corresponding Authors.
Email: zjhuang@dedu.ecnu.edu.cn, hqian@cs.ecnu.edu.cn.

Likert-scale items [22], with the general diagnostic process illustrated in Figure 1(a). More critically, polytomous options in the form of the Likert scale [11] are fundamentally different from objective binary right or wrong data of cognitive diagnosis, as shown in Figure 1. The ordinal options of the Likert scale can reflect subtle changes in trait level but bring complex modeling challenges. ***The first key challenge is to accurately capture the ordinal relationships among polytomous options and map them onto continuous trait levels.*** The ordinal relationship between Likert-scale options (e.g., strongly disagree to strongly agree) encodes essential semantic information, reflecting the intensity of the underlying trait. To enhance trait diagnosis accuracy, it is crucial to effectively model and utilize these ordinal structure options, rather than treating them merely as discrete classes. ***The second key challenge is to disentangle individual response style biases from the estimation of true trait levels.*** In subjective self-report questionnaires, responses are influenced not only by the participant's actual traits but also by subjective tendencies and social desirability effects [2]. These may include tendencies to choose more favorable options (social desirability bias), or habitual preferences toward higher, lower, or midpoints (e.g., acquiescence, disacquiescence, or midpoint response styles) [35, 23]. This style interference is far more prevalent and severe in subjective self-report data than in objective exercise data. Without accounting for response style biases, the model confuses the response style with the true trait level, making the trait estimation not accurate enough [36]. Although recent work by Li et al. [16] proposed the PCDF model, which re-encodes polytomous responses into binary indicators using a cumulative category response function and applies binary cognitive diagnosis models, it does not explicitly model the ordinal structure of the data nor address the presence of response style bias. Consequently, the method does not fully address these core challenges, which constrains its effectiveness in accurate trait estimation.

To this end, this paper proposes the Style-Aware Polytomous Diagnosis (SAPD) model. To tackle the first challenge of capturing ordinal relationships among polytomous options, SAPD introduces the Ordinal Option Graph (OOG), which explicitly models the ordered structure of Likert-scale options and enables their alignment with continuous trait levels. To address the second challenge of disentangling response style bias, SAPD incorporates two components: the Style-Aware Relational Graph (SARG), which implicitly embeds individual style tendencies by modeling interactions among participants, items, options and traits. The Response Style Corrector (RSC) explicitly adjusts for individual response biases during inference, enabling more accurate trait estimation. Extensive experiments on five real-world datasets encompassing diverse trait dimensions validate the effectiveness of SAPD in improving prediction accuracy, mitigating response bias, and producing interpretable trait-level estimates.

The subsequent sections respectively review the related work, present the preliminaries, introduce the proposed SAPD, show the empirical analysis, and finally conclude the paper.

2 Related Work

Cognitive Diagnosis Models. Cognitive diagnosis models (CDMs) encompass a wide range of modeling approaches [20, 33, 38]. Traditional methods such as Item Response Theory [6] and Multidimensional IRT [30] leverage latent factor models to infer students' mastery level. With the advancement of deep learning, numerous neural CDMs have been proposed [14, 37, 21]. Models like NCDM [31] and KaNCD [32] replace traditional interaction functions with multilayer perceptrons to capture complex nonlinear interactions between students and exercises. SCD [28] incorporates the symbolic tree to explicably represent the complicated student-exercise interaction function and gradient-based optimization methods. Recent models [37, 7, 19] incorporate graph-based structures to represent student–exercise relations and apply graph neural networks to capture deeper structural dependencies. RCD [7] constructs hierarchical relationships of the student-exercise-concept graph and utilizes attention mechanisms to integrate node interactions. ORCDF [25] further introduces a response graph that encodes response signals as edge types, extracting signal-specific interaction patterns. Although these neural CDMs have shown promising results in cognitive diagnostic tasks, they are generally designed for binary response formats and lack the ability to generalize to more complex, subjective, and ordinal response data commonly found in non-cognitive assessments.

Polytomous Trait Diagnosis Models. With the widespread use of polytomous ordinal data in educational and psychological assessment [22, 9], researchers have developed diagnostic models tailored for ordinal responses [16]. Traditional statistical methods, such as the Graded Response Model [26], which is an extension of IRT [6], use cumulative logistic functions [27] to estimate the probability distribution over ordinal options. Multidimensional variants such as the Multidimensional Graded Response Model [10] aim to model multiple latent traits simultaneously. Although these methods perform well on small-scale, well-structured datasets, they often struggle with large-scale, high-dimensional data or highly interactive scenarios [1]. In response, Li et al. proposed PCDF [16], a general framework for splitting and merging polytomous data using Cumulative Category Response Function [27] theory. This framework can be seamlessly integrated with the existing dichotomous cognitive diagnostic model [31]. However, PCDF remains a data re-encoding approach and cannot achieve flexible diagnosis of different options on the model. Moreover it overlooks the role of individual response styles [2] during the assessment process, which significantly limits the precision and validity of the estimation of traits.

Response Style Correction. Response styles and their biases in trait diagnosis have been widely recognized in psychometrics and education [36]. The acquiescence response style, for example, can obscure participants' true trait levels due to the preference to agreeing responses [23]. Traditional correction methods typically employ statistical approaches [34]. Kreitchmann et al. [13] employed multidimensional IRT modeling to control response biases on Likert scales by introducing latent factors. However, this method was limited by strong assumptions and item imbalance, yielding only partial correction. Recently, the strength of deep learning in pattern recognition has encouraged researchers to explore more flexible modeling paradigms. Wang et al. proposed the Affective-Cognitive Diagnosis (ACD), integrating affective states (e.g., concentration, confusion) into cognitive modeling. ACD primarily targets binary cognitive tasks and does not directly address subjective response biases associated with polytomous Likert-scale data. Moreover, affective states and response styles are conceptually distinct and follow different underlying mechanisms [2]. Thus, effectively integrating response style estimation and correction into deep learning-based polytomous trait diagnosis remains a critical and underexplored challenge.

3 Preliminaries

In this section, we first formally introduce the fundamental elements of the polytomous trait diagnostic task. And then we introduce the formal problem definition of the polytomous trait diagnostic task.

Polytomous Trait Diagnostic. Suppose there are N participants,

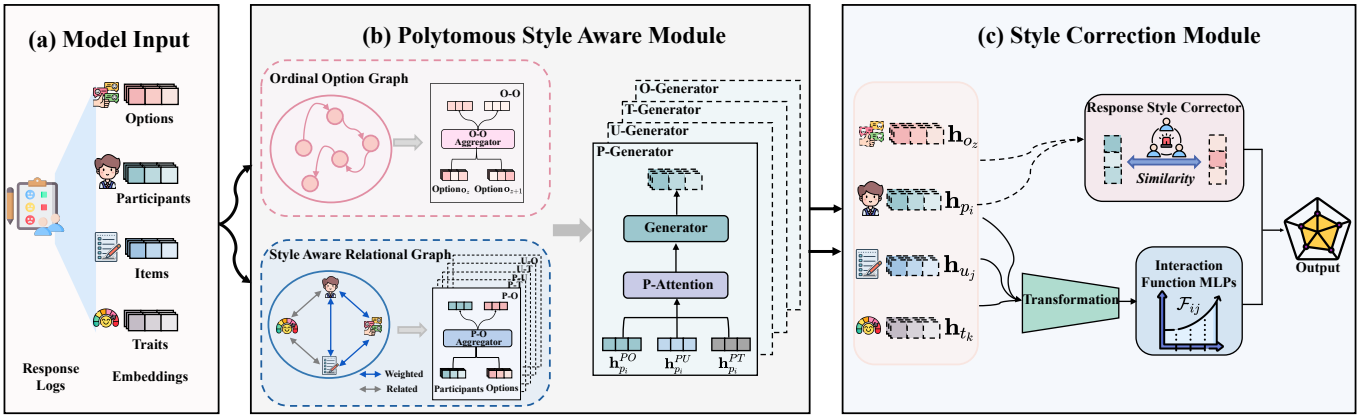


Figure 2. The procedure of the proposed Style-Aware Polytomous Diagnosis (SAPD) model.

M item units, Z options, and K trait dimensions involved in the trait assessment scenario. These can be represented as $P = \{p_1, \ldots, p_N\}$, $U = \{u_1, \ldots, u_M\}$, $O = \{o_1, \ldots, o_Z\}$ and $T = \{t_1, \ldots, t_K\}$, denoting the sets of participants, item units, response options, and trait dimensions, respectively. Each participant in the questionnaire assessment is required to complete a series of items. The corresponding response logs are denoted as a set of triplets $R = \{(p, u, o)|p \in P, u \in U, o \in O\}$, where o denotes the ordinal response option selected by the participant p for the item u. These options are typically represented as integer values ranging from 0, where higher values indicate stronger agreement or endorsement of the item's content. In addition, $\boldsymbol{Q}$ represents the relationship between items and traits, which can be regarded as a binary matrix $\boldsymbol{Q} = (\boldsymbol{Q}_{ij})_{M\times K}$, where $\boldsymbol{Q}_{ij} \in \{0, 1\}$ means whether u_i relates to t_j or not and $\boldsymbol{Q}_{ij}$ is the element in i-th row and j-th column of $\boldsymbol{Q}$.

Response Style. Response style refers to the individual-specific tendencies that participants exhibit when selecting the response options in self-report questionnaires. Common types of response styles include *acquiescence* (a tendency to agree with items), *disacquiescence* (a tendency to disagree), and *midpoint* (a preference for selecting middle options). In this paper, the response style of a participant p_i is formally defined as a latent vector $\boldsymbol{s}_{p_i} \in \mathbb{R}^Z$, where Z denotes the number of ordinal response options. $\boldsymbol{s}_{p_i}$ captures participants' characteristic preference distribution over polytomous responses.

Problem Definition. Given the observed triplet logs $R = \{(p, u, o)|p \in P, u \in U, o \in O\}$ and the trait mapping matrix $\boldsymbol{Q} \in \{0, 1\}^{M\times K}$, the goal of the polytomous trait diagnostic task is to infer the participants' trait levels $\boldsymbol{\theta} \in \mathbb{R}^{N\times K}$, while considering the influence of their latent response style $\boldsymbol{S} \in \mathbb{R}^{N\times Z}$.

4 Methodology: The Proposed SAPD Model

This section presents the Style-Aware Polytomous Diagnosis (SAPD) model. To model the ordinal nature of Likert-scale responses, SAPD introduces the Ordinal Option Graph (OOG), which explicitly encodes the ordered relationships among options. To address individual response style bias, SAPD incorporates two complementary components: the Style-Aware Relational Graph (SARG), which implicitly encodes personalized patterns through high-order interactions, and the Response Style Corrector (RSC), which explicitly adjusts for style bias during inference. We also detail the training objective, including the composite loss and optimization strategy. An overview of the SAPD architecture is shown in Figure 2.

In trait diagnosis, we utilize response logs and the item-trait association matrix. To represent the diagnostic process, we decompose each response entry into four components: participants, items, options and traits. Each component is encoded using trainable embeddings $\boldsymbol{H}_p \in \mathbb{R}^{N\times d}$, $\boldsymbol{H}_u \in \mathbb{R}^{M\times d}$, $\boldsymbol{H}_t \in \mathbb{R}^{K\times d}$, and $\boldsymbol{H}_o \in \mathbb{R}^{Z\times d}$. For example, $\boldsymbol{h}_{p_i} \in \mathbb{R}^{1\times d}$, denotes the row vector of i-th participant.

4.1 Polytomous Style Aware Module

Ordinal Option Graph. In polytomous diagnostic tasks, the questionnaire options typically exhibit an inherent ordinal structure, which directly correlates and implies a progressive relationship with participants' latent trait levels. For a given individual, selecting an option with a higher numerical label generally indicates a higher level of the associated latent trait. To effectively leverage this important information, this paper proposes the Ordinal Option Graph (OOG). Formally, based on the ordinal semantic order options, we construct a directed unweighted graph $\mathcal{G}_O = (\mathcal{V}_O, \mathcal{E}_O)$. The node set $\mathcal{V}_O = \{o_1, o_2, \ldots, o_Z\}$ corresponds to all Z ordinal response options. The edge set $\mathcal{E}_O$ captures the sequential relationships between adjacent options. Specifically, for each pair of adjacent options o_z and o_{z+1} $(1 \le z < Z-1)$, we add a directed edge $(o_z, o_{z+1}) \in \mathcal{E}_O$. This results in a chain-structured graph: $o_1 \to o_2 \to \cdots \to o_Z$. As an unweighted graph, all edge weights are implicitly set to 1. This structure directly encodes the ordinal relationship between options as graph connections, as shown in Figure 2(b).

To learn semantically enriched node representations over this structure, we design a unidirectional message-passing mechanism on the OOG. Each node aggregates information solely from its immediate predecessor, preserving the strict order of semantic flow. We initialize each node o_z with a trainable embedding and update the representations of the nodes through a layer-wise aggregation scheme. In the l-th layer, the embedding of node o_z is updated by combining its own representation and that of its predecessor o_{z-1} using an order-aware aggregator, i.e.,

$$\boldsymbol{h}_{o_z}^{OO^{(l+1)}} = \begin{cases} \text{AGG}\left(\left\{\boldsymbol{h}_{o_z}^{OO^{(l)}}\right\}\right), & \text{if } z = 1 \\ \text{AGG}\left(\left\{\boldsymbol{h}_{o_z}^{OO^{(l)}}, \boldsymbol{h}_{o_{z-1}}^{OO^{(l)}}\right\}\right), & \text{if } 1 < z \le Z \end{cases}. \quad (1)$$

For the starting node o_1, which has no incoming edge, the update relies solely on its own representation. The number of layers l determines the semantic receptive field of each node, i.e., how many

lower-level options a node can incorporate during message propagation. Given that the significance of neighbors from l gradually diminishes as l increases, we employ a descending accumulation method to integrate the aggregated outcomes from different l. It can be expressed as $\boldsymbol{h}_{o_z}^{OO} = \sum_{l=0}^{L} \frac{1}{l+1} \boldsymbol{h}_{o_z}^{OO^{(l)}}$, where L represents the predefined maximum receptive field.

Style-Aware Relational Graph. In polytomous diagnostic tasks, participants often exhibit response tendencies during their answering process. To more precisely characterize individual-level response preferences and model the high-order relational structure among items, response options and psychological traits, we propose the Style-Aware Relational Graph (SARG).

Formally, SARG is defined as a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node sets $\mathcal{V} = P \cup O \cup U \cup T$ comprise four types of entities: Participants (P), Options (O), Items (U) and Traits (T). The edge sets $\mathcal{E}$ include five types of relation-specific subgraphs: $\mathcal{E}_{PO}$, $\mathcal{E}_{PU}$, $\mathcal{E}_{UO}$, $\mathcal{E}_{PT}$, and $\mathcal{E}_{UT}$. Among these, $\mathcal{E}_{PO}$, $\mathcal{E}_{PU}$, and $\mathcal{E}_{UO}$ are weighted edges, capturing personalized response style biases such as a participant's preference for specific options, their tendency toward certain items, or the option distributions of particular items. In contrast, $\mathcal{E}_{PT}$ and $\mathcal{E}_{UT}$ are unweighted (related) edges, reflecting structural associations between participants and traits, as well as items and their associated trait dimensions. To enable effective message passing and representation learning on the SARG structure, we perform relation-specific aggregation operations on each subgraph. Concretely, for any node x, its embedding in layer $l+1$, denoted as $\boldsymbol{h}_x^{(l+1)}$, is updated by combining its own representation in layer l with aggregated information from its neighbors, as formulated below

$$\boldsymbol{h}_{p_i}^{PO^{(l+1)}} = \text{AGG}\left(\boldsymbol{h}_{p_i}^{PO^{(l)}}, \left\{ w_{po} \cdot \boldsymbol{h}_{o_z}^{PO^{(l)}} \mid o \in \mathcal{N}_{PO}(p_i) \right\}\right), \tag{2}$$

where w_{po} denotes the relative frequency with which participant p_i selects option o, and $\mathcal{N}_{\text{PO}}(p_i)$ represents the set of option nodes connected to p_i via $\mathcal{E}_{PO}$. Following the same pattern, we derive representations for other subgraphs. We employ the same descending accumulation method to integrate the aggregated outcomes from different l, which can be expressed as $\boldsymbol{h}_{p_i}^{PO} = \sum_{l=0}^{L} \frac{1}{l+1} \boldsymbol{h}_{p_i}^{PO^{(l)}}$.

To further adaptively integrate interaction features from multiple relational perspectives and achieve accurate estimation of both individual-specific style and latent trait levels, we introduce a weighted generation module based on an attention mechanism. This generator is designed to fuse heterogeneous information captured from different subgraphs, each reflecting a unique aspect of participant response behavior. For each participant p_i, we first obtain three intermediate embeddings: $\boldsymbol{h}_{p_i}^{PO}$, $\boldsymbol{h}_{p_i}^{PU}$, and $\boldsymbol{h}_{p_i}^{PT}$, which are derived from the subgraphs $\mathcal{G}_{PO}$, $\mathcal{G}_{PU}$, and $\mathcal{G}_{PT}$, respectively.

To determine the relative contribution of each view to the final representation, we compute attention scores using a trainable vector $\boldsymbol{a}_p \in \mathbb{R}^{1\times d}$ and trainable parameters $\boldsymbol{W}_p^g, \boldsymbol{b}_p^g \in \mathbb{R}^{d\times d}$, as follows

$$\alpha^{PO} = \boldsymbol{a}_p \cdot \tanh\left(\boldsymbol{h}_{p_i}^{PO}\boldsymbol{W}_p^g + \boldsymbol{b}_p^g\right), \tag{3}$$

$$\alpha^{PU} = \boldsymbol{a}_p \cdot \tanh\left(\boldsymbol{h}_{p_i}^{PU}\boldsymbol{W}_p^g + \boldsymbol{b}_p^g\right), \tag{4}$$

$$\alpha^{PT} = \boldsymbol{a}_p \cdot \tanh\left(\boldsymbol{h}_{p_i}^{PT}\boldsymbol{W}_p^g + \boldsymbol{b}_p^g\right). \tag{5}$$

The attention weights are then normalized using the softmax function, obtaining $\tilde{\alpha}^{PO}$, $\tilde{\alpha}^{PU}$, and $\tilde{\alpha}^{PT}$. Then, the participant's unified representation $\boldsymbol{h}_{p_i}$ is generated from the three views using their corresponding attention weights.

$$\boldsymbol{h}_{p_i} = \tilde{\alpha}^{PO} \cdot \boldsymbol{h}_{p_i}^{PO} + \tilde{\alpha}^{PU} \cdot \boldsymbol{h}_{p_i}^{PU} + \tilde{\alpha}^{PT} \cdot \boldsymbol{h}_{p_i}^{PT}. \tag{6}$$

This attention-guided fusion process enables the model to selectively emphasize the most informative relational features, enhancing both response style perception and latent trait estimation.

4.2 Style Correction Module

Based on the informative trait and option embeddings obtained by SARG, we further propose a Response Style Corrector (RSC) to explicitly mitigate response style bias in prediction.

Trait-Aware Interaction Function. Specifically, in the prediction phase, we first apply a trait-aware interaction function that simulates the interaction between participants and items to predict the probability that a participant p_i selects a specific option o_z for a given item u_j. The interaction function is based on specific trait patterns, where the number of dimensions equals the number of traits being assessed. This paper employs simple linear transformations to map the features $\boldsymbol{h}_{p_i}$, $\boldsymbol{h}_{u_j}$, and $\boldsymbol{h}_{t_z}$ from dimension d to K.

$$\tilde{\boldsymbol{h}}_{p_i} = \boldsymbol{h}_{p_i}\boldsymbol{W}_p + \boldsymbol{b}_p\,, \tilde{\boldsymbol{h}}_{u_j} = \boldsymbol{h}_{u_j}\boldsymbol{W}_u + \boldsymbol{b}_u\,, \tilde{\boldsymbol{h}}_{t_z} = \boldsymbol{h}_{t_z}\boldsymbol{W}_t + \boldsymbol{b}_t\,, \tag{7}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable parameters. Subsequently, we feed the transformed participant representation $\tilde{\boldsymbol{h}}_{p_i}$ and item representation $\tilde{\boldsymbol{h}}_{u_j}$ into the interaction function. The Interaction Function (IF) is designed to predict the likelihood that a participant selects a specific option for a given item, which can be formulated as

$$\mathcal{F}_{ij} = \text{sigmoid}\Big(\text{MLP}\big((\tilde{\boldsymbol{h}}_{p_i} - \tilde{\boldsymbol{h}}_{u_j}) \odot \boldsymbol{Q}_{u_j}\big)\Big), \tag{8}$$

where $\boldsymbol{Q}_{u_j} \in \mathbb{R}^{1\times K}$ signifies the traits associated with the j-th item, and $\odot$ denotes element-wise multiplication. $\mathcal{F}_{ij}$ constitutes the basic prediction of the option scores for participant i on item j b.

Response Style Corrector. Considering that many participants exhibit response tendencies independent of their true target traits, Response Style Corrector (RSC) derives a style interaction score vector by measuring the similarity between the participant embedding and the option embedding matrix $\boldsymbol{H}_O = [\boldsymbol{h}_{o_1}, \ldots, \boldsymbol{h}_{o_Z}]$,

$$\boldsymbol{s}_{p_i} = \boldsymbol{h}_{p_i}\boldsymbol{H}_o^{\top} \in \mathbb{R}^Z. \tag{9}$$

Each component $\boldsymbol{s}_{p_i}$ quantifies the participant's content-independent preference for option. $\boldsymbol{s}_{p_i}$ is then additively combined with the logits of interaction function to produce the final prediction. This additive operation preserves the discriminative power of the trait-aware interaction while explicitly compensating for the bias introduced by individual response styles, thereby improving both the accuracy and robustness of latent trait level estimation.

4.3 Model Training

During the training phase of the polytomous trait diagnosis task, we optimize the model using a joint loss function. As the primary objective, we adopt the Categorical Cross-Entropy Loss (CE) to measure the discrepancy between the predicted probability distribution $\hat{y}_{ij}$ and the ground truth label y_{ij}. The loss is formally defined as

$$\mathcal{L}_{CE} = -\frac{1}{N'}\sum_{i=1}^{N'}\sum_{j=1}^{Z} y_{ij}\log(\hat{y}_{ij}), \tag{10}$$

where N' denotes the number of samples, Z is the number of response options. Notably, the options are not merely mutually exclusive categories, but are inherently ordered, reflecting ascending or

descending levels of latent traits. To capture this ordinal structure, we incorporate an auxiliary supervision signal via the Earth Mover's Distance Loss, which quantifies the cumulative discrepancy between the predicted and true distributions. The EMD loss is defined as

$$\mathcal{L}_{EMD} = \frac{1}{N'} \sum_{i=1}^{N'} \sum_{j=1}^{Z} |\text{CDF}(\hat{y}_i)_j - \text{CDF}(y_i)_j| \ , \qquad (11)$$

where $\text{CDF}(\cdot)$ denotes the cumulative distribution function, designed to capture the global shift between distributions. The final training objective is a weighted sum of the above two components $\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{EMD}}$, where α is a tunable hyperparameter that balances the contribution of the EMD term during joint optimization. By minimizing this composite loss, the model is encouraged not only to accurately estimate each participant's trait level but also to align the predicted probability distributions with the ordinal structure inherent in the response options, thereby enhancing the consistency and interpretability of the inferred trait representations.

5 Experiments

In this section, we first present the construction of five real-world datasets and detail the evaluation metrics used. We then conduct extensive and rigorous experiments on these datasets to answer the following research questions. The code and appendix of this paper are available at https://github.com/yxwang19/SAPD. The repository also provides a lightweight LightGCN variant of SAPD.

•**Q1:** How does SAPD compare to existing methods in predicting participants' response options?

•**Q2:** To what extent does each component contribute to the performance of SAPD?

•**Q3:** Does SAPD offer interpretability in diagnosing participants' trait states?

•**Q4:** How do hyperparameters influence SAPD?

•**Q5:** Can SAPD effectively disentangle participants' response styles from their latent trait levels?

5.1 Dataset Description

This paper conducts extensive experiments on five real-world datasets: Suzhou, Houston, Moscow, BIG5, and EQSQ, each encompassing diverse trait dimensions. These datasets span both Organization for Economic Co-operation and Development (OECD) and Open-Source Psychometrics Project (OSPP), covering diverse psychological constructs and participant demographics. Detailed statistics for each dataset are summarized in Table 1.

Table 1. Statistics of real-world datasets for experiments.

Dataset Source	OECD			OSPP	
Dataset	**Suzhou**	**Houston**	**Moscow**	**BIG5**	**EQSQ**
#Participants	5,354	5,316	6,025	14,666	10,888
#Items	120	120	120	50	120
#Traits	15	15	15	5	2
#Traits per item	1	1	1	1	1
#Options	5	5	5	5	4
#Response logs	642,480	637,920	723,000	733,300	1,306,560

Suzhou. This dataset originates from the Social and Emotional Skills Survey (SSES) conducted by the OECD in Suzhou, China. After filtering out incomplete or invalid responses, it includes complete questionnaire records for 5,354 participants aged 10 and 15. Each participant completed a questionnaire using 5-point Likert scale.

Houston. This dataset is collected from the same OECD SSES initiative, conducted in Houston, United States. Following a preprocessing protocol consistent with the Suzhou dataset, we preserved 5,316 complete questionnaires after removing incomplete responses.

Moscow. The Moscow dataset targets adolescents in Russia and aims to explore cross-national differences and commonalities in social-emotional competencies. After processing, a total of 6,025 high-quality responses were retained.

BIG5. The BIG5 dataset is sourced from the Big Five Personality Test hosted by OSPP. It includes 50 items rated on a 5-point Likert scale. To improve age diversity, we expanded the participant age range to 10-50 years, yielding 14,666 valid responses.

EQSQ. The EQSQ dataset is based on the Empathizing–Systemizing Test, consisting of 120 items rated on a 4-point Likert scale. Similarly, we restricted the participant age range to 10–50 years, ultimately collecting 10,888 valid samples that profile individual tendencies across empathizing and systemizing dimensions.

5.2 Experimental Setup

This subsection presents the experimental setup, including comparison methods, evaluation metrics, and implementation details.

Compared Methods. Existing methods for polytomous trait diagnosis often adapt CDMs through specific data transformation strategies. To evaluate the effectiveness of SAPD, we construct a comprehensive set of compared methods by combining five representative CDMs, IRT, MIRT, NCDM, KaNCD and RCD, with three widely used adaptation strategies: Linear-Split (LS), One-vs-All (OVA), and PCDF. Each CDM is paired with all three strategies, resulting in a total of 15 comparison methods.

Linear-Split (LS): Converts ordinal labels to continuous values, trains a CDM to predict these values, and linearly maps the predicted scores back into discrete ordinal categories.

One-vs-All (OVA): Transforms each polytomous option into Z binary classification labels, each indicating whether the response matches a specific label. The final prediction is derived by aggregating the binary output.

PCDF: Based on the Cumulative Category Response Function framework, this method partitions data according to cumulative ordinal levels, allowing CDMs to process polytomous responses.

Evaluation Metrics. Evaluating trait diagnosis models is challenging, as individuals' true trait levels are unobservable. Following standard psychometric practice [5], the effectiveness of the model is assessed indirectly through its ability to predict responses to unseen items. SAPD adopts this widely accepted validation paradigm by comparing prediction accuracy in held-out test sets against competitive methods, demonstrating the rationality of its modeling. Its ability to accurately recover participants' underlying traits and response styles contributes directly to its superior predictive performance. Moreover, visualizations of learned style and trait representations provide empirical evidence that SAPD disentangles response styles from trait estimates. A complementary case study further confirms its ability to deliver accurate and interpretable diagnoses.

To comprehensively evaluate the performance of SAPD in polytomous option prediction, we adopt three commonly used evaluation metrics: Accuracy (ACC), Weighted Accuracy (WACC), and Mean Absolute Error (MAE). Specifically, ACC measures the overall proportion of correctly predicted options. WACC is based on a weighted classification approach and is more suitable for assessing prediction performance in polytomous data, as it better reflects the severity of misclassifications. MAE quantifies the average absolute deviation

Table 2. Overall predict performance in polytomous trait diagnosis scenario. In each column, an entry with the best mean value is marked in bold and underline for the runner-up. The standard deviation is not shown in the table since it is very small (less than 0.01). If the mean value of the best model significantly differs from the runner-up, passing a t-test with a significance level of 0.05, then we denote it with "*" at the corresponding position.

	Suzhou			Houston			Moscow			BIG5			EQSQ		
Methods	ACC (↑)%	WACC (↑)%	MAE (↓)	ACC (↑)%	WACC (↑)%	MAE (↓)	ACC (↑)%	WACC (↑)%	MAE (↓)	ACC (↑)%	WACC (↑)%	MAE (↓)	ACC (↑)%	WACC (↑)%	MAE (↓)
IRT-LS	40.90	66.34	0.7717	40.27	66.12	0.7705	41.66	67.47	0.7280	29.93	59.14	0.9505	37.47	65.83	0.7416
IRT-OVA	36.44	61.81	0.9347	36.15	62.58	0.8870	36.76	63.48	0.8848	31.85	57.57	1.0689	37.89	65.15	0.7737
IRT-PCDF	41.78	65.80	0.8118	41.14	65.83	0.8006	42.53	67.32	0.7472	32.10	57.40	1.0788	38.29	65.51	0.7635
MIRT-LS	43.19	67.79	0.7355	42.75	67.70	0.7321	45.63	69.99	0.6658	41.52	66.51	0.7860	40.55	66.56	0.7060
MIRT-OVA	37.82	62.14	0.9404	36.81	62.17	0.9192	37.99	63.78	0.8519	31.92	56.97	1.1025	39.90	65.94	0.7619
MIRT-PCDF	48.18	69.16	0.7525	47.37	68.99	0.7478	48.66	70.74	0.6756	40.24	63.62	0.8967	41.07	66.89	0.7360
NCDM-LS	45.72	69.02	0.7149	45.60	69.18	0.7051	47.99	71.13	0.6458	39.22	64.68	0.8257	37.90	65.96	0.7406
NCDM-OVA	36.72	61.78	0.9391	36.25	62.77	0.8797	37.05	63.69	0.8414	32.84	58.08	1.0498	38.97	65.81	0.7580
NCDM-PCDF	46.13	69.06	0.7165	46.18	69.47	0.7013	48.38	71.37	0.6373	39.75	63.78	0.8842	38.23	65.03	0.7821
KaNCD-LS	45.97	69.39	0.7002	45.68	69.46	0.6920	48.26	71.53	0.6309	40.67	66.04	0.7813	40.44	67.61	0.7001
KaNCD-OVA	36.15	61.49	0.9457	36.02	62.43	0.8916	36.47	63.02	0.8629	31.88	57.44	1.0617	38.96	65.29	0.7775
KaNCD-PCDF	46.57	69.37	0.7008	46.69	69.68	0.6952	48.84	71.67	0.6294	40.66	64.29	0.8669	39.51	66.29	0.7457
RCD-LS	41.04	66.48	0.7659	40.35	66.23	0.7671	42.56	68.06	0.7130	32.69	61.13	0.8973	37.38	65.83	0.7408
RCD-OVA	36.65	62.32	0.9085	36.36	62.80	0.8788	36.80	63.60	0.8403	32.22	57.98	1.0582	38.15	65.42	0.7668
RCD-PCDF	42.29	66.18	0.8005	41.68	66.11	0.7856	42.83	67.74	0.7291	34.98	60.20	0.9843	38.40	65.57	0.7618
SAPD	**51.03***	**71.07***	**0.6960***	**51.13***	**71.32***	**0.6835***	**52.17***	**73.20***	**0.6082***	**44.78***	**67.36***	**0.7801***	**47.72***	**70.72***	**0.6597***

Table 3. Overall prediction performance of ablation study on five datasets. Details are as same as Table 2.

	Suzhou			Houston			Moscow			BIG5			EQSQ		
Methods	ACC (↑)%	WACC (↑)%	MAE (↓)	ACC (↑)%	WACC (↑)%	MAE (↓)	ACC (↑)%	WACC (↑)%	MAE (↓)	ACC (↑)%	WACC (↑)%	MAE (↓)	ACC (↑)%	WACC (↑)%	MAE (↓)
SAPD-w/o-RSC	48.34	70.29	0.6970	48.42	70.44	0.6879	50.10	72.30	0.6219	42.74	66.86	0.7832	42.76	68.60	0.6835
SAPD-w/o-OOG	50.64	70.78	0.7025	50.84	71.07	0.6924	51.88	72.77	0.6217	44.30	66.71	0.8072	47.20	70.11	0.6651
SAPD-w/o-SARG	50.54	70.68	0.7081	50.76	71.13	0.6874	51.75	72.69	0.6184	44.03	66.36	0.8210	47.49	70.21	0.6667
SAPD	**51.03**	**71.07**	**0.6960**	**51.13**	**71.32**	**0.6835**	**52.17**	**73.20**	**0.6082**	**44.78**	**67.36**	**0.7801**	**47.72**	**70.72**	**0.6597**

between predicted and true labels, making it particularly appropriate for ordinal prediction tasks where maintaining label order is crucial.

Implementation Details. During training, all model parameters are initialized using Xavier normal initialization [8] and optimized with the Adam optimizer [12] with a learning rate set to 0.0003. The response data from each participant was randomly split into training and test sets in an 8:2 ratio. All experiments are independently repeated five times using different random seeds.

5.3 Experimental Results and Analysis

We conduct comprehensive experiments to analyze the experimental results and address the aforementioned research questions.

Participant Performance Prediction (To Q1). Table 2 presents the comparison results between our proposed SAPD method and the 15 baseline methods on five real-world datasets, evaluated using ACC (↑), WACC (↑), and MAE (↓). The best performance for each metric and dataset is highlighted in bold. Due to the typically low standard deviation of the compared methods, we do not report them in the table. As shown, SAPD consistently achieves the optimal results on all five datasets compared to all compared methods, thereby thoroughly validating its effectiveness in predicting participant performance in polytomous trait diagnosis tasks. Specifically, SAPD consistently outperforms the second-best-performing method, achieving an average accuracy improvement of approximately 4% across the five datasets. Among the compared methods, the OVA strategy methods generally exhibit the poorest performance, suggesting that simply decomposing polytomous labels into multiple binary classification tasks leads to a significant loss of ordered information. In contrast, SAPD effectively addresses this issue by explicitly modeling option ordering relationships and response style bias, thereby capturing participant-specific factors more accurately. This enables the model to improve both prediction accuracy and ordinal consistency, maintaining leading performance across all evaluation metrics.

Ablation Study (To Q2). To quantify the contribution of each core component to SAPD, we created three ablated variants, SAPD-w/o-OOG, SAPD-w/o-SARG, and SAPD-w/o-RSC, removing OOG, SARG and RSC, respectively. Table 3 reports the performance of these ablated variants and the entire SAPD model in terms of ACC, WACC, and MAE. Among the variants, removing RSC results in the most substantial performance drop: ACC and WACC drop by an average of 2% across all datasets, while MAE increases accordingly. This highlights the importance of disentangling the response style bias for accurate trait estimation. Excluding OOG, which disregards the ordinal nature of options, reduces ACC and WACC by an average of 0.4% and also worsens MAE, confirming the necessity of modeling option order. When SARG is removed, ACC and WACC drop by around 0.6%, indicating that high-order relational structures provide complementary cues that enhance representation robustness. Overall, the entire SAPD model, which keeps OOG, SARG, and RSC, achieves the best scores on all metrics. The results demonstrate that each component plays a distinct and complementary role, and all are essential to achieve optimal performance.

Interpretability Performance (To Q3). To assess the interpretability of SAPD in estimating latent traits and disentangling style bias, we conducted related visualization experiments on BIG5 dataset, which has the largest sample size in five datasets. Figure 3 shows the population-level distributions of two randomly selected traits, Openness and Extroversion, estimated by SAPD. Trait levels are visualized as normalized histograms overlaid with Gaussian density curves. Both distributions exhibit bell-shaped patterns, with most individuals centered around medium levels and fewer at the extremes. This aligns with the common assumption of normally distributed traits in psychometrics, indicating that SAPD yields psychologically plausible trait estimations after correcting for response style bias. Figure 4(a) further investigates the response style representations learned by SAPD. We assign each participant a label based on their response, and apply t-SNE to reduce the dimensionality of the inferred response style obtained by SAPD. The three style groups of disacquiescence, midpoint, and acquiescence exhibit a discernible pattern from left to right, following the semantic progression of Likert options from low to high. This supports SAPD's effectiveness in capturing ordinal structure among response options. To evaluate the disentanglement between style and trait, Figure 4(b) focuses on participants labeled with the midpoint style. Their estimated trait embeddings are visualized via t-SNE, revealing a continuous distribu-

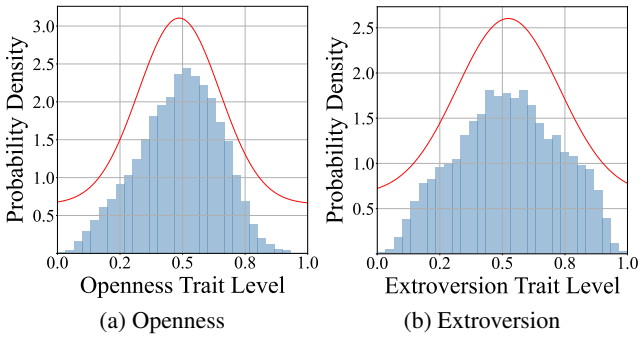


Figure 3. Population-level trait distributions inferred by SAPD.

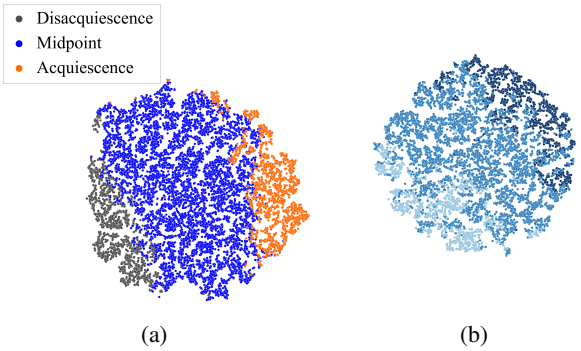


Figure 4. *t*-SNE visualizations of response style and trait inferred by SAPD on BIG5 dataset. (a) Response style visualization of all participants. (b) Trait level visualization of Midpoint Response Style participants.

tion. This indicates that even within a single response style group, SAPD preserves meaningful variation in trait estimates, demonstrating its ability to effectively disentangle response style from latent traits. Additional visualizations are provided in Appendices 1–3.

Hyperparameter Analysis (To Q4). We analyze two key hyperparameters: the number of graph aggregation layers l, and the weighting coefficient α for $\mathcal{L}_{EMD}$. The parameter l determines the receptive field of each node that how many neighbors are incorporated during message passing. As shown in Figure 5, we evaluate $l \in \{1, 2, 3, 4\}$ across all datasets. The best ACC and WACC are consistently achieved at $l = 2$, while larger l introduces noise from distant nodes and degrades accuracy. We also analyze the impact of α, which balances the contribution of $\mathcal{L}_{\text{EMD}}$ in the joint objective. We vary α over {0.0, 0.2, 0.4, 0.6, 0.8, 1.0} and report the results in Figure 6 on BIG5 dataset. Performance peaks at α = 0.8 for ACC, WACC and MAE, demonstrating that appropriately emphasizing the ordinal consistency leads to more accurate predictions. Additional results on other datasets are provided in Appendix 4.

Case Study (To Q5). To illustrate SAPD's diagnostic effectiveness and its ability to correct response style bias in real-world settings, we conducted a case study on Houston dataset. This dataset originates from the OECD assessment and includes expert evaluations as ground truth. As shown in Figure 7(a), we compare the estimated trait levels across five social-emotional dimensions using radar charts. The gray area denotes the expert evaluation, while the red and blue curves correspond to the estimation of SAPD and a PCDF variant adapted to NCDM, respectively. SAPD clearly aligns more closely with "ground truth" across all dimensions, while PCDF shows greater deviations due to its inability to account for response style effects. Figure 7(b) further quantifies this comparison using bar graphs of absolute deviations from expert scores. SAPD consistently

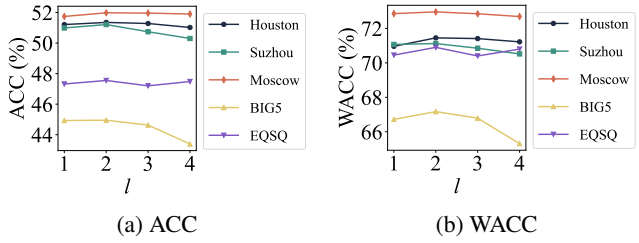


Figure 5. ACC and WACC performance under different l for each dataset.

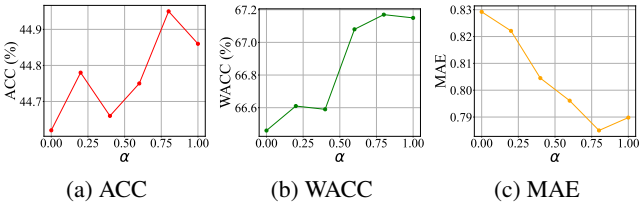


Figure 6. ACC, WACC, and MAE performance under different values of α on BIG5 dataset.

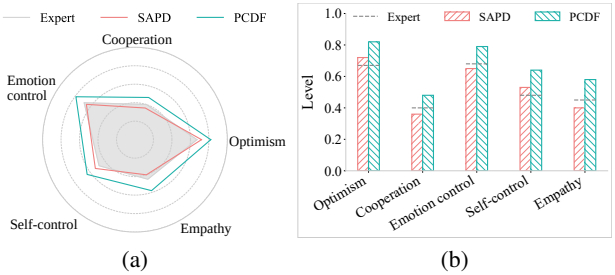


Figure 7. Case Study of the typical participant on Houston dataset.

yields the smallest deviation, indicating a superior estimation accuracy. These results demonstrate that SAPD effectively disentangles the response style bias from the trait inference, leading to a more accurate and reliable polytomous trait diagnosis. Appendix 5 provides an additional dis-acquiescence case on Houston dataset.

6 Conclusion

This paper introduces the SAPD model, which accurately infers participants' latent trait levels by disentangling response-style biases. To model the ordinal structure of Likert-scale options, SAPD constructs an Ordinal Option Graph (OOG) that preserves option ordering in representation learning. To reduce bias from individual response styles, it incorporates a Style-Aware Relational Graph (SARG) to capture personalized response patterns, and a Response Style Corrector (RSC) to adjust for stylistic deviations during inference. Extensive experiments on five real-world datasets demonstrate that SAPD consistently outperforms state-of-the-art methods in terms of accuracy, weighted accuracy, and mean absolute error, further validating its effectiveness in mitigating response style bias and enhancing trait estimation. As social-emotional skills continue to gain importance in education, mental health, and long-term development, improving the accuracy and interpretability of trait diagnostic models like SAPD becomes increasingly essential. In future work, we will design quantitative interpretability metrics to systematically evaluate model explanations and further enhance diagnostic validity.

Ethical Statement

The research presented does not involve human subjects or raise concerns related to privacy, security, or legal compliance. The datasets used in this study are from publicly available datasets that have undergone rigorous ethical review and anonymization during their release. These datasets contain no private student information and are widely used in various educational research studies.

Acknowledgements

We would like to thank the anonymous reviewers for constructive comments. This work is supported by the National Natural Science Foundation of China (No. 62476091), the Key Program in Education of the National Social Science Fund of China (No. ABA220028), the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science–MOE (No. KLATASDS2508), and the General Program in Education of the National Social Science Fund of China (No. BEA230071).

References

[1] D. Andrich. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573, 1978.

[2] U. Böckenholt. Measuring response styles in likert items. *Psychological Methods*, 22(1):69, 2017.

[3] X. Chen, L. Wu, F. Liu, L. Chen, K. Zhang, R. Hong, and M. Wang. Disentangling cognitive diagnosis with limited exercise labels. In *Advances in Neural Information Processing Systems*, pages 18028–18045, New Orleans, LA, USA, 2023.

[4] Z. Dong, J. Chen, and F. Wu. Llm-driven cognitive diagnosis with solo taxonomy: A model-agnostic framework. *Frontiers of Digital Education*, 2(2):20, 2025.

[5] S. E. Embretson and S. P. Reise. *Item Response Theory for Psychologists*. Psychology Press, 2013.

[6] S. E. Embretson and S. P. Reise. *Item Response Theory*. Psychology Press, 2013.

[7] W. Gao, Q. Liu, Z. Huang, Y. Yin, H. Bi, M.-C. Wang, J. Ma, S. Wang, and Y. Su. RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 501–510, Virtual, 2021.

[8] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 249–256, Sardinia, Italy, 2010.

[9] T. G. Halle and K. E. Darling-Churchill. Review of measures of social and emotional development. *Journal of Applied Developmental Psychology*, 45:8–18, 2016.

[10] S. Jiang, C. Wang, and D. J. Weiss. Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, 7:109, 2016.

[11] A. Joshi, S. Kale, S. Chandel, and D. K. Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396, 2015.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, 2015.

[13] R. S. Kreitchmann, F. J. Abad, V. Ponsoda, M. D. Nieto, and D. Morillo. Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of likert items. *Frontiers in Psychology*, 10: 2309, 2019.

[14] J. Li, F. Wang, Q. Liu, M. Zhu, W. Huang, Z. Huang, E. Chen, Y. Su, and S. Wang. HierCDF: A Bayesian network-based hierarchical cognitive diagnosis framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 904–913, Virtual, 2022.

[15] M. Li, H. Qian, J. Lv, M. He, W. Zhang, and A. Zhou. Foundation model enhanced derivative-free cognitive diagnosis. *Frontiers of Computer Science*, 19(1):191318, 2025.

[16] X. Li, S. Guo, J. Wu, and C. Zheng. An interpretable polytomous cognitive diagnosis framework for predicting examinee performance. *Information Processing & Management*, 62(1):103913, 2025.

[17] S. Liu, T. Hu, H. Chai, Z. Su, and X. Peng. Learners' interaction patterns in asynchronous online discussions: An integration of the social and cognitive interactions. *British Journal of Educational Technology*, 53 (1):23–40, 2022.

[18] S. Liu, H. Qian, M. Li, and A. Zhou. QCCDM: A Q-augmented causal cognitive diagnosis model for student learning. In *Proceedings of the 26th European Conference on Artificial Intelligence*, pages 1536–1543, Kraków, Poland, 2023.

[19] S. Liu, J. Shen, H. Qian, and A. Zhou. Inductive cognitive diagnosis for fast student learning in web-based intelligent education systems. In *Proceedings of the ACM Web Conference 2024*, pages 4260–4271, Singapore, 2024.

[20] Y. Liu, T. Zhang, X. Wang, G. Yu, and T. Li. New development of cognitive diagnosis models. *Frontiers of Computer Science*, 17(1):171604, 2023.

[21] H. Ma, Y. Yao, C. Wang, S. Song, and Y. Yang. AD4CD: Causal-guided anomaly detection for enhancing cognitive diagnosis. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pages 12337–12345, Philadelphia, PA, USA, 2025.

[22] T. Nemoto and D. Beglar. Likert-scale questionnaires. In *JALT 2013 Conference Proceedings*, pages 1–6, Tokyo, Japan, 2014.

[23] R. Primi, D. Santos, F. De Fruyt, and O. P. John. Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology*, 72(3):447–465, 2019.

[24] T. Qi, M. Ren, L. Guo, X. Li, J. Li, and L. Zhang. ICD: A new interpretable cognitive diagnosis model for intelligent tutor systems. *Expert Systems with Applications*, 215:119309, 2023.

[25] H. Qian, S. Liu, M. Li, B. Li, Z. Liu, and A. Zhou. ORCDF: an oversmoothing-resistant cognitive diagnosis framework for student learning in online education systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2455–2466, Barcelona, Spain, 2024.

[26] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1):1–97, 1969.

[27] F. Samejima. Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, 60(4):549–572, 1995.

[28] J. Shen, H. Qian, W. Zhang, and A. Zhou. Symbolic cognitive diagnosis via hybrid optimization for intelligent education systems. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 14928–14936, Vancouver, Canada, 2024.

[29] Y. Su, S. Shen, L. Zhu, L. Wu, Z. Huang, Z. Cheng, Q. Liu, and S. Wang. Global and local neural cognitive modeling for student performance prediction. *Expert Systems with Applications*, 237(Part C): 121637, 2024.

[30] J. B. Sympson. A model for testing with multidimensional items. In *Proceedings of the 1977 Computerized Adaptive Testing Conference*, volume 14, Minneapolis, MN, 1978.

[31] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Chen, Y. Yin, Z. Huang, and S. Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, USA, 2020.

[32] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Yin, S. Wang, and Y. Su. Neuralcd: A general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8312–8327, 2023.

[33] F. Wang, W. Gao, Q. Liu, J. Li, G. Zhao, Z. Zhang, Z. Huang, M. Zhu, S. Wang, W. Tong, and E. Chen. A survey of models for cognitive diagnosis: New developments and future directions. *arXiv preprint arXiv:2407.05458*, 2024.

[34] B. Weijters, N. Schillewaert, and M. Geuens. Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, 36:409–422, 2008.

[35] B. Weijters, M. Geuens, and N. Schillewaert. The stability of individual response styles. *Psychological Methods*, 15(1):96, 2010.

[36] E. Wetzel, J. R. Böhnke, and A. Brown. *Response Biases*. Oxford University Press, 2016.

[37] S. Yang, M. Chen, Z. Wang, X. Yu, P. Zhang, H. Ma, and X. Zhang. Disengcd: A meta multigraph-assisted disentangled graph learning framework for cognitive diagnosis. In *Advances in Neural Information Processing Systems 38*, Vancouver, Canada, 2024.

[38] L. Zhang and P. Chen. A neural network paradigm for modeling psychometric data and estimating IRT model parameters: Cross estimation network. *Behavior Research Methods*, 56(7):7026–7058, 2024.